IA DE CONFIANCE

OPPORTUNITÉ STRATÉGIQUE POUR UNE SOUVERAINETÉ INDUSTRIELLE ET NUMÉRIQUE

Julien CHIARONI et Arno PONS

THINK-TANK DIGITAL NEW DEAL

juin 2022

COLLECTION DIGITAL NEW DEAL "NUMÉRIQUE DE CONFIANCE"

“

EN IMPOSANT L'IA
DE CONFIANCE COMME
VALEUR ÉTALON,
L'UE PEUT ÉDICTER
SES PROPRES CONDITIONS
EXTRATERRITORIALES
AU MARCHÉ MONDIAL.

SOMMAIRE

PRÉFACE 3

I. IDENTIFIER LES DÉFIS À RELEVER

1.1 QU'EST-CE QUE L'IA ?

1.1.1 Les défis humanistes de l'IA 5
1.1.2 Une absence de consensus sur la définition 7
1.1.3 En quoi l'IA est-elle différente des algorithmes dits "classiques" 7
1.1.4 Un champ scientifique et technique varié 9

1.2 LA NÉCESSITÉ DE RENFORCER LA CONFIANCE DANS L'IA

1.2.1 Qu'est-ce que la confiance dans les systèmes à base d'IA ? 15
1.2.2 Sur quoi repose la confiance dans l'IA ? 19
1.2.3 La difficulté à auditer et certifier des systèmes à base d'IA 28

II. ADRESSER UN DOUBLE ENJEU POLITIQUE

2.1 LA CONFIANCE, UN ENJEU DE SOUVERAINETÉ NUMÉRIQUE

2.1.1 L'autonomie stratégique, ambition raisonnable de la souveraineté 33
2.1.2 Protéger, condition sous-jacente à la confiance 37
2.1.3 Atteindre l'autonomie stratégique par l'écosystème de confiance 40

2.2 LA CONFIANCE, UN ENJEU DE COMPÉTITIVITÉ POUR L'EUROPE

2.2.1 Faire de la confiance la valeur étalon du marché 43
2.2.2 Un déploiement prudent de l'IA expliqué par des difficultés d'industrialisation et un manque de confiance 43
2.2.3 Une analyse du marché de l'IA de confiance sur des filières industrielles stratégiques pour l'Europe 46

III. BÂTIR UNE STRATÉGIE INDUSTRIELLE PAR L'IA DE CONFIANCE

3.1 UNE STRATÉGIE OFFENSIVE PAR LA RÉGULATION

3.1.1 Faire de la confiance un véritable avantage compétitif pour les européens 51
3.1.2 Le règlement IA comme première pierre à l'édifice d'une UE ambitieuse 53
3.1.3 Une approche volontariste par la norme 56
3.1.4 Garantir l'équilibre entre réglementation, normes et innovation par les sandboxes 61

3.2 UNE STRATÉGIE INDUSTRIELLE PAR LA COOPÉRATION

3.2.1 Accroître les efforts de RDI et de formation en IA de confiance 63
3.2.2 Créer une InfraTech de l'IA de confiance 66
3.2.3 Favoriser l'adoption via les data spaces et des cas d'application industriels 72
3.2.4 Une gouvernance industrielle, unifiée IA & Data pour le numérique de confiance 75

RÉCAPITULATIF DES RECOMMANDATIONS 81

CONCLUSION 83

REMERCIEMENTS 84

PRÉFACE

La crise Covid a jeté une lumière crue sur la « Désindustrialisation de la France[1] », soulignant sa dépendance à des produits manufacturés aussi basiques que les masques. La nécessité de faire renaître une industrie nationale forte et souveraine s'est imposée au point que le ministre de l'Économie et des finances du premier Gouvernement Borne, s'est adjoint la fonction de souveraineté industrielle. Mais Bruno Le Maire est aussi ministre de la souveraineté numérique, montrant désormais le lien entre industrie et numérique, et la dimension stratégique de l'intelligence artificielle : « L'intelligence artificielle est une question existentielle pour nos nations, celles qui ne maîtriseront pas l'IA seront vassalisées et perdront leur souveraineté[2] »

Nous devons pour cela naturellement réfléchir au niveau continental, en confortant la position de l'Europe en tant que pôle mondial d'excellence dans le domaine de l'IA, et porteuse de ses valeurs humanistes :

Pour que l'IA soit garante de ces valeurs, nous devons faire preuve envers elle d'une grande exigence éthique et politique. « L'IA des Lumières » telle que l'entend le think-tank Digital New Deal, est une « IA de confiance » qui refuse que des biais idéologiques – qu'ils proviennent d'une Silicon Valley purement marchande et servant les intérêts des oligopoles du Web ou d'une Chine aux relents totalitaires - viennent orienter aujourd'hui sa définition, pour être traduits ensuite en biais technologiques avec tous les impacts que l'on connaît : économiques, sociétaux, ou écologiques. D'où l'importance de traduire nos valeurs en standards mondiaux grâce au paquet réglementaire européen[3] qui crée cet espace de confiance au bénéfice des citoyens et des entreprises. Car alors, toute société étrangère sera obligée de respecter ce cadre de confiance pour commercialiser son algorithme au sein de l'UE.

Et pour que notre excellence en recherche d'IA soit traduite en capacité de leadership industriel, la réglementation seule ne suffit pas, nous devons absolument créer une coopération industrielle au niveau européen à la hauteur de nos ambitions. Il est tout à fait possible d'écrire cette nouvelle page pour l'industrie française et européenne d'ici à 2030. La France a déjà lancé des programmes ambitieux que cette note se propose de compléter pour accroître sa dimension continentale et garantir son impact.

Les auteurs présentent ici une méthode pour y parvenir, en faisant d'une part de l'IA de confiance la pointe de la flèche d'un écosystème de confiance, garant de notre autonomie stratégique ; et d'autre part en capitalisant sur notre culture industrielle (systèmes critiques) pour saisir cette opportunité historique de conquête du marché global de l'IA

Olivier Sichel,
Président Digital New Deal

[1] *Désindustrialisation de la France*, 1995-2015, Nicolas Dufourcq, Odile Jacob, 2022
[2] *Déclaration du ministre de l'Économie lors de l'inauguration de l'Institut Interdisciplinaires d'IA*, octobre 2019
[3] RGPD (Règlement Général sur la Protection des Données), DSA (Digital Services Act), DMA (Digital Markets Act), DGA (Data Governance Act), DA (Data Act), AI Act (Artificial Intelligence Act)

LA VOCATION DE L'EUROPE EST DE DÉFINIR UNE IA DES LUMIÈRES : HUMANISTE, TRANSPARENTE ET RESPONSABLE.

I. IDENTIFIER LES DÉFIS À RELEVER

L'Intelligence Artificielle est au cœur des préoccupations et des fantasmes, sa simple définition constitue un défi en soi. Il est donc crucial que nous apportions la nôtre, afin de pouvoir offrir au monde une vision européenne, c'est-à-dire humaniste, de ce sujet matriciel qui devient de plus en plus structurant dans les champs économiques, sociaux, écologiques et démocratiques. L'Europe ne doit pas reproduire les erreurs du passé en se laissant imposer une définition « siliconienne » de l'IA, comme elle l'avait laissé faire pour Internet avec toutes les conséquences que l'on connaît. Nous devons être capables de penser et opérer une « IA des Lumières » qui soit humaniste, transparente et responsable. Un Humanisme tel que nous l'entendons au XXIe siècle, c'est-à-dire qui place l'Humain, ses valeurs, mais aussi ses interactions avec son milieu naturel, au centre de tout. L'humanisme a toujours eu pour objectif l'épanouissement de l'Homme, et sa confiance dans sa capacité à évoluer de manière positive dans le respect de son environnement, un environnement entendu aujourd'hui au sens le plus large et écologique du terme.

1.1. QU'EST-CE QUE L'IA ?

Pour simplifier la lecture de ce rapport, **nous emploierons l'expression « l'intelligence artificielle » ou « l'IA » pour qualifier les systèmes à base d'IA,** ou reposant sur des technologies à base d'IA.

1.1.1. Les défis humanistes de l'IA

L'IA se déploie dans tous les aspects de notre vie

Le potentiel de l'intelligence artificielle (IA) et de ses développements alimente un imaginaire puissant, reflétant les angoisses provoquées par les grands bouleversements technologiques

À ce potentiel répond l'angoisse de dépossession face à une technologie amenée à se déployer partout : santé, finance, industrie, sécurité, transports, commerce, service public, tous les pans de nos vies sont concernés. S'il est parfois exagéré, le potentiel de ces technologies de rupture est toutefois bien réel et déjà perceptible dans de nombreux domaines de nos vies quotidiennes.

Dans cette « vie algorithmique[4] », **l'intelligence artificielle déborde déjà très largement des programmes de recherche universitaires** et devient progressivement un élément qui conditionne l'organisation des connaissances, la production de biens et services, et même les prises de décision, au point de devenir un facteur d'organisation des collectivités et un vecteur de puissance pour les États[5].

Les progrès exponentiels des technologies de l'information (loi de Moore[6]), la connectivité toujours plus étendue (loi de Metcalfe[7]), ainsi que l'accès à des masses de données toujours plus importantes (*Big Data*), accélèrent de manière continue les avancées de l'IA.

[4] *La Vie algorithmique. Critique de la raison numérique*, Éric Sadin, éditions L'Echappée, 2015
[5] *Pourquoi l'intelligence artificielle est le futur de la croissance*, Mark Purdy et Paul Daugherty Accenture, 2016
[6] Loi empirique de Gordon Moore sur l'évolution exponentielle de la puissance de calcul, selon laquelle la puissance d'un processeur est multipliée par deux tous les deux ans.
[7] Loi théorique et empirique de l'effet de réseau énoncée par Robert Metcalfe, affirmant que l'utilité d'un réseau est proportionnelle au carré du nombre de ses utilisateurs (plus il y a d'utilisateurs sur un réseau, plus celui-ci a de la valeur).

Les rapports entre les humains et l'IA

Cette nouvelle vie algorithmique regorge de défis pour nos sociétés. Quelle sera, par exemple, notre place dans un monde où les systèmes à base d'IA seront omniprésents ? Comment permettre à l'humain de conserver son attention, son intérêt, son sentiment d'utilité et sa dignité au travers de sa créativité, de son mérite et de sa responsabilité[8] ? Face au réchauffement climatique qui s'accélère, comment mettre l'IA au service de la transition écologique ? Quels seront également les impacts sur le monde du travail et l'économie ?

La lutte contre le réchauffement climatique peut bénéficier des apports combinés de l'IA et de la Data. En effet, l'objectif de zéro émissions nettes de gaz à effet de serre (GES) à l'horizon 2050 implique de découpler la croissance économique de l'usage intensif des ressources physiques[9]. C'est précisément ce que peut faire l'intelligence artificielle par l'analyse de grandes masses de données (*Big Data*). L'IA pourrait ainsi jouer un rôle de premier plan en apportant plus d'efficience et de transparence dans de nombreux domaines (Smart Cities & Digital Twins, transports, industrie, construction, gestion des déchets, agriculture, économie circulaire, énergie, etc.), ou en améliorant notre compréhension de phénomènes complexes (fusion nucléaire[10], etc.). Toutefois, nous ne pouvons exclure la contribution « négative » de l'IA[11], et plus largement du numérique, que seule une approche frugale, en données et en énergie, couplée à des avancées techniques, pourra réduire.

Concernant le monde du travail, le déploiement de l'IA à grande échelle fait émerger la crainte d'une perte massive d'emplois. La destruction créatrice schumpétérienne[12] suppose que de nouveaux emplois se créent presque naturellement quand d'autres, rendus obsolètes par l'automatisation, sont détruits. Mais nos sociétés seront-elles en mesure d'absorber ces destructions face au rythme effréné des évolutions technologiques ? Le temps d'adaptation des filières est en effet un facteur incompressible, et la capacité des individus à se former et se reconvertir est forcément limitée. Qu'en sera-t-il également du rapport entre l'humain et l'IA dans le monde du travail ?

Toutes ces questions, parfois philosophiques et éthiques, ne doivent ni être éludées, ni instrumentalisées. Il ne tient qu'à nous de construire des rapports sains avec ces technologies et d'anticiper leurs répercussions sur l'économie. **L'IA n'a de sens que si son utilisation est alignée avec les valeurs et principes que défendent les communautés humaines.** Elle sera alors acceptée par les individus, les marchés et la société, et deviendra un vecteur de prospérité.

Pour ce faire, le développement de l'IA doit dès aujourd'hui prendre en compte des problématiques sociétales de long terme telles que la protection des citoyens, de l'environnement, de l'emploi ou encore de la démocratie. Et **la notion de « confiance »** devient **alors essentielle : sous-jacent fondamental aux relations dans nos sociétés, composante indispensable de la réponse aux défis humanistes, et du respect des droits fondamentaux défendus par l'Union européenne.**

L'objet de ce rapport n'est pas d'être exhaustif sur les défis humanistes de l'IA, ni d'en proposer une analyse précise. Plusieurs livres et analyses traitent spécifiquement de ces sujets. En revanche, **ce rapport porte principalement sur la « confiance », notion aussi complexe que fondamentale, y compris dans le champ de l'IA.**

---

[8] Laurence Devillers - 3 dimensions de l'interaction humain-machine, suite aux travaux du laboratoire sur le futur du travail opéré par l'institut d'innovation Matrice.

[9] *The Role of Artificial Intelligence in the European Green Deal*, Policy Department for Economic, Scientific and Quality of Life Policies Directorate-General for Internal Policies, European Parliament, Mai 2021

[10] https://www.inria.fr/fr/fusion-nucleaire-comment-simplifier-la-simulation-des-plasmas

[11] N. Thompson et coll. Deep Learning Diminishing Returns. https://spectrum.ieee.org/deep-learning-computational-cost

[12] Désigne le processus continuellement à l'œuvre dans les économies et qui voit se produire de façon simultanée la disparition de secteurs d'activité économique conjointement à la création de nouvelles.

1.1.2. Une absence de consensus sur la définition

L'IA prend une place de plus en plus importante dans les débats publics, mais fait remarquable : **la définition de l'intelligence artificielle ne fait pas consensus,** et nous ne disposons pas à l'heure actuelle d'une définition légale précise et générale.

Des définitions multiples

Marvin Minsky (1927-2016), l'un des fondateurs de l'Intelligence artificielle, définissait l'intelligence artificielle comme suit : **« l'IA consiste à faire faire à une machine ce que l'homme fait moyennant une certaine intelligence »**. L'AAAI (Association for the Advancement of Artificial Intelligence) a retenu comme définition : **« La compréhension scientifique des mécanismes sous-jacents à la pensée et au comportement intelligent et leur incarnation dans les machines »**.

La plupart des définitions actuelles de l'Intelligence Artificielle (IA) intègrent la notion d'apprentissage. De nombreux raisonnements humains reposent sur une approche empirique d'observations statistiques ou probabilistes. Un processus décisionnel automatique reposant sur une telle approche peut donc raisonnablement être qualifié d'IA, dans son acception d'imitation partielle de la capacité de décision humaine. **L'IA est alors considérée comme un ensemble de mécanismes techniques et scientifiques permettant de reproduire la plupart des processus cognitifs humains tels que l'apprentissage, l'intuition, l'auto-amélioration, la créativité, la planification de tâches ou encore la compréhension et la génération de langage naturel.**

Mais compatibles entre elles

Ainsi, de nombreuses définitions existent, par exemple celles de l'ISO-24765, de l'Académie des Technologies[13] ou de la Commission européenne[14], mais toutes sont en réalité compatibles entre elles. On y retrouve les capacités cognitives données à des machines (matériel et logiciel) de réaliser des tâches demandant de l'intelligence quand elles sont exercées par des humains. Des exemples d'application de ces capacités se trouvent dans le véhicule autonome, les agents conversationnels, ou la reconnaissance d'images. Les technologies au service de ces applications reposent par exemple sur la représentation des connaissances, le raisonnement, l'apprentissage automatique, la planification, etc…

1.1.3. En quoi l'IA est-elle différente des algorithmes dits "classiques"

Les algorithmes « classiques »

L'informatique en général repose sur des algorithmes. Un algorithme peut être considéré comme un ensemble de règles ordonnant un nombre fini d'opérations pour résoudre un problème. Pour fonctionner, un ordinateur a besoin qu'un programmeur lui fournisse un programme informatique composé de multiples algorithmes, écrits dans un langage que la machine peut interpréter. Les algorithmes traduisent des entrées (inputs) en sorties (outputs). **En algorithmie classique, la machine applique à une entrée les étapes de l'algorithme, dans l'ordre demandé jusqu'à obtenir une sortie. C'est une sorte de « recette de cuisine » dont toutes les composantes sont connues et décrites de manière parfaitement explicite.**

Une nouvelle logique apportée par l'IA symbolique, puis connexionniste

Ce que nous appelons communément l'Intelligence Artificielle (ou IA) rompt avec l'algorithmique classique, où le programmeur décrit de manière directe et explicite la manière de résoudre un problème. Historiquement, la conception d'algorithmes d'IA émerge dans les

[13] *Rapport Renouveau de l'Intelligence artificielle et de l'apprentissage automatique*, paru en 2018
[14] *Réglementation européenne sur l'intelligence artificielle*, Commission européenne, 2021

années 1950 au travers de deux courants. **L'IA à base de connaissances**, qualifiée aujourd'hui de Good Old Fashioned AI (GOFAI) ou d'**IA symbolique**, qui se base quasi exclusivement sur le raisonnement symbolique et la logique. Elle se distingue de l'**IA dirigée par les données**, appelée aussi **IA statistique et connexionniste**, sous les feux de la rampe ces dernières d'années avec la collecte massive des données (*Big Data*) et l'arrivée de l'IA subsymbolique (et du *deep learning*), bien qu'aussi ancienne.

L'IA symbolique comme la définit Nicholas Asher (Directeur scientifique de l'Institut Interdisciplinaire en Intelligence Artificielle ANITI), **« utilise le raisonnement formel et la logique ; c'est une approche cartésienne de l'intelligence, où les connaissances sont encodées à partir d'axiomes desquels on déduit des conséquences. La prédiction doit être juste même si l'on ne dispose pas de données exhaustives »**. Ce paradigme reste pertinent pour la résolution de problèmes complexes sous contraintes, dans un contexte d'incertitude.

L'IA basée sur les données et l'apprentissage automatique, notamment connexionniste, en revanche, cherche à apprendre à partir d'exemples ou à découvrir des régularités contenues dans des données, sans les connaître au préalable. Les approches connexionnistes ont également une capacité de généralisation leur permettant de traiter un cas, ou de reconnaître une forme dans des conditions différentes de celles de l'apprentissage.

Leurs domaines d'emploi diffèrent cependant. **Alors que l'IA connexionniste est l'IA des sens (perception), l'IA symbolique est celle du sens (raisonnement).** Plusieurs travaux cherchent aujourd'hui à hybrider ces deux paradigmes, comme le souligne Nicholas Asher : **"L'addition de ces deux courants, IA symbolique et IA connexionniste, constitue le défi d'aujourd'hui", on parle d'IA hybride.** Par exemple, l'apprentissage par renforcement consiste à récompenser les comportements souhaités et/ou à sanctionner les comportements non désirés avec des stratégies de récompense ou de sanction basées sur des connaissances métiers ou heuristiques issues de l'IA symbolique.

Le tournant du *machine learning*

Le développement des techniques de *machine learning*, à partir des années 80, marque un tournant : on passe d'une logique de programmation à une logique d'apprentissage. L'apprentissage machine s'affranchit des règles expertes et fonctionne à partir d'exemples : le programmeur fournit à la machine des séries d'exemples (*inputs, outputs*) qui vont "entraîner" ou permettre de construire un modèle statistique capable de déterminer un output à partir d'un nouvel *input* donné. Le succès du *machine learning* est tel qu'il y a aujourd'hui souvent confusion entre IA et *machine learning.*

Le *machine learning* est particulièrement adapté pour traiter des problèmes en monde ouvert (*in vivo*), comme la reconnaissance d'image ou le traitement automatique du langage (TAL, Traitement Automatique du Langage), où l'ensemble des situations possibles est difficile à connaître et à maîtriser pour un être humain. Sa performance va par contre dépendre des techniques algorithmiques employées et, dans de nombreux cas de figure, du volume et de la qualité des données qui vont servir d'exemple, et qui sont fournies ou annotées par des humains. Il faut un très grand nombre de photos (certaines avec l'objet visé et d'autres non) correctement annotées par des humains pour qu'un programme de reconnaissance d'image fonctionne de manière satisfaisante.

Aussi puissant soit-il pour la reconnaissance d'image, **le *machine learning* est très loin d'être pertinent pour résoudre tous les types de problèmes.** L'apprentissage machine est par exemple moins pertinent pour les problèmes comportant un nombre fini et limité de cas de figure ou de paramètres potentiels.

1.1.4. Un champ scientifique et technique varié

Les sous-domaines clés de l'IA

L'intelligence artificielle est donc plurielle. Pour le dire simplement, cette technologie, ou plutôt ces technologies recouvrent un ensemble de disciplines, de techniques et de supports disparates poursuivant des objectifs apparentés à certaines facultés humaines. Dans son livre blanc *Intelligence Artificielle. Les défis actuels et l'action d'Inria*[15], l'Institut national de recherche en sciences et technologies du numérique (Inria) français propose d'ailleurs une structuration en sous-domaines. **Certains de ces sous-domaines de l'IA ont pris une importance considérable au cours des années 2000 grâce à l'exploitation du *Big Data*.** En voici quelques-uns :

- **L'apprentissage supervisé (*supervised learning*)** : Dans ces algorithmes, les données d'*input* (données d'entraînement), sont étiquetées au préalable par des humains pour indiquer à la machine à quoi elles correspondent. Le modèle est entraîné à partir de ces données, et doit parvenir à réaliser des classements ou des prédictions.
- **L'apprentissage non supervisé (*unsupervised learning*)** : Ici, les données d'*input* ne sont pas étiquetées et n'ont pas de résultat connu. Le modèle doit lui-même observer des structures et des tendances dans les données.
- **L'apprentissage par renforcement (*reinforcement learning*)** : L'apprentissage par renforcement consiste pour une IA à apprendre les actions à effectuer à partir d'expériences, de façon à optimiser une récompense quantitative au cours du temps. L'agent est plongé au sein d'un environnement et prend ses décisions en fonction de son état courant ; en retour, l'environnement procure à l'agent une récompense qui peut être positive ou négative.
- **Les réseaux de neurones** : Un réseau de neurones artificiels est un ordinateur construit pour répliquer le fonctionnement des connexions du système nerveux du cerveau. Le neurone artificiel est conçu comme un automate doté d'une fonction de transfert qui transforme ses entrées en sorties selon des règles précises. Assemblés en réseau, ces neurones artificiels peuvent opérer rapidement des classements et apprendre progressivement à les améliorer. L'idée forte de cet assemblage est de pouvoir, à partir d'un ensemble de calculs rudimentaires et locaux au niveau de chaque neurone, apprendre une fonction plus complexe en multipliant les couches de neurones.
- **L'apprentissage profond (*deep learning*)** : L'amélioration continue des performances d'apprentissage machine passe un cap en 2010 avec le développement du *deep learning*. Domaine de pointe du *machine learning*, cette technique correspond à un type spécifique d'apprentissage automatique basé sur l'utilisation de grands réseaux neuronaux (profonds) pour apprendre des représentations abstraites des données d'entrée, par l'utilisation de couches multiples. Elle permet donc à la machine de reconnaître par elle-même des concepts complexes. Né de la combinaison des algorithmes d'apprentissage avec les réseaux de neurones et l'utilisation du *Big Data*, le *deep learning* a révolutionné l'IA. Les applications se retrouvent dans les moteurs de recherche, le diagnostic médical, ou encore les véhicules autonomes.
- **Le Traitement Automatique du Langage, parlé ou écrit (TAL en français et *Natural Language Processing* – NLP en anglais)** : Parmi les applications les plus répandues, le TAL recouvre plusieurs types de tâches linguistiques : analyse de parole ou de texte pour en extraire du

[15] *Intelligence Artificielle. Les défis actuels et l'action d'Inria*, livre blanc, coord. Bertrand Braunschweig, 2020 2e édition.

sens, traduction automatique, génération textuelle, interrogation automatisée de bases de données (questions-réponses), etc. Il fonctionne grâce à plusieurs techniques (l'indexation sémantique ou *Part-of-Speech tagging*, la reconnaissance d'entités nommées *Named Entity Recognition, le Parsing*, l'analyse grammaticale ou sémantique), et se retrouve par exemple dans les fonctions de reconnaissance vocale de nos smartphones.

Qu'ils soient optimisés pour des réseaux de neurones ou qu'ils relèvent de conceptions plus classiques, **l'efficacité des algorithmes reste néanmoins dépendante de leur capacité à traiter une quantité toujours plus importante de données et de variables.** Dans les cas d'explosion combinatoire (trop de variables pour un algorithme), il est nécessaire de faire des choix stratégiques pour obtenir un résultat dans un temps raisonnable.

L'importance de la représentation des connaissances

Aujourd'hui, même si l'IA connexionniste a nettement pris le dessus, l'IA symbolique reste également active. La conception de systèmes à base de connaissances capables de réaliser des fonctions de raisonnement symbolique constitue un domaine majeur en IA. De tels systèmes nécessitent en particulier une représentation adéquate des connaissances utiles, ainsi que des mécanismes efficaces de raisonnement. Faisant appel aux modèles et méthodes de la logique du premier ordre (dite calcul des prédicats), l'IA symbolique a donné lieu aux réseaux sémantiques et ontologies, aux systèmes à base de connaissances, aux systèmes experts ou à la programmation par contraintes. On peut aussi y associer la logique floue, même si elle présente la particularité d'être associable aussi bien à du raisonnement formel qu'à du *machine learning*.

Reposant sur l'idée que **« l'intelligence est surtout liée à la connaissance plus qu'à un problème de raisonnement »**, Edward Feigenbaum, père du premier système expert DENDRAL définit en 1977, l'ingénierie des connaissances Knowledge Engineering (IC) comme **« l'art d'acquérir, de modéliser et de représenter la connaissance en vue de son utilisation par un ordinateur »**. Pour cela, on peut s'appuyer sur :

- **Des représentations logiques** construites selon une syntaxe précise. Ainsi, une base de connaissances est un ensemble de formules décrivant le domaine sur lequel s'appliquent des règles de raisonnement, comme dans le langage PROLOG[16].
- **Des réseaux sémantiques**, des graphes conceptuels[17] bénéficiant de mécanismes de raisonnement induits par les opérations de graphes comme l'homomorphisme de graphes.
- **Des ontologies** qui constituent en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts. On peut dire que « l'ontologie est aux données ce que la grammaire est au langage »[18]. Elles sont aujourd'hui utilisées pour modéliser et partager un ensemble de connaissances dans un domaine donné, comme par exemple dans le Web sémantique[19] ou en génie logiciel.
- **L'inférence, ou le raisonnement,** qui repose sur des opérations de déduction à partir d'informations implicites. Ainsi, ce mécanisme permet de créer des liens entre les informations afin d'en tirer une assertion, une conclusion ou une hypothèse. Par exemple, l'inférence bayésienne est un raisonnement permettant de déduire la probabilité de survenance ou non d'un événement.

[16] Alain Colmerauer et Philippe Roussel développent à Marseille en 1972 le langage PROLOG (Programmation Logique), au départ pour traiter le langage. Ce programme est une suite de clauses de Horn sur lesquelles opère un mécanisme de raisonnement utilisant le principe de résolution. Comme LISP, Prolog utilise massivement la structure de liste et est naturellement récursif.

[17] Les graphes conceptuels sont introduits par John F. Sowa (chercheur à IBM) en 1984 pour formaliser la différence entre les concepts individuels (instances), les concepts génériques, et les classes (types).

[18] https://fr.wikipedia.org/wiki/Ontologie_(informatique)

[19] Le Web sémantique est une extension du Web standardisée par le World Wide Web Consortium (W3C)2. Ces standards encouragent l'utilisation de formats de données et de protocoles d'échange normés sur le Web - Wikipedia

Les méthodes de raisonnement par cas (*Case-based Reasoning*), introduites au début des années 80 par Roger Schank, reposent sur l'idée que l'on raisonne parfois en utilisant des analogies. Ces approches connaissent un nouveau regain car elles ont le bon goût d'être plus facilement explicables.

A la frontière entre programmation mathématique et IA, la programmation par contraintes (PPC) est apparue à la fin des années 1980, pour la résolution de problèmes combinatoires complexes tels que les problèmes de planification, d'ordonnancement et d'allocation de ressources. Cette technologie repose sur le paradigme de séparation de la modélisation du problème avec sa résolution. **« En informatique, de toutes les approches en programmation, la programmation par contraintes se rapproche le plus de l'idéal : l'utilisateur décrit le problème, l'ordinateur le résout. »**[20]. La modélisation du problème peut inclure la connaissance métier et se fait par le biais d'un ensemble de relations logiques : les contraintes. Des mécanismes de propagation des contraintes dans un arbre de branchement permettent la réduction du domaine de décisions. Le solveur de contraintes calcule alors une ou plusieurs solutions en instanciant chacune des variables à une valeur satisfaisant simultanément toutes les contraintes. Aujourd'hui, de nombreuses applications sont déployées, comme dans la grande distribution qui optimise depuis longtemps sa logistique et sa gestion des stocks.

Enfin, **l'IA symbolique est souvent utilisée pour concevoir des systèmes d'aide à la décision**. Rappelons qu'un problème de décision consiste en un choix, ou un classement, entre plusieurs hypothèses mutuellement exclusives résultant d'un processus qui tient compte des connaissances que l'on a sur l'état du monde, des préférences et/ou de l'objectif à atteindre. Ces connaissances peuvent être empreintes d'incertitude et les préférences sont par nature nuancées. Un outil simple pouvant être mis en œuvre est alors l'arbre de décision, qui représente un ensemble de choix sous la forme graphique d'un arbre. Les différentes alternatives sont alors les feuilles de l'arbre et sont atteintes en fonction de décisions prises à chaque étape. Cependant, la définition et l'utilisation d'un ou de plusieurs critères de sélection sont nécessaires. Contrairement à la situation monocritère qui peut être résolue assez facilement, la décision multicritère nécessite des méthodes plus élaborées. Parmi les techniques utilisées on peut citer la méthode du « *What-if* » ou l'agrégation multicritère. Cette dernière consiste à évaluer globalement les différents candidats ou solutions proposées, à partir de la fusion des appréciations partielles.

[20] E. Freuder

L'INTELLIGENCE ARTIFICIELLE EN QUELQUES DATES

1943-1955 En 1943, les travaux de McCulloch et Pitts introduisent un modèle de neurone artificiel. Quelques années plus tard, Hebb propose une règle pour modifier des connexions entre neurones, puis Minsky et Edmonds construisent le **premier réseau de neurones**.

Turing publie en 1950, son article « Computing Machinery and Intelligence » qui explore le problème et propose une expérience **(test de Turing)** dans le but d'identifier le moment où une machine sera capable d'imiter la conversation humaine.

1956 Le congrès sur l'IA au Dartmouth College réunit les meilleurs spécialistes internationaux de l'époque. L'expression « *intelligence artificielle* » est utilisée pour la première fois, et désigne « *Tout aspect de l'intelligence, et notamment de l'apprentissage [qui] peut être décrit de manière tellement précise qu'une machine peut le simuler*[21] ». Le congrès de Dartmouth est considéré comme le **moment fondateur de la recherche fondamentale en IA.**

1960's La période qui suit est celle de l'essor de l'IA. En 1958, John McCarthy crée le langage informatique LISP (nom créé à partir de *list processing*) qui permet de faciliter la programmation d'IA. Un grand nombre de programmes sont développés principalement aux États-Unis (Stanford, MIT, Carnegie-Mellon), mais aussi en Écosse (Édimbourg) et au Japon pour résoudre divers problèmes tels que :

- « *Logic Theorist* » et « *Geometry Theorem Prover* » qui sont en mesure de **prouver certains théorèmes mathématiques,**
- « *General Problem Solver* » qui réussit quant à lui à **résoudre des puzzles avec un raisonnement semblable à celui de l'humain.** Cette période voit une poursuite de la recherche sur les réseaux de neurones et l'invention du **premier robot capable de raisonner sur ses propres actions**, le Shakey.

En 1965, Lotfi Zadeh propose une extension de la logique classique qui permet la modélisation des imperfections des données et se rapproche dans une certaine mesure de la flexibilité du raisonnement humain : **la logique floue (« *fuzzy logic* »).**

1970's Au début de cette décennie les capacités des programmes d'IA sont limitées, et les programmes les plus performants peinent à manipuler des versions simplistes des problèmes qu'ils sont censés résoudre. **La puissance et la mémoire de l'époque sont un frein aux applications pratiques** : le langage naturel y est limité à 20 mots car la mémoire ne peut en contenir plus. L'IA connaît alors une période moins florissante. En 1969, Minsky et Papert publient « *Perceptrons* » dans lequel ils démontrent les limitations des réseaux de neurones à une seule couche. **De nombreux fonds et subventions sont réduits** et plusieurs projets abandonnés en raison du pessimisme sur les possibilités réelles de l'IA. Ce moment constitue le **« premier hiver » de l'IA**.

Mais en dépit de ces années sombres, **les travaux ne s'interrompent pas pour autant**. Les systèmes experts apparaissent entre 1969 et 1979, imitant un expert afin de résoudre des problèmes par des règles. Le premier système expert, DENDRAL, est créé pour **déterminer la structure d'une molécule à partir de sa formule** et des résultats de sa spectrométrie de masse. Par la suite, d'autres systèmes d'expert voient le jour comme MYCIN qui avait pour but de **diagnostiquer des infections sanguines à un niveau proche de celui des médecins humains** experts dans ce domaine.

1980's En 1982, Hopfield propose des réseaux de neurones associatifs et on assiste à une **renaissance du connexionnisme**. En 1986, Rumelhart, Hinton et Williams publient l'algorithme de la rétropropagation du gradient qui permet d'optimiser les paramètres d'un réseau de neurones à plusieurs couches. À partir de ce moment, **la recherche sur l'apprentissage automatique (*machine learning*) à base de réseaux de neurones connaît un essor fulgurant.**

En parallèle, le logiciel Chinook devient le **1er programme informatique à résoudre entièrement le jeu de dames** : quelle que soit la situation initiale, l'ordinateur est désormais certain de gagner la partie. On note en même temps le **développement des algorithmes d'apprentissage qui sont à la base de l'apprentissage profond (*deep learning*)**. Durant la décennie, la puissance de calcul va augmenter de manière fulgurante, tandis que de nombreux secteurs vont commencer à utiliser des ordinateurs. Cette augmentation de la matière première (donnée), et des moyens (capacités de calcul), va accélérer le développement de la discipline.

[21] *Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, J. McCarthy, M. L. Minsky, N. Rochester, C. E. Shannon, 31 août 1955.

Le **« deuxième hiver » de l'IA** arrive à la fin des années 80 et au début des années 90. Les limites des systèmes experts à base de connaissances, provoquent l'abandon de nombreux projets et une forte réduction des investissements publics et privés.

1990's Malgré les restrictions budgétaires de la fin des années 80, les recherches en IA se poursuivent particulièrement dans le sous-domaine de l'apprentissage (*machine learning*), basé sur l'analyse statistique de grandes quantités de données. De telles tâches incluent la **reconnaissance d'image parmi un ensemble ou l'identification et la compréhension de mots dans une langue humaine.**

1997 L'ordinateur **Deep Blue**, développé par IBM, **bat Gary Kasparov aux échecs** dans un match en 6 parties concrétisant 40 ans plus tard la « prophétie » de 1957 d'Herbert Simon.

2000's Les années 2000 constituent **un virage sans précédent avec l'arrivée du Web 2.0 puis du Big Data**. Les techniques abandonnées dans les années 1980 sont réexploitées, avec cette fois-ci des moyens décuplés.

2010 **L'apprentissage profond (*deep learning*)** envahit progressivement la discipline, et se développe dans de nombreux domaines **(reconnaissance visuelle, traduction automatique, robotique, etc.)**. Combiné avec les données d'entraînement en masse (*Big Data*), **le *deep learning* permet d'obtenir des résultats inespérés** dans des domaines très divers : jeux, systèmes de recommandation, reconnaissance d'image, de la parole, traitement du langage naturel (génération, traduction, résumés automatiques, questions-réponses), diagnostic médical, prévision de séries temporelles.

2011 **Watson**, programme informatique d'IA conçu par IBM et dont la 1ère version ne mobilise pas d'apprentissage profond (*deep learning*), **devient le champion du jeu télévisé US « *Jeopardy!* »** en battant ses concurrents humains. Deep Blue et Watson sont tous les deux des supercalculateurs. Deep Blue (1997) reste sur le calcul de probabilités, tandis que Watson grâce à ses 90 serveurs et ses 2880 cœurs[22] offre une puissance de calcul permettant de répondre aux questions de *Jeopardy!* dans des temps équivalents à ses concurrents humains.

2015 **AlphaGo**, développé par l'entreprise DeepMind (rachetée en 2014 par Google), **bat Fan Hui, le champion européen de Go**. L'algorithme d'AlphaGo utilise des **techniques d'apprentissage sur la base d'exemples de nombreuses parties de jeu entre humains enregistrées. Puis en 2017 l'algorithme augmente encore ses capacités en surpassant tous les joueurs, y compris le champion du monde Lee SeDol, et en faisant preuve d'inventivité dans ses choix. Dernière évolution marquante** avec le successeur **AlphaZero : l'IA apprend par elle-même, sans exemple préalable de parties humaines.**

2020's **Les années 2020 consacrent le développement des enjeux de confiance et d'explicabilité dans l'IA**.

2022 La startup française NukkAI **bat les meilleurs joueurs de bridge du monde** au moyen d'un programme d'**IA hybride expliquant ses raisonnements.**

L'exemple du Bridge expliqué

Nook, développé par l'entreprise française NukkAI, devient le **premier programme d'IA à dépasser 8 champions du monde de Bridge,** dans la phase de jeu de la carte. Le jeu de Bridge avait jusqu'alors résisté à l'IA et aux méthodes purement numériques (à base de *deep learning* par exemple) car c'est un jeu à information incomplète où il faut tirer des inférences des actions et non-actions des adversaires.

Nook est une IA hybride dans le sens où elle est composée de trois modules reposant chacun sur des paradigmes d'IA différents : **1)** un module symbolique qui permet de restreindre la combinatoire et d'expliquer les décisions **2)** un module de recherche arborescente **3)** un module de réseaux de neurones modélisant les adversaires. **A chaque étape du jeu il est possible d'accéder à « ce que Nook avait en tête au moment de sa décision » ce qui permet à l'humain de comprendre pourquoi la machine a effectué un coup brillant par exemple.**

Par ailleurs, **l'utilisation de méthodes symboliques pour restreindre la combinatoire permet à Nook d'être très peu énergivore** : le challenge NukkAI a consommé 200.000 fois moins de ressources que celui sur le jeu de Go.

[22] *IBM Watson : tout savoir sur le super ordinateur*, CFTC-IBM, 05.07.2019

LA CONFIANCE CONSTITUE UN MUR POUR L'IA, DANS LA MESURE OÙ SON ENVERS LA MÉFIANCE APPARAÎT COMME UN FACTEUR DE RALENTISSEMENT, ET POURRAIT MENER À UN NOUVEL HIVER DE L'IA.

1.2. LA NÉCESSITÉ DE RENFORCER LA CONFIANCE DANS L'IA

1.2.1. Qu'est-ce que la confiance dans les systèmes à base d'IA ?

La dimension systémique de la confiance

La confiance est fondamentalement une relation entre deux (ou plusieurs) parties. Elle est le ciment social qui relie individus et organisations au sein d'une même communauté. **Mais cette confiance ne peut exister que dans un cadre plus global qui favorise la confiance et la protège : l'écosystème de confiance.**

Cet écosystème, constitué d'un ensemble d'acteurs de confiance, construit et entretient un corpus de valeurs, principes et règles, et développe un ensemble de mécanismes pour surveiller les comportements, juger et punir les entorses à la confiance. Il a pour but d'encourager et protéger les relations entre parties au sein d'un environnement donné en développant des règles, des standards et les moyens de les mettre en œuvre. Ce faisant, l'écosystème doit gérer les risques, surveiller les activités et anticiper les nouveaux comportements. Un écosystème de confiance repose sur un réseau étendu et organisé de tiers de confiance. Ces tiers sont indépendants, et peuvent être des organisations ou des personnes partageant des « principes » communs : une régulation, une accréditation, un audit, une certification, une norme, une juridiction, et les organes de mise en œuvre qui rendent possibles et crédibles les exigences et évaluation de fiabilité dans un environnement donné.

L'organe de gouvernance, le plus souvent un État, mais cela peut également s'appliquer à un consortium ou une organisation, **met en œuvre une analyse extensive des menaces que peut rencontrer son écosystème de confiance, et définit des critères pour s'assurer que l'écosystème et ses tiers sont bien dignes de confiance eux-mêmes**. Des tiers injustes, négligents, incompétents, biaisés ou corrompus peuvent ainsi menacer l'équilibre de l'écosystème.

La confiance possède donc une dimension systémique : un outil, une application ou une relation s'inscrivent dans un ensemble bien plus large incluant toutes les parties prenantes, voire la société dans son ensemble.

EFFONDREMENT DE LA CONFIANCE DANS L'ÉCOSYSTÈME : LE CAS DU BOEING 737 MAX

Le 29 octobre 2018, le vol Lion Air 610 s'écrase en mer au large de l'Indonésie. Le 10 mars 2019, le vol 302 d'Ethiopian Airlines s'écrase six minutes après son décollage d'Addis-Abeba. Ces deux accidents, impliquant la nouvelle version du Boeing 737 Max, ont coûté la vie à 346 personnes.

L'enquête du *Department of Justice* (DoJ) américain a depuis mis en évidence un défaut sur le logiciel anti-décrochage (MCAS), ainsi qu'une volonté délibérée de l'entreprise de dissimuler des informations au certificateur afin d'accélérer la mise en service du 737 Max. Le DoJ démontre que les défauts du logiciel anti-décrochage ont été aggravés par une formation insuffisante des pilotes. Par ailleurs, l'un des pilotes chargé de réaliser les essais est accusé d'avoir volontairement induit en erreur le régulateur américain de l'aéronautique (*Federal Aviation Administration* - FAA) en omettant de signaler les difficultés liées au logiciel de pilotage. Le pilote se serait même vanté d'avoir pu tromper ses interlocuteurs de la FAA pour pouvoir faire certifier le système...

Les conséquences financières de ces dysfonctionnements ont été radicales pour Boeing, qui affichait une perte de plus de 20 milliards de dollars depuis 2019, et a enregistré 565 annulations de commandes entre 2019 et 2020.

L'impact d'un tel dysfonctionnement se traduit par une perte de confiance de la société non seulement envers l'acteur incriminé (ici Boeing), mais aussi dans les mécanismes de régulation et de certification dans leur ensemble (les États, la FAA, etc.). Cette situation augmente le risque de voir le transport aérien en général ne plus être digne de confiance. Des logiciels aux pilotes d'essai, en passant par l'entreprise, le régulateur sectoriel, et in fine les États parties prenantes, la rupture de confiance s'affiche ici comme un risque systémique, et si la faillite d'une entreprise peut entraîner la faillite de l'ensemble du système, c'est en réalité un risque sociétal.

Les défaillances en série du Boeing 737 Max montrent dans quelle mesure la confiance, traduite dans des mécanismes de régulation, est une institution invisible qui soutient toute la société. Il n'y a donc pas de risque isolé lorsqu'il s'agit de confiance.

Sources : *Suspension du vol Boeing 737 Max*, Wikipédia, *Crash des 737 Max : Boeing admet avoir trompé le régulateur européen*, Le Figaro, *737 Max : un ancien pilote d'essai de Boeing inculpé pour fraude*, Les Échos

Pourquoi est-il difficile d'avoir confiance dans l'IA ?¬

La notion de confiance évolue et se décline désormais à des produits et services, matériels et immatériels, issus de l'activité humaine. Ainsi on veut avoir confiance dans une voiture, dans un vaccin, dans une IA, etc. Dans ces cas précis, la confiance s'est essentiellement développée autour de la notion de fiabilité, à savoir le fait que le produit fonctionne sans défaillance le plus longtemps possible. A cela sont venues s'ajouter de nombreuses notions telles que la sûreté ou la capacité à se prémunir d'événements menaçant la sécurité des biens et des personnes. **Dans le monde technologique et numérique, les caractéristiques de la confiance se sont donc multipliées et on attend désormais des produits qu'ils soient valides, robustes, résilients, équitables, compréhensibles, éthiques, fiables, et reproductibles. L'IA n'y échappe pas et comporte une série de risques inhérents, dont la difficulté à rendre compte des résultats des algorithmes, ce qui alimente l'opacité de l'IA** : « On ne sait pas, aujourd'hui, prouver que les conclusions d'un système entraîné par apprentissage sur une base de données sont les bonnes, qu'elles sont robustes à des petites variations, qu'elles ne sont pas entachées de biais, etc.[23] »

Dans la suite de cette section, nous présentons de façon ciblée certaines problématiques. Elles ne doivent pas être comprises comme une liste exhaustive des points à résoudre pour la constitution d'une IA de confiance, mais comme une illustration de la complexité du sujet, et la nécessité d'y parvenir.

L'effet « boîte noire »

« Une grande partie des préoccupations éthiques soulevées par l'IA tiennent à l'opacité de ces technologies.[24] »

L'un des atouts de l'IA contemporaine est sa capacité inédite à traiter des jeux de données de très grande dimension, combinant un nombre extrêmement élevé de variables. Ceci permet de traiter automatiquement de nouveaux types de données tels que l'image numérique, le langage, les structures moléculaires, etc. Mais cette puissance possède son revers : **l'un des problèmes fondamentaux de l'IA réside dans la difficulté à expliquer non seulement son fonctionnement, mais surtout ses résultats et la manière dont elle y parvient. C'est ce que certains désignent sous le terme de « boîte noire ».**

[23] *Les 5 murs de l'IA, la confiance*, Bertrand Braunschweig, Le Monde, 2022
[24] *Donner un sens à l'intelligence artificielle*, Mission parlementaire Villani, 2018

Dans le cas du *deep learning,* qui utilise des réseaux de neurones profonds, il n'y a pas besoin de règles établies à l'avance : la machine construit elle-même un modèle grâce à l'usage statistique des données, comme mentionné précédemment. Plus opaque, le *deep learning* est cependant plus efficace et affiche des performances supérieures aux modèles d'apprentissage plus simples (règles formelles, arbres décisionnels simples, réseaux bayésiens[25]). Même si nous pouvons comprendre les fonctions qui composent le *deep learning*, leur accumulation devient rapidement complexe. **Une « boîte noire » (sans accès aux caractéristiques du réseau) ou une « boîte blanche » (avec un accès aux caractéristiques du réseau) se crée au gré des opérations. Et, dans les deux cas, il s'avère difficile de rendre compte précisément de ce que la machine calcule.**

Pourtant, il est nécessaire d'expliquer les algorithmes d'intelligence artificielle. Winston Maxwell, chercheur en droit à l'Institut des Mines-Télécoms, avance deux raisons. **D'une part, « les individus ont le droit de comprendre et de contester une décision algorithmique[26] »**. Si des systèmes d'IA sont utilisés dans des services (privés et publics), on doit pouvoir expliquer la décision à la personne concernée et, en cas de litige, l'individu et la justice doivent pouvoir comprendre le résultat de l'algorithme pour pouvoir le contester juridiquement. **D'autre part, « il faut garantir qu'une institution de contrôle comme la Commission nationale informatique et libertés (CNIL), ou un tribunal, puisse comprendre le fonctionnement de l'algorithme,** à la fois dans l'ensemble et dans un cas particulier.[27] »

La question des biais algorithmiques

Les comportements cognitifs humains sont loin d'être parfaitement rationnels et sont affectés de dizaines de biais documentés par les sciences cognitives. Les modèles d'IA, qu'ils soient codés symboliquement ou par les données, incorporent également des biais. Certains sont dus à la formulation inadéquate des problèmes par les concepteurs des systèmes d'IA ; d'autres à l'imparfaite disponibilité ou sélection des données d'apprentissage dans les systèmes d'IA statistiques. Par rapport aux biais humains, qui ne sont pas nécessairement uniformément répartis selon les individus et qui peuvent s'équilibrer, se contrebalancer, ou se corriger par l'éducation, **les biais algorithmiques posent un problème spécifique : leur réplicabilité aisée et le risque d'une très large diffusion, et donc la possibilité d'émergence de biais systémiques.**

Le problème des biais algorithmiques est ainsi devenu visible à l'échelle mondiale avec les géants du numérique et l'exemple du ciblage publicitaire de Google (recommandation d'offres d'emploi moins rémunérées aux femmes, reproduction des discriminations déjà présentes dans les données). Il fait courir le risque de voir s'installer une méfiance générale sur l'IA, un préjudice pouvant freiner considérablement son développement, et surtout impacter nos sociétés.

De manière générale, la définition du biais dépend de son contexte. En intelligence artificielle, la notion de biais se réfère à l'idée que chaque cas demande un traitement adapté. En ce sens, le biais est ce qui permet aux systèmes de *machine learning* de juger si une situation est différente d'une autre (« discriminer »), et permet d'adapter le comportement du système. Le biais est donc fondamental pour le processus d'apprentissage et l'adaptation du comportement à une situation particulière[28].

En revanche, au niveau sociétal, le terme de biais se réfère à l'injustice que peut provoquer la différence de traitement. Pour éviter la confusion, l'ISO préfère utiliser le terme d'inéquité (*unfairness*) en IA. Aussi, les biais dans les données d'entraînement sont une source majeure de biais dans les systèmes d'IA. Par ailleurs, les biais cognitifs humains affectent la collecte des données, leur traitement, l'architecture du système, le modèle d'entraînement et d'autres choix de développement.

[25] *Éclairer la boîte noire des algorithmes*, Florence d'Alché-Buc, Winston Maxwell, Antonin Counillon, l'MTech Institut Mines-Télécoms, 22 février 2021
[26] ibid
[27] ibid
[28] ISO-IEC 2022 DIS 22989 Information technology — Artificial intelligence concepts and terminology, p.39

En résumé, certains biais sont essentiels au bon fonctionnement d'un système d'IA, mais des biais indésirables peuvent être introduits involontairement et peuvent mener à des résultats injustes.

Les algorithmes de *deep learning,* qui reposent sur des données massives (*Big Data*) pour personnaliser les contenus et aider à la décision, **font redouter une reproduction technologique des inégalités sociales préexistantes, en outre construites sur une « vision du passé » (les données)**. Or si nous devons vivre au quotidien avec des systèmes d'IA et leur faire confiance, ces systèmes doivent se conformer aux lois et aux normes sociales que nous nous sommes donnés. **Par ailleurs, cette exigence éthique interroge le statut de ces algorithmes dans la société : doivent-ils rester neutres au risque d'amplifier des biais, ou doivent-ils corriger des biais existant dans la société ?** La nécessaire conformité des algorithmes à nos normes ne peut néanmoins surgir uniquement de la volonté : **« cette exigence nécessite le développement de procédures, outils et méthodes permettant d'auditer ces systèmes afin d'en évaluer la conformité à notre cadre juridique.[29] »**

UN CHATBOT DEVIENT NAZI EN QUELQUES HEURES SUR LES RÉSEAUX SOCIAUX

En mars 2016, Microsoft introduit son chatbot « Tay », un agent conversationnel à base d'IA, sur le réseau social Twitter. Tay n'est pas le premier chatbot créé par le géant américain, qui bénéficiait de l'expérience de son premier né « Xiaoice » en Chine, entraîné sur 40 millions d'utilisateurs.

D'après Microsoft, Tay a été modelé sur la personnalité type d'une jeune femme américaine âgée de 18 à 24 ans, et devait proposer une « expérience positive ». Dans les 24 heures suivant sa mise en ligne sur Twitter, Tay a subi une « attaque coordonnée, lancée par quelques individus » pour exploiter sa vulnérabilité. Concrètement, le chatbot est passé de tweets tels que « humans are super cool » à un appel au féminicide, à l'apologie du suprémacisme blanc, au meurtre antisémite et à un « mème » vantant le « cool » d'Hitler.

Le chatbot n'a pas « conscience » de ce dont il parle, et c'est dans cette faille que se sont engouffrées les personnes mises en cause. Précision de taille : les individus en question, qui ont découvert la faille, sont des internautes de 4chan et 8chan, deux forums en ligne désormais connus pour abriter les débats des suprémacistes blancs américains.

En résumé, un agent conversationnel entraîné avec des usagers xénophobes produira un comportement, des outputs, de même nature. S'il est impossible de préserver une IA de la formation des biais qui existent en monde ouvert, l'expérience Tay montre la neutralité de l'agent conversationnel, qui évolue en miroir avec les internautes. L'expérience interroge aussi la possibilité d'assigner des règles préexistantes à des techniques de *machine learning* qui génèrent leurs propres règles.

Sources : *L'IA de Microsoft est-elle réellement devenue raciste au contact des internautes ?*, Erwan Lecomte, 25.3.2016, Sciences et Avenir, *Microsoft explique les raisons qui ont fait de son chatbot Tay un fan d'Hitler, l'entreprise parle d'une attaque coordonnée*, developpez.com, 27.3.2016.

La génération de résultats inéquitables peut être due, notamment, à la nature et la distribution des corpus d'apprentissage lors de la conception. Comme l'illustre l'anecdote du chatbot Tay, ces résultats peuvent être également obtenus dans le cas où le système apprend des données fournies en entrée en conditions d'opération. Certaines bonnes pratiques de conception doivent être mises en œuvre afin de se prémunir, du moins en partie, des biais liés aux données. **Dans la boîte à outils du concepteur, il est donc nécessaire de prévoir une**

[29] *Donner un sens à l'intelligence artificielle*, Mission parlementaire Villani, 2018

estimation de la résilience face à un détournement volontaire, une estimation de robustesse soit le maintien de la performance en situation "normale", **et également la mise en œuvre de mécanismes « garde-fous »** permettant dans certains cas prévisibles de limiter les résultats néfastes (d'où l'importance du contrôle humain[30]). **Ces garde-fous ne sont sans doute pas une solution imparable**, mais la réglementation s'orientant vers une approche par les risques (via l'AI Act, évoqué plus loin dans ce rapport), et une analyse des risques pratiquée de façon systématique pour tout système d'IA permettrait de s'assurer que le système est conçu de façon adaptée au cas d'usage final, qui comporte des risques inhérents.

Le concepteur a une part de responsabilité dans l'absence de biais (si tant est qu'il dispose des outils d'ingénierie et d'inspection adéquats), **mais l'utilisateur doit également utiliser le système en présentant un niveau de maîtrise adapté**, qu'il soit suffisamment informé afin d'utiliser le système de façon correcte, et qu'il dispose également d'une conscience de ses responsabilités en tant qu'utilisateur et membre de la société.

1..2.2. Sur quoi repose la confiance dans l'IA ?

LES INITIATIVES MONDIALES DE DÉFINITION DE L'IA DE CONFIANCE

La confiance est au premier plan des discussions mondiales récentes entre chercheurs, industriels et ONG du domaine. Le momentum commence en 2018 avec la publication de 45 travaux de tous horizons, et au moins 117 documents portant sur les principes d'IA publiés entre 2015 et 2020, dont la majorité par des entreprises[31].

Ces travaux fournis par de grandes institutions, et des industriels, constituent une large base d'études pour formaliser à la fois les concepts et les usages que nous ferons de l'IA. De nombreux groupes de travail ont produit leurs recommandations sur le sujet, en particulier dans des organisations multilatérales. La production de normes pour l'IA fait l'objet d'un travail simultané des plus grandes économies mondiales et des organes de normalisation. De la même manière, la mise à disposition d'éléments pour implémenter l'intelligence artificielle là où elle est jugée utile, est commune parmi les organisations multilatérales. La plupart des initiatives produisent naturellement des travaux de développement concret de l'IA. Enfin l'éthique, sujet couramment associé aux problématiques de l'IA, n'est pas nécessairement l'angle privilégié, de même que la sensibilisation du public à la problématique, que seul le *Partnership* des *Big Tech* (voir ci-dessous) déclare officiellement prendre en charge. Toutefois, le Conseil de l'Europe a engagé depuis quelques années des travaux en vue d'un texte multilatéral fixant des règles relatives à l'IA du point de vue des droits humains.

Le Partenariat sur l'Intelligence Artificielle

Fin 2016, **Amazon, Google, Facebook, IBM et Microsoft annoncent leur alliance au sein d'une organisation *not-for-profit*,** le *Partnership on AI to Benefit People and Society,* dans l'objectif d'éduquer le public sur l'IA et de mettre en commun les recherches et les bonnes pratiques dans ce domaine. Le Partenariat se concentre notamment sur les problématiques éthiques et légales au prisme des défis posés par l'IA. La longue liste des partenaires[32] du Partnership on AI suggère une volonté de mettre à disposition des ressources communes pour avancer dans la direction souhaitée d'une IA acceptable d'un point de vue éthique.

UNESCO study on ethics of AI

En novembre 2021, **193 pays se sont engagés derrière la *Recommandation de l'UNESCO sur l'IA*.** L'UNESCO fonde son approche sur des valeurs « désirables en soi », **déduites des impacts éthiques de l'IA** à partir des axes de respect, de protection de la dignité, des droits humains et des libertés fondamentales. L'approche repose sur 10 principes[33] fondamentaux. **Cette recommandation reste dans le domaine de la soft law** : le respect de ses principes reste volontaire, mais les sanctions sont inexistantes car c'est une régulation non contraignante.

[30] L'un des lignes directrices proposées par le groupe d'expert de haut niveau à la commission européenne
[31] *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report.*Stanford University, p.41
[32] https://partnershiponai.org/partners/
[33] *193 countries adopt first-ever agreement on the Ethics of Artificial Intelligence*, UN News, 25 novembre 2021

The European Commission's HLEG. The High Level Expert Group on AI of the European Commission

Le travail du groupe d'experts européens sur l'IA (AI HLEG) a été décisif dans l'approche choisie par la Commission pour aborder l'IA. Les recommandations du groupe ont servi de canevas aux initiatives législatives de la Commission et des États membres, parmi lesquelles la Communication *on Building Trust in Human Centric Artificial Intelligence, le White Paper on Artificial Intelligence: a European approach to excellence and trust* et le *Coordinated plan on AI*[34].

The OECD's Expert Group and Observatory. OECD created an AI Group of Expert (AIGO)

L'Observatoire des politiques de l'intelligence artificielle de l'OCDE se forme à la suite de la publication de la *Recommandation on Artificial Intelligence* (*OECD AI Principles*[35]), première ébauche de standard international sur l'IA, adopté en mai 2019 par certains pays membres. Ces principes ont servi de base à l'adoption des G20 Principles[36] en juin 2019. À la longue liste des États membres participants[37] s'ajoutent des centres de recherche du MIT, Harvard ou l'Inria pour la France, mais aussi les géants IBM, Microsoft, Google/DeepMind ou Facebook.

Le Global Partnership on Artificial Intelligence (GPAI)

Le *Global Partnership on Artificial Intelligence* (Partenariat Mondial sur l'IA) est une initiative aux multiples parties prenantes lancée à la suite du G7 de juin 2020 par la France et le Canada. Le GPAI vise à approfondir la Recommandation de l'OCDE publiée en mai 2019 en soutenant la recherche fondamentale en IA, et ses applications industrielles et commerciales. Le Partenariat Mondial s'organise autour de plusieurs groupes de travail thématiques, dont le *Responsible AI Working Group*[38]. Le groupe de travail répond à un mandat connexe à la mission générale du GPAI, à savoir **promouvoir et contribuer au développement, à l'usage et à une gouvernance responsable de systèmes d'IA centrés sur l'humain, en cohérence avec les Objectifs de Développement Durable.** Le GPAI compte aujourd'hui 25 pays membres.

La Global AI Action Alliance (GAIA) du World Economic Forum (WEF)

En janvier 2021, le Forum Économique Mondial (WEF) lance la Global AI Action Alliance (GAIA) afin d'accélérer le déploiement d'une intelligence artificielle inclusive, transparente et de confiance. Cette coopération rassemble plus de **100 parties prenantes composées d'entreprises, d'organisations internationales**, d'ONG et d'universitaires œuvrant pour la **maximisation des bénéfices imputables à l'IA et la minimisation de ses risques**[39].

Les organismes de normalisation ISO IEC IEEE ITU CEN-CENELEC SAE

Le monde de la normalisation s'est largement saisi du sujet de l'IA. Au niveau international, **l'ISO et l'IEC ont créé un comité commun sur l'IA (le SC 42), qui développe de façon offensive des normes « horizontales » sur l'IA**, allant de la terminologie à la gouvernance, au management de l'IA par les organisations, et comprenant les spécifications, de l'IA de confiance, l'ingénierie, l'interopérabilité, etc. Au niveau européen, le CEN et le CENELEC ont formé en 2021 le Joint *Technical Committee 21 "Artificial Intelligence* » (JTC 21). **La première ambition du JTC 21 est d'accompagner la régulation européenne de l'IA par le développement de normes « harmonisées[40] »**. Les futures normes européennes vont donc fortement modeler le marché européen, voire mondial. De son côté, l'IEEE *(Institute of Electrical and Electronics Engineers)* a lancé la *Global Initiative on Ethically Aligned design of AI Systems*, découlant de la publication d'*Ethically Aligned Design*[41], une analyse scientifique de principes de haut niveau et de recommandations actionnables.

[34] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai
[35] OECD AI Principles overview, https://oecd.ai/en/ai-principles
[36] https://oecd.ai/en/list-of-participants-oecd-expert-group-on-ai
[37] Responsible AI Working Group Report, GPAI — Montreal Summit 2020, executive summary
[38] *World Economic Forum launches Global artificial intelligence alliance*, Hindustan Times, 28 janvier 2021
[39] *World Economic Forum launches Global artificial intelligence alliance*, Hindustan Times, 28 janvier 2021
[40] Une norme harmonisée est une norme publiée au journal officiel de l'Union Européenne et liée à un acte législatif européen. Cette norme s'applique à l'ensemble des pays de la zone UE. La conformité d'un produit ou d'un service à une norme harmonisée constitue une présomption de conformité à la réglementation européenne.
[41] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf

Plus de 200 groupes de travail identifiés à travers le monde

Partie prenante du GPAI, **le think-tank Future Society recense 200 groupes de travail et initiatives sur l'intelligence artificielle**[42], tous attachés à l'idée d'une IA centrée sur l'humain et son bien-être. Le champ de l'intelligence artificielle "Responsable" telle que définie par le mandat du groupe de travail Future Society est vaste, ce qui s'illustre dans l'écosystème pléthorique d'initiatives cherchant à fournir des orientations sur la manière dont l'intelligence artificielle devrait se développer et être adoptée, ou sur la manière de l'utiliser pour mettre en œuvre l'agenda *AI for Social Good*[43] au service des Objectifs de Développement Durable[44]. Ces programmes **portés par le monde universitaire, les secteurs public et privé, la société civile et les ONG** ont pour point commun de dessiner les contours d'une intelligence artificielle responsable à travers des mécanismes formels ou informels.

Les attributs de la confiance dans l'IA

Lorsqu'il s'agit de définir ce qu'implique la confiance dans l'IA, la principale difficulté est de comprendre sur quoi repose cette confiance. Et donc quels en sont les attributs ? Un grand nombre de documents (normes, guides techniques, livres blancs, articles scientifiques) traite de la question, mais **nous notons plusieurs limitations dans la littérature :**

- **Absence d'une liste exhaustive et pertinente des attributs de confiance.** Responsabilité, gouvernance, sécurité, résistance aux attaques, absence de biais, etc. Les attributs sont nombreux, chaque document défend des objectifs potentiellement différents, et s'adresse aussi à des publics différents. Globalement, nous pouvons estimer que les normes et guides techniques sont globalement plutôt destinés aux concepteurs et fournisseurs, quand les articles scientifiques s'adressent aux chercheurs. Il est ainsi difficile d'avoir une vue d'ensemble cohérente des attributs de confiance pour chaque acteur de l'écosystème de l'IA de confiance, et pour chaque étape du cycle de vie du système (de la conception jusqu'à l'utilisation). Certains attributs ne seront pas pertinents selon le cas d'application, le type de modèle, ou l'acteur impliqué. Il est nécessaire de disposer d'attributs de confiance pertinents permettant de garantir le niveau de confiance de chaque acteur.

- **Absence de hiérarchie des attributs**. Selon les textes, les attributs sont présentés à différents niveaux de granularité, ce qui constitue un frein majeur pour leur adoption (tant par les concepteurs d'IA, que par des inspecteurs de la conformité du système). Par exemple, certains textes énoncent la notion d'intégrité comme un attribut de la confiance ; toutefois l'intégrité ne constitue pas un attribut en soi, car elle est composée d'un nombre d'éléments (complétude, exactitude, etc.), qu'il convient d'analyser séparément afin d'estimer l'intégrité du système. Ainsi, de nombreux attributs sont de « trop haut niveau » pour être opérationnels, tels que la gouvernance, la responsabilité, etc., qui nécessitent une décomposition en sous-éléments réellement observables, quantifiables.

- **Des attributs trop peu « outillés »**. Les concepteurs doivent disposer de guides leur permettant de comprendre comment développer des solutions de confiance, et les inspecteurs doivent connaître les points d'observation pour l'évaluation de la conformité. Les considérations générales, ou les attributs de trop haut niveau ne permettent pas l'obtention de consensus. Sans pour autant brider l'innovation en fixant des obligations de moyens, il est nécessaire que

[42] *Areas for Future Action in the Responsible AI Ecosystem, The Future Society*, GPAI and CEIMIA, décembre 2020, p.5
[43] L'AI for Social Good est une initiative conjointe du Computing Community Consortium, de l'Office of Science and Technology Policy (OSTP) de la Maison Blanche et de l'Association for the Advancement of Artificial Intelligence (AAAI), sous la forme de bourse académique ayant donné lieu à 5 ateliers sur le sol américain. *Artificial Intelligence for Social Good,* Gregory D. Hager, Ann Drobnis, Fei Fang, Rayid Ghani, Amy Greenwald, Terah Lyons, David C. Parkes, Jason Schultz, Suchi Saria, Stephen F. Smith et Millind Tambe, Mars 2017
[44] *Areas for Future Action in the Responsible AI Ecosystem, The Future Society*, GPAI and CEIMIA, décembre 2020, p.12

les objectifs à atteindre soit décomposés au niveau de granularité minimum pour permettre, par exemple, au concepteur de comprendre comment rendre son système transparent dans son propre cas d'usage, et à l'inspecteur de disposer de points de mesure clairs et adaptés. Il ne s'agit pas de fixer des obligations de moyens qui brideraient l'innovation, ou d'établir des seuils minimums d'acceptabilité qui ne s'adapteraient peut-être pas à tous les cas, mais de fournir des ensembles d'outils d'ingénierie et d'inspection.

S'appuyant sur différentes sources (normes, standards, etc.), les équipes académiques et industrielles travaillant dans le cadre du Grand Défi National « IA de confiance » (Grand Défi National « *sécuriser, certifier et fiabiliser les systèmes fondés sur l'IA* » issu de la Stratégie Nationale en IA du plan France 2030 - SGPI[45]), **proposent d'extraire quelques attributs jugés prioritaires pour l'IA de confiance**. Cette sélection d'attributs prioritaires semble faire raisonnablement consensus dans la communauté, mais montre l'hétérogénéité des points de vérifications nécessaires, qui concernent la responsabilité de différentes parties prenantes de l'écosystème IA, et dont les modes de vérification varient entre estimation quantifiée (performance, etc.) et observation d'expert. En outre, certains attributs sont des critères généralistes, incluant eux-mêmes des listes d'attributs qui ne sont pas toujours définis précisément et de façon mesurable par l'industrie IA.

Dans ce contexte, le programme Confiance.ai, pilier technologique du Grand Défi National « IA de confiance », ambitionne de développer un environnement pour le développement d'IA de confiance. Il travaille donc, parmi bien d'autres sujets techniques, sur la définition, la structuration et les métriques des attributs de confiance dans le cadre du déploiement de l'IA dans les systèmes critiques. Ses attributs sont aujourd'hui regroupés en fonction des capacités qu'ils caractérisent : la technique, l'interaction, l'éthique et les intermédiaires de la confiance (comme la certification). En voici quelques exemples :

Attributs techniques :

- **Fiabilité (*Reliability*)** : Propriété relative à un comportement et à des résultats prévus et cohérents. Selon le contexte ou le secteur, et également selon le produit ou le service spécifique, les données et la technologie utilisées, différentes caractéristiques s'appliquent et doivent être vérifiées pour s'assurer que les attentes des parties prenantes sont satisfaites. Les caractéristiques de la fiabilité comprennent la disponibilité, la résilience, la sécurité, la confidentialité, la sûreté, la responsabilité, la transparence, l'intégrité, l'authenticité, la qualité, et la facilité d'utilisation. La fiabilité est un attribut qui peut être appliqué aux services, produits, technologies, données et informations et, dans le contexte de la gouvernance, aux organisations.

  Sources : ISO-27000, ISO 5723BSI, ISO 24028 (3.42), HLEG 2019 ALTAI.

- **Fiabilité (*Dependability*) :** Capacité de fournir un service auquel on peut faire confiance. Cela implique : disponibilité pour un service correct ; continuité d'un service correct ; absence de conséquences catastrophiques sur le ou les utilisateurs et l'environnement (sûreté) ; absence de divulgation non autorisée d'informations (confidentialité) ; absence d'altérations inappropriées du système (intégrité) ; capacité à subir des modifications et des réparations (maintenabilité) ; existence simultanée de la disponibilité pour les utilisateurs autorisés uniquement, de la confidentialité et de l'intégrité (sécurité).

- **Conformité (*Compliance*)** : Démonstration qu'une caractéristique ou une propriété d'un produit satisfait aux exigences énoncées.

  Source : CENELEC-EN50126

[45] Secrétariat Général Pour l'Investissement

LA CONFIANCE EST
UNE ARME DÉFENSIVE
CONTRE LES MONOPOLES
QUE L'ON SUBIT,
ET OFFENSIVE POUR
LES COOPÉRATIONS
QUE L'ON CHOISIT.

- **Traçabilité (*Traceability*)** : 1. Degré auquel une relation peut être établie entre deux ou plusieurs produits du processus de développement, en particulier les produits ayant un prédécesseur, un successeur ou une relation maître-subordonné les uns par rapport aux autres ; par exemple, le degré auquel les exigences et la conception d'un composant logiciel donné correspondent. 2. Degré auquel chaque élément d'un produit de développement logiciel établit sa raison d'être ; par exemple, le degré auquel chaque élément d'un diagramme à bulles fait référence à l'exigence qu'il satisfait.

  Sources : IEEE-610.12-1990, Adapted from IEEE glossary of Software Engineering Terminology.

- **Précision (*Accuracy*)** : 1. Une évaluation qualitative de la justesse, ou de l'absence d'erreur. 2. Une mesure quantitative de l'ampleur de l'erreur. 3. Dans le cadre du système de gestion de la qualité, l'exactitude est une évaluation de la justesse.

  Sources : ISO/IEC/IEEE 24765:2010.

- **Qualité des données (*Data Quality*)** : 1. Mesure dans laquelle les caractéristiques des données répondent aux besoins exprimés et implicites lorsqu'elles sont utilisées dans des conditions spécifiées. 2. Mesure dans laquelle les données sont exemptes de défauts et possèdent les caractéristiques souhaitées.

  Sources : ISO-25024:2015, Projet DEEL

- **Sûreté (*Safety*)** : 1. Attente selon laquelle un système ne conduit pas, dans des conditions définies, à un état dans lequel la vie humaine, la santé, les biens ou l'environnement sont mis en danger. 2. Capacité d'avoir des niveaux de risque acceptables en ce qui concerne les dommages causés aux personnes, aux entreprises, aux logiciels, aux biens ou à l'environnement. 3. Absence de risque inacceptable de préjudice, c'est-à-dire de blessures humaines ou de mort. Absence de risque non tolérable.

  Sources : ISO/IEC/IEEE 12207:2017, ISO 9126-1:2001, EN 50129

- **Robustesse (*Robustness*)** : La robustesse est une caractéristique importante pour assurer la confiance des utilisateurs car elle permet de maintenir les niveaux de performance des systèmes d'IA dans des conditions d'utilisation très différentes. Elle se définit comme 1. la capacité d'un système à maintenir son niveau de performance dans diverses circonstances, 2. (Gobal) la capacité du système à réaliser la fonction prévue en présence d'entrées anormales ou inconnues / (Local) la mesure dans laquelle le système fournit des réponses équivalentes pour des entrées similaires.

  Sources : ISO-22989:2021, projet DEEL

- **Justesse (*Correctness*) :** 1. Le degré auquel un système ou un composant est exempt de fautes dans sa spécification, sa conception et son implémentation. 2. Le degré auquel le logiciel, la documentation ou d'autres éléments répondent aux exigences spécifiées. 3. Le degré auquel le logiciel, la documentation ou d'autres éléments répondent aux besoins et aux attentes des utilisateurs, qu'ils soient spécifiés ou non.

  Source : ISO-24765:2017

- **Maintenabilité (*Maintainability*)** : Capacité à étendre/améliorer un système donné tout en maintenant sa conformité aux exigences inchangées.

  Source : projet DEEL

- **Vérifiabilité (*Verifiability*)** : Capacité à évaluer une mise en œuvre des exigences afin de déterminer qu'elles ont été satisfaites.

  Source : ARP4754A

- etc.

Attributs interactions

- **Explicabilité (*Explainability*)** : Mesure dans laquelle le comportement d'un modèle peut être rendu compréhensible pour les humains. Propriété d'un système d'IA capable de présenter les facteurs importants qui influencent les résultats du système d'IA, de manière compréhensible par un humain. Explication : les systèmes fournissent des preuves ou des raisons pour tous les résultats. Significatif : Les systèmes fournissent des explications qui sont compréhensibles pour les utilisateurs individuels. Exactitude de l'explication : l'explication reflète correctement le processus utilisé par le système pour générer les résultats. Limites des connaissances : le système ne fonctionne que dans les conditions pour lesquelles il a été conçu ou lorsque le système atteint un niveau de confiance suffisant dans ses résultats.

  Sources : hal-03176080, ISO22989, NISTIR8312.

- **Transparence (*Transparency*)** : 1. Présentation ouverte, complète, accessible, claire et compréhensible des informations. 2. Propriété d'un système ou d'un processus à impliquer l'ouverture et la responsabilité. 3. Propriété d'une organisation selon laquelle les activités et les décisions appropriées sont communiquées aux parties prenantes concernées de manière complète, accessible et compréhensible. 4. Propriété d'un système selon laquelle les informations appropriées sur le système sont communiquées aux parties prenantes concernées.

  Sources : ISO 16759:2013, ISO 27036-3, ISO 22989

- **Responsabilité (*Accountability*)** : Fait de répondre de ses actions, décisions et performances.

  Source : ISO-24028:2020

- **Supervision et contrôle (*Oversight and control*)** : Action délibérée ou processus en vue d'atteindre des objectifs spécifiques, et surveillance régulière de cette action.

  Source : ISO-IECJTC1-SC42-WG1-N1298

- **Utilisabilité (*Usability*)** : Degré selon lequel un produit ou un système peut être utilisé par des utilisateurs spécifiés pour atteindre des objectifs spécifiés avec efficacité, efficience et satisfaction dans un contexte d'utilisation spécifié.

  Source : ISO/IEC 25010

- etc.

Attributs éthiques

- **Équité (*Fairness*) :** 1. Traitement ou comportement impartial et juste, sans favoritisme ni discrimination. 2. L'équité fait référence à une variété d'idées connues sous le nom d'équité, d'impartialité, d'égalitarisme, de non-discrimination et de justice. L'équité incarne un idéal d'égalité de traitement entre des individus ou entre groupes d'individus. C'est ce que l'on appelle généralement l'équité "substantive". Mais l'équité englobe également une perspective procédurale, à savoir la capacité de demander et d'obtenir réparation lorsque les droits et libertés individuels sont violés.

  Sources : Oxford English Dictionary, HLEG 2019 ALTAI.

- **Vie privée (*Privacy*)** : Garantie de préservation de la vie privée d'un individu lorsque celui-ci subit une intrusion ou une collecte illégale de ses données.

  Sources : ISO/IEC 2382:2015, 21262637

- **Diversité et inclusion (*Diversity and inclusion*)** : Attention portée à la représentativité des données d'entraînement en prenant en compte la diversité des profils (ex : représentativité ethnique, parité, âge, religion etc.).

- **Techniques subliminales** : Les techniques subliminales sont les stimuli (images ou sons) incorporés dans un objet ou un message pour être perçus à un niveau non-conscient, dans le but d'influer sur le comportement des individus. Ces techniques ont été utilisées dans la publicité ou la propagande. L'efficacité de ces techniques et leurs objectifs sont sujets à débat. Dans le cadre de l'IA, l'usage de ces techniques serait interdit au titre de l'article 5 du projet de règlement sur l'IA (AI Act[46]).

  Sources : ISO/IEC 2382:2015, 21262637

- etc.

Attributs pour intermédiaires d'écosystème de confiance

- **Assurance qualité :** Ensemble des activités, tout au long du cycle de vie du projet, nécessaires pour donner une confiance suffisante dans le fait qu'un produit ou un service est conforme aux exigences des parties prenantes ou qu'un processus respecte la méthodologie établie. Les synonymes potentiels sont : assurance, assurance produit, assurance développement, assurance conception.

  Source : Daniels, S. E., Johnson, K., & Johnson, C. (2002). Quality glossary. Quality Progress, 35(7), 43.

- **Audit:** 1. Un examen systématique et indépendant visant à déterminer si les procédures spécifiques aux exigences d'un produit sont conformes aux dispositions prévues, sont mises en œuvre efficacement et permettent d'atteindre les objectifs spécifiés. 2. Processus méthodique, indépendant et documenté permettant d'obtenir des preuves d'audit et de les évaluer de manière objective pour déterminer dans quelle mesure les critères d'audit sont satisfaits. Un audit peut être interne (audit de première partie), externe (audit de seconde ou de tierce partie), ou combiné (associant deux disciplines ou plus).

  Sources : CENELEC-EN50126, ISO-27000:2018

- **Certification :** 1. Attestation d'une tierce partie concernant des produits, des processus, des systèmes ou des personnes. 2. Une garantie écrite qu'un système ou un composant est conforme à ses exigences spécifiées et est acceptable pour une utilisation opérationnelle. 3. Démonstration formelle qu'un système ou un composant est conforme aux exigences spécifiées et est acceptable pour une utilisation opérationnelle. 4. Processus de confirmation qu'un système ou un composant est conforme à ses exigences spécifiées et qu'il est acceptable pour une utilisation opérationnelle.

  Source : ISO-24765:2017

- **Conformité à la réglementation (*Lawfulness & Compliance*)** : Démonstration qu'une caractéristique ou une propriété d'un produit satisfait aux exigences énoncées par la réglementation. Capacité à faire appliquer la réglementation en vigueur.

  Sources : EN 50126.

- etc.

Le facteur humain pour garantir la confiance

L'humain joue un rôle majeur dans l'écosystème de confiance. Ainsi que le note la norme ISO/IEC TR 24028 *Overview of trustworthiness in artificial intelligence*, il est nécessaire de

[46] https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:52021PC0206&from=FR

spécifier qui a confiance en qui, et dans quels aspects du développement et de l'utilisation de l'IA[47]. Les individus impliqués doivent avoir confiance dans l'IA et dans les autres acteurs de l'écosystème, mais également inspirer confiance. **L'identification et la définition de tous les acteurs humains est une première étape en vue de la constitution de l'écosystème.**

Parmi ceux-ci, nous pouvons déjà lister :

- **L'utilisateur (l'opérateur)** : Ses actions contribuent à ce que l'algorithme d'IA produise un résultat. Ce résultat devra être correct et en adéquation avec des attentes de performance (un "bon" résultat), mais il doit également être en phase avec les valeurs de l'utilisateur (valeurs morales, déontologiques, etc.).
- **L'individu impacté par la prise de décision IA** : Il subit les conséquences de la prise de décision sans avoir toujours la possibilité de peser sur celle-ci. Il est intéressé par son propre cas et souhaite être traité de façon morale, équitable, comprendre les facteurs qui ont joué en sa faveur ou défaveur.
- **Le concepteur** : Il souhaite concevoir et implémenter un algorithme le plus performant, réplicable et robuste et cherchera une vision quasi mathématique de la confiance.
- **L'auditeur /certificateur** : Il a pour mission de contrôler le système et identifier les dysfonctionnements ou anomalies afin de garantir un bon niveau de performance uniquement associé à des standards ou à la législation.
- etc.

Le rapport *Lignes directrices en matière d'éthique pour une IA digne de confiance*[48] rédigé par le groupe d'experts indépendants de haut niveau pour l'intelligence artificielle (*High-level Expert Group on Artificial Intelligence* - HLEG AI) mandaté en 2018 par la Commission, identifie également d'autres parties prenantes, telles que les prestataires ou la société au sens large.

Pour assurer la solidité de l'écosystème, il est nécessaire que chacune des parties prenantes soit identifiée et formée à l'IA de confiance. Ceci peut paraître évident pour les individus agissant dans un contexte professionnel, tels que les concepteurs ou les auditeurs qui pourront acquérir ces notions dans le cadre de leur formation. Il est essentiel toutefois que chaque maillon de l'écosystème ait connaissance de la façon dont il peut inspirer la confiance et contribuer à la réussite de l'ensemble. Dans le cas de l'utilisateur du système, il est essentiel que celui-ci soit formé à la bonne utilisation et qu'il soit informé des caractéristiques de l'IA de façon appropriée.

La personne impactée par la prise de décision du système doit également être consciente du fait que le traitement est effectué par une IA, et doit connaître ses possibilités de recours en cas de contestation du résultat. Sur ces points, l'AI Act propose notamment que les systèmes d'IA à haut risque suivent la procédure réglementaire de marquage CE[49], qui inclut la mise à disposition d'une notice d'utilisation. Le texte suggère également que dans certains cas où les résultats de l'IA pourraient être potentiellement préjudiciables pour l'individu, celui-ci soit informé qu'il interagit avec une IA. En outre, le Parlement européen a émis une recommandation en 2020 concernant la responsabilité civile pour l'IA[50], demandant la reconnaissance de la responsabilité de l'opérateur du système d'IA, et notant l'importance d'identifier tous les maillons de la chaîne de responsabilité (utilisateur, concepteur, etc.), et de proposer des moyens de

[47] *It is insufficient to simply refer to the 'trustworthy AI', but to specify who trusts whom in what aspects of AI development and use*. ISO/IEC TR 24028.
[48] Lignes directrices en matière d'éthique pour une IA digne de confiance. HLEG IA, 2019.
[49] Le marquage CE (*Conformité Européenne*) a été créé dans le cadre de l'harmonisation des législations techniques européennes. Le marquage CE est un marquage réglementaire indiquant que le fabricant engage sa responsabilité sur la conformité du produit à l'ensemble des exigences fixées par la législation de l'Union européenne applicable à ce produit. - Wikipedia
[50] Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission sur un régime de responsabilité civile pour l'intelligence artificielle, 2020/2014(INL).

recours en cas de préjudice. Le texte met ainsi en exergue le rôle de l'utilisateur dans la bonne utilisation du système : comme pour tout produit issu de l'industrie, l'utilisateur d'un système d'IA a une part de responsabilité dans la bonne ou mauvaise utilisation du produit. En 2020, le Parlement européen a proposé un projet de rapport[51] mettant l'accent sur la transmission des savoirs relatifs à l'IA et notamment sur l'importance d'éduquer les populations sur ce que sont les IA à haut risque et sur la façon de les utiliser.

Il est nécessaire enfin de prendre également en compte la notion de culture sectorielle : si certains champs disciplinaires sont habitués à l'usage de traitements statistiques, ou d'outils informatisés, tels que la production industrielle et le domaine médical, l'appropriation de solutions d'IA peut représenter une barrière à franchir dans un grand nombre de secteurs.

La formalisation des acteurs de l'écosystème de l'IA de confiance doit donc nécessiter, d'une part, l'identification de tous les maillons et à quel endroit du cycle de vie de l'IA ils interviennent, et également la détermination des attentes de chacun des acteurs (performance, respect des valeurs, respect de la réglementation, etc.), ainsi que de fixer des prérequis pour que chaque acteur puisse contribuer efficacement à la confiance (formation, information, etc.).

1.2.3. La difficulté à auditer et certifier des systèmes à base d'IA

La difficulté à auditer l'IA

Si l'on se place du point de vue d'un auditeur externe à l'organisation qui a produit l'algorithme ou l'assemblage d'algorithmes donnant lieu à une décision (comme l'accord d'un crédit, le classement d'un employé, le prix d'une scie sauteuse en période de soldes, le rang d'une recommandation de vidéo ou de partenaires de rencontre), la tâche d'audit s'avère complexe et ceci pour plusieurs raisons.

L'approche par pure transparence du code (comme pour Parcoursup[52]), ou du modèle généré, n'a pas grand sens en IA, les algorithmes de Google DeepMind rivalisant avec les meilleurs outils de traitement automatique de la langue ont par exemple plus de 280 milliards de paramètres à ce jour. Et même des outils de plus vieille génération, comme les algorithmes de *yield management (pricing dynamique)* des compagnies aériennes, mobilisent des couches de technologies et des modèles linéaires et stochastiques[53] avec plusieurs millions d'équations. **Les auteurs eux-mêmes de ces codes fonctionnent souvent en silo** : une équipe de développeurs se concentrant sur un aspect de l'algorithme, avec une métrique de performance locale.

Ils laissent parfois les souches algorithmiques s'auto-sélectionner à la performance (A/B testing dans un contexte de systèmes de recommandation par exemple[54]) ; les outils de génération de modèles de *machine learning*, tels que ceux proposés par des plateformes comme *Dataiku*[55] ou *DataRobot*[56] aident à la **production de code de moins en moins manuelle.** Ces évolutions dans le développement logiciel rendent complexe la lisibilité du code produit et **les auteurs du code eux-mêmes ont du mal à procéder à des audits en profondeur.**

Peut-on alors auditer « en boîte noire » un algorithme, sans connaître ni sa technologie ni ses « règles », implicites ou explicites ? :

- **Légalement et techniquement**, il faut d'abord pouvoir accéder à l'algorithme dans ses conditions d'usage réelles. Sortir l'algorithme de la plateforme pour le mettre dans un bocal et l'observer « à

[51] *Projet de rapport du Parlement européen intitulé « L'intelligence artificielle dans les domaines de l'éducation, de la culture et de l'audiovisuel*, 2020/2017(INI).
[52] Parcoursup est une plateforme Web destinée à recueillir et gérer les vœux d'affectation des futurs étudiants de l'enseignement supérieur français. En 2020, elle gère près de 660 000 étudiants et 15 500 formations. - Wikipedia
[53] Qui se rapporte à l'étude des phénomènes aléatoires dépendant du temps.
[54] Using A/B testing to measure the efficacy of recommendations generated by Amazon Personalize
[55] Société d'origine française spécialisée dans l'analyse des données et le développement des méthodes prédictives en environnement Big Data
[56] Plateforme de gestion centralisée pour optimiser les résultats d'entreprises à l'aide de la puissance de l'IA

froid » ne permet que des tests partiels. Auditer en conditions réelles peut poser des problèmes si l'acteur audité ne facilite pas l'accès. Les évolutions du cadre réglementaire européen (DSA[57] et DMA[58], entre autres) semblent s'orienter dans le sens d'une plus forte auditabilité et donc d'une participation volontaire et obligatoire de celui-ci en face de l'auditeur.

- **Contradictoirement**, il faut que la technique de sonde employée par l'auditeur ne soit pas détectable par l'algorithme à auditer. En effet, si le comportement est robotique, ou ne serait-ce que trop atypique, l'algorithme pourra se sur-adapter à la requête et donc biaiser son comportement, voire le débiaiser. On trouve des exemples de ce type d'adaptation dans le transport ou le e-commerce où les tarifs fournis aux robots de *scraping*[59] ne sont pas systématiquement les mêmes que les tarifs fournis au grand public, leur probabilité d'achat étant fortement plus faible. Créer des cas d'usage de synthèse, indétectables, est un sujet pointu en data science[60], d'autant plus complexe que le nombre de dimensions qui caractérisent l'usage est élevé.

- **Mathématiquement**, détecter un biais ou une déloyauté revient à explorer un très large espace de requêtes possibles pour l'algorithme (toutes les configurations historiques client, tout le catalogue de produits, toutes les conditions d'usage influençant potentiellement la réponse algorithmique). L'objectif est de trouver des circonstances dans lesquelles, le biais ou la tromperie, est « flagrant ». Ces circonstances devront à la fois être représentatives des usages de la plateforme et induire un préjudice significatif. L'audit d'algorithmes en « boîte noire » pour la détection de biais ou de fraudes[61] devient un axe de recherche qui prend sa place dans les conférences d'IA.

- **À valeur probante**, même si la détection isole des zones de flagrant délit significatives (dans le cadre du test opéré), il faut encore que la méthode d'échantillonnage soit reproductible, donc statistiquement probante auprès de l'auditeur. Les propriétés statistiques vérifiées lors des tests devront alors s'adapter aux standards de preuve des autorités de régulation. Par exemple, dans un contexte de collusion, comment prouver que deux algorithmes s'envoient des signaux pour s'indiquer l'un à l'autre de renoncer à une stratégie tarifaire dans un marché donné ?

L'évolution du cadre réglementaire, la pression politique et médiatique et les abus répétés de position dominante des acteurs *Big Tech*, font penser que les circonstances sont réunies pour que naisse un « écosystème de l'audit algorithmique ». Cet écosystème « RegTech » devrait ainsi pouvoir émerger, et l'Europe a une carte certaine à jouer en usant de sa sensibilité, sa culture informatique et mathématique, et ses laboratoires de recherche déjà positionnés autour de ces sujets, même si l'Institut le plus avancé sur ce sujet paraît être *AI Now* de Kathy Crawford à New York University.

La difficulté à certifier l'IA

Fondamentalement, nous distinguons certification réglementaire et certification volontaire. Dans le premier cas, le cadre réglementaire impose que le produit soit certifié avant d'être mis sur le marché, tandis que dans le second cas la certification a lieu en fonction des choix de l'entité qui met le produit sur le marché (avantage compétitif, reconnaissance d'une qualité demandée par un client, etc.). Dans tous les cas, **la certification repose sur un ensemble d'exigences techniques et de moyens de vérifier le respect de ces exigences. La certification est toujours opérée par un organisme tiers.**

Pour un système à base d'IA, il s'agit donc de vérifier que celle-ci répond à un ensemble de critères fixés. La certification peut concerner le produit d'IA en lui-même (l'algorithme, la solution IA), mais également les processus associés (développement, entraînement, etc.), une personne (le développeur, le *data scientist,* etc.) ou un système (système de management, etc.).

[57] Digital Services Act
[58] Digital Market Act
[59] Technique d'extraction automatique de contenus (structurés le plus souvent sur un ou plusieurs sites web)
[60] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3634235
[61] https://www.cs.bu.edu/faculty/crovella/paper-archive/minimization-audit-Neurips21.pdf

CERTIFICATION DE PROCESSUS POUR L'IA – LNE

En 2021, le laboratoire national de métrologie et d'essais (LNE) a créé une certification volontaire des processus de conception, développement, évaluation et maintien en conditions opérationnelles des solutions d'IA par apprentissage. Fruit d'un consensus entre développeurs, évaluateurs et utilisateurs finaux d'algorithmes, cette certification permet aux développeurs et intégrateurs de prouver le respect des exigences de performance, réglementation, confidentialité et éthique de leurs clients.

La vérification est réalisée via un audit, pour une certification valable un an. Par exemple, l'inspection vérifie la qualité et la pertinence des stratégies mises en œuvre pour l'apprentissage, la qualité des données, les capacités du demandeur à documenter convenablement ses processus, la bonne identification des personnes clés impliquées dans les processus, ou encore la qualité de l'information au client.

Source : *Certification de processus pour l'IA*, Laboratoire National de Métrologie et d'Essais

Pour définir de façon appropriée une exigence relative, par exemple l'explicabilité de l'IA, il faut définir techniquement ce qu'est l'explicabilité et comment la mesurer (et donc les métriques associées). De même pour la robustesse, la performance, la sécurité, ou l'équité. Comme évoqué précédemment, **les critères auxquels doit répondre une intelligence artificielle de confiance ne sont pas encore fixés de façon pertinente et exhaustive, et les outils de développement et d'inspection n'existent pas dans tous les cas.** Chaque attribut de confiance doit être identifié, être associé à une métrique pour permettre son inspection, et éventuellement proposer des seuils de conformité pour chacun de ces attributs. **À l'heure actuelle, certifier qu'une IA est de confiance est possible, mais cette certification ne s'appuierait que sur un ensemble limité de critères, ceux pour lesquels nous disposons de suffisamment de connaissances.**

Certains attributs de confiance sont relatifs à la performance du système. Si l'on parle de certification réglementaire, c'est-à-dire imposée par la loi, il ne s'agira pas de déterminer si le système est efficace ou non pour réaliser sa tâche, mais plutôt si les erreurs et anomalies ne viennent pas en infraction à la réglementation. **Plusieurs stratégies sont alors possibles pour démontrer une performance : les preuves formelles ou les statistiques.** Chacune de ces méthodes présente des avantages et des inconvénients qu'il convient de bien maîtriser avant de choisir une stratégie de certification d'un produit.

Les méthodes par preuves formelles, utilisées en algorithmie classique, consistent à formaliser le système au sein d'un modèle présentant tous les comportements possibles du système. L'inspection consiste à déterminer des ensembles de propriétés et à les vérifier sur le modèle. En IA, **s'il existe des travaux bien avancés pour l'évaluation par preuves formelles,** comme par exemple pour l'évaluation de robustesse sur les réseaux de neurones (ISO/IEC TR 24029-1 « *Assessment of the robustness of neural network* », et ISO/IEC TR 24029-2 « *Part 2: Methodology for the use of formal methods* » en cours de développement), la mise au point de méthodes pour d'autres métriques de performance et pour d'autres types d'algorithmes d'IA **constitue encore un champ de recherche.**

L'évaluation par la réalisation d'essais consiste à placer le système dans un environnement type, et observer ses comportements face à des stimuli pertinents. Pour un système d'IA, il peut s'agir de fournir des ensembles de données d'entrée, ou de recourir à un simulateur. Dans le cas d'un essai sur un échantillon de données (ou scénarios), **les problématiques associées peuvent être la disponibilité des données, ou encore de s'assurer que la base de**

données est parfaitement représentative des capacités du système (qu'elle entre bien dans son domaine de fonctionnement). En outre, il est nécessaire également de s'assurer que la base de données contient des entrées « non souhaitées », c'est-à-dire qui auront été identifiées à l'analyse de risques du système comme étant des entrées possibles mais pouvant générer un comportement préjudiciable. D'autres questions se posent, notamment du point de vue de la distribution des bases de données et la quantité nécessaire pour une vérification pertinente.

L'usage d'un simulateur permet de réaliser des vérifications massives sans être contraint par la difficulté de collecter un nombre suffisant de données ; toutefois, la simulation nécessite en premier lieu que l'environnement d'utilisation du système soit modélisé et que ce modèle soit lui-même qualifié. **Dans le cas du véhicule autonome par exemple, la modélisation d'un environnement de conduite complet, permettant d'évaluer l'IA, est au stade de la recherche** (voir par exemple le projet 3SA[62]) et de premières offres commerciales (Dassault Systèmes, AVSimulation, Ansys, etc.).

La certification de l'IA est en lien direct avec la confiance : cela signifie, pour tous les acteurs de l'écosystème de confiance, qu'un organisme tiers a validé la conformité de l'IA. Nous comprenons qu'il est possible de certifier une IA (le système lui-même, des processus, etc.), mais qu'un certain nombre d'éléments doivent au préalable nécessiter des avancées. Au même titre qu'il est nécessaire d'outiller les concepteurs afin de leur permettre de développer des IA de confiance, **il est essentiel que les inspecteurs soient en mesure de s'appuyer sur des outils et méthodes permettant de caractériser la confiance.** La vérification, la validation et la certification des systèmes classiques (qui ne relèvent donc pas de l'IA) constituent déjà des tâches difficiles, même s'il existe déjà des technologies exploitables dans l'industrie. L'application aux systèmes d'IA complexes est une tâche importante à laquelle il convient de s'attaquer pour être en capacité d'utiliser ces systèmes dans des environnements critiques tels que les avions, les centrales nucléaires, les trains, les hôpitaux, etc.

D'autre part, l'**IA s'inscrivant dans un écosystème de confiance composé de différents acteurs et éléments, la validation devra s'appuyer sur des collaborations avec des spécialistes de diverses disciplines** des sciences informatiques, ainsi qu'avec des scientifiques d'autres champs de compétences contribuant à l'IA comme les psychologues, les sociologues, les biologistes (en biomimétique, notamment), les mathématiciens, etc. Nous notons également qu'**il ne s'agit pas uniquement de certifier le système, mais également les ensembles de données ayant contribué à son apprentissage, le cas échéant, le système et les sous-systèmes**, ainsi que les outils ayant servi à la création des applications d'IA finales.

[62] 3SA – Simulation pour la Sécurité des systèmes du véhicule Autonome

L'AUTONOMIE STRATÉGIQUE EST UNE CAPACITÉ À GÉNÉRER ET DÉFENDRE UN ÉCOSYSTÈME DE CONFIANCE QUI ORGANISE NOS INTERDÉPENDANCES.

II. ADRESSER UN DOUBLE ENJEU POLITIQUE

« La confiance est une arme défensive contre les monopoles que l'on subit, et offensive pour les coopérations que l'on choisit. »

La notion de confiance est double. Prise sous l'angle de la régulation, elle est évidemment défensive, un véritable « mal nécessaire » mais non suffisant pour que l'Europe puisse devenir une troisième voie numérique crédible, attractive, et compétitive.

Mais la confiance est aussi la source d'une stratégie offensive. La confiance n'est en effet pas qu'un ensemble règles qui viennent encadrer et sécuriser nos relations contractuelles avec des plateformes étrangères, **c'est aussi une promesse qui conditionne et fédère nos écosystèmes.**

Imposer les termes de la confiance pour nous protéger de ceux que l'on subit, n'est pas la même chose que bâtir la confiance pour nous lier à ceux que l'on choisit. Ce sont ces deux faces d'une même pièce avec lesquelles l'Union européenne doit savoir jouer.

2.1. LA CONFIANCE, UN ENJEU DE SOUVERAINETÉ NUMÉRIQUE

2.1.1. L'autonomie stratégique, ambition raisonnable de la souveraineté

« La souveraineté c'est la capacité, seul ou à plusieurs, à faire respecter ses valeurs, ses intérêts, et surtout ses lois. »

La souveraineté peut se définir comme la capacité à analyser, décider ou agir en fonction d'un ensemble de valeurs, de principes, d'intérêts et d'objectifs sans influence extérieure indue, manipulation ou contrainte.

La souveraineté s'applique à l'espace numérique tout autant qu'aux espaces terrestres, maritimes et spatiaux, et obéit aux mêmes aspirations d'étendre l'État de droit à son champ. Elle peut être entendue comme **la capacité des États « à se faire obéir, à imposer leurs lois, à apparaître comme devant être respectés dans l'espace numérique[63]. »**

Cependant, la définition de la « souveraineté numérique » souffre d'une trop grande variété d'interprétations. En l'absence d'une acception juridique précise, **cette notion reste finalement dépendante de définitions subjectives**, empêchant ainsi toute forme de consensus. Or le thème est central aujourd'hui !

Qu'est ce que l'autonomie stratégique ?

« On peut comprendre l'autonomie stratégique comme une volonté d'affranchissement de notre dépendance vis-à-vis d'une puissance extérieure, même amicale. C'est une vision assez gaullienne de la géopolitique selon laquelle l'alliance n'empêche pas l'autonomie. La décennie précédente a été jalonnée de ces évènements qui, par accumulation, ont jeté la lumière sur un impensé européen, recouvrant une dépendance généralisée non seulement aux géants du numérique, mais aussi au droit des puissances étrangères, fussent-elles alliées[64] » Laurence Houdeville et Arno Pons, Digital New Deal, 2021

[63] Pierre Trudel, professeur à l'université de Montréal
[64] *Cloud de confiance, un enjeu d'autonomie stratégique pour l'Europe*, Laurence Houdeville et Arno Pons, Digital New Deal, 2021.

La souveraineté peut ainsi se définir en fonction du niveau d'autonomie, c'est-à-dire sur sa capacité à choisir le niveau et la nature de ses dépendances. À l'image de la souveraineté énergétique où un État choisit de ne pas dépendre totalement d'un seul fournisseur, ni d'une seule énergie, la souveraineté numérique cherche elle aussi **à contrôler ses dépendances**. Soit en **« faisant soi-même »**, dans les rares cas d'une indépendance totale, soit plutôt en **« faisant avec les autres »**, dans un cadre de confiance maîtrisé. Et possiblement aussi en **« faisant pour les autres »** afin de se rendre indispensable, et d'ainsi pouvoir dissuader quiconque de menacer l'équilibre de la confiance par sa maîtrise d'une composante indispensable à la souveraineté de chacun.

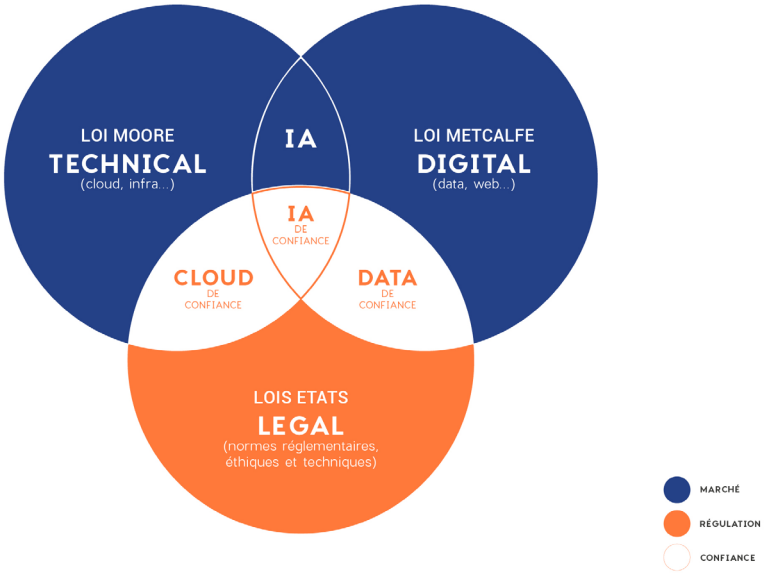
La souveraineté numérique est donc une gestion de nos dépendances technologiques et économiques. Ce n'est en effet pas un acquis, mais une quête. Nos niveaux de dépendances nous invitent à penser la souveraineté comme un idéal politique, un cap stratégique exigeant et nécessaire qui doit guider nos choix. **En somme, la souveraineté est une ambition, l'autonomie stratégique un objectif, et le niveau de dépendance une mesure.**

L'écosystème de confiance, périmètre naturel de la souveraineté numérique et de la prospérité

La confiance n'est pas un concept éthéré et indépendant du reste du monde : **elle ne peut se développer que dans un écosystème « de confiance » fait de règles** (lois, éthique, normes techniques..), de tiers de confiance, d'acteurs permettant de définir les règles et de juger les litiges, d'acteurs qui partagent et respectent cet ensemble de règles et d'acteurs qui anticipent les menaces et comportements futurs pesant sur l'écosystème de confiance. Un écosystème de confiance protège un pays (ou un groupe de pays), son économie, ses citoyens et doit être protégé par la loi, le droit, voire par la force si nécessaire. Pour un pays démocratique, c'est donc la qualité de son écosystème de confiance qui constitue le socle de son autonomie stratégique et donc de sa souveraineté.

« L'autonomie stratégique est une capacité à générer et défendre un écosystème de confiance qui organise nos interdépendances. »

ÉCOSYSTÈME DE CONFIANCE (CLOUD + DATA + IA) = AUTONOMIE STRATÉGIQUE



Gérer ses dépendances, c'est choisir ses interdépendances et nourrir son corollaire, l'indispensabilité. Pour protéger cet équilibre délicat, la quête de souveraineté consiste ainsi à construire et sécuriser un « écosystème de confiance ». Un environnement de confiance basé sur une maîtrise technologique d'une infrastructure matérielle et immatérielle commune, un cadre réglementaire sécurisant, une capacité à se défendre, et bien évidemment un choix éclairé de parties prenantes fiables et complémentaires. Des partenaires solidaires qui partagent les mêmes valeurs de respect des données, d'ouverture et de transparence, au point d'assumer auprès d'eux des dépendances librement consenties pour pallier nos faiblesses. Cette gestion éclairée des dépendances doit faire l'objet d'une gouvernance dédiée et permettre une interopérabilité. **À défaut de garantir une chaîne de valeur technologique pleinement souveraine (auto-suffisante), le grand enjeu de cette gouvernance équilibrée sera de sanctuariser une « chaîne ou un réseau de confiance » internationale.** Un mode de gouvernance qui peut composer sur certaines strates, mais ne jamais subir les influences extérieures.

« Pour que la France soit à la hauteur de son ambition européenne, à la hauteur aussi de son histoire, elle doit rester souveraine ou décider elle-même, sans les subir, les transferts de souveraineté qu'elle consentirait, tout comme les coopérations contraignantes dans lesquelles elle s'engagerait.[65] » Emmanuel Macron

L'EXEMPLE DE L'INDISPENSABILITÉ JAPONAISE

Si cette évolution stratégique est récente en Europe, elle est en revanche ancienne pour d'autres puissances, en particulier pour le Japon, historiquement protectionniste et désormais très exposé à l'imprévisibilité et à l'arbitraire du pouvoir chinois. Un groupe d'étude du Parti Libéral Démocrate japonais a rédigé un rapport enjoignant le gouvernement à construire une stratégie de sécurité économique, en réponse à aux menaces que fait peser son puissant voisin.

« Le rapport propose de poursuivre deux objectifs stratégiques : 1) la préservation de l'autonomie stratégique, qui se traduit principalement par des mesures de protection ; 2) le renforcement de ***« l'indispensabilité stratégique » du Japon, concept qui vise « de façon stratégique, à accroître le nombre de secteurs dans la structure industrielle mondiale, où le Japon est essentiel à la communauté internationale.***[66] ***»***

Ainsi, ***« le positionnement technologique et industriel du Japon (…) doit lui permettre de dissuader ou de riposter à d' éventuelles actions de coercition économique,*** *comme le Japon a eu à en souffrir en 2010 lorsque la Chine a interrompu ses exportations de terres rares.* »

Autrement dit, le Japon tente d'exercer, à travers cette indispensabilité stratégique fondée sur ses choix économiques, une forme de **« *soft deterrence* – dissuasion douce »**. Celle-ci doit permettre de préserver le pays des rapports de force brutaux et se maintenir dans une centralité stratégique pour les rivaux comme pour les alliés.

Source : *L'ambition japonaise d'une stratégie de sécurité économique : une voie à suivre, Nicolas Regaud, Brève stratégique* - 20, 15 avril 2021, IRSEM, https://www.irsem.fr/publications-de-l-irsem/breves-strategiques/breve-strategique-n-20-2021.html

La souveraineté ne doit surtout pas être empruntée sur son flanc protectionniste : elle doit être entendue au contraire de manière positive et constructive. La quête de souveraineté, c'est celle de l'autonomie, pas celle de l'autarcie. Ce n'est pas tant imposer ses choix que sa capacité à ne pas subir ceux des autres. Cela implique donc le prérequis de la prise en compte de l'autre, et de la gestion de la relation à l'autre. La politique visant la souveraineté technologique ne consiste pas à se couper de l'extérieur, ce qui serait une hérésie dans un monde dématérialisé et totalement interdépendant, mais au contraire à gérer nos niveaux d'interdépendance et les interopérabilités associées.

[65] *Discours du président de la République Emmanuel Macron*, 7 février 2020 sur la stratégie de défense et de dissuasion devant les stagiaires de la 27e promotion de l'École de Guerre

[66] *L'ambition japonaise d'une stratégie de sécurité économique : une voie à suivre*, Nicolas Regaud, Brève stratégique - 20, 15 avril 2021, IRSEM

Tout l'art de la souveraineté consiste à choisir de qui nous souhaitons partiellement dépendre, et à quel niveau, c'est-à-dire en qui nous avons suffisamment confiance. Ce n'est pas un hasard si c'est en Europe que la notion de « numérique de confiance » se dessine, continent où la libre circulation est le socle de notre vie économique et politique, un lieu où la prospérité est plus qu'ailleurs indissociable de la notion d'échange. Entre la Chine qui prône l'autosuffisance, et les États-Unis qui assument une forme d'ingérence via l'extraterritorialité du droit américain, **il s'agit donc pour l'Union européenne d'inventer une forme de souveraineté qui ne soit synonyme ni de protectionnisme, ni de féodalisme, mais au contraire dans la définition d'une autonomie stratégique partagée, ouverte et multipartite, faite de dépendances acceptées et réciproques avec des partenaires de confiance, une troisième voie.**

La souveraineté est en quelque sorte l'intérêt du capital confiance, un capital qu'il faut donc investir et dépenser avec prudence, car comme le rappelait Jean-Paul Sartre **« La confiance se gagne en gouttes et se perd en litres ».** La souveraineté doit pouvoir en effet s'appuyer sur un écosystème de confiance qui permette de sceller les membres au sein de sa communauté de valeurs et d'intérêts, et d'imposer ses choix à ceux qui sont en dehors. **L'Union européenne a su patiemment construire ce capital confiance au fil des décennies, il s'agit maintenant de le transposer au numérique.**

Le monde a besoin d'une troisième voie entre les deux impérialismes technologiques, car les pays doivent avoir la possibilité d'adhérer à des valeurs universelles compatibles avec le multilatéralisme. En Europe plus qu'ailleurs, il est souhaitable que la souveraineté des États démocratiques soit associée à la confiance des citoyens dans leur gouvernement et leurs institutions, car sans elle il n'y a pas de modèle humaniste possible. **La souveraineté sans la confiance c'est l'exercice unilatéral de la force**, l'atteinte aux échanges multilatéraux, et donc l'entrave à la prospérité telle que nous la concevons dans l'Union européenne.

La multitude comme expression de l'écosystème de confiance

Choisir la multitude, c'est choisir le nombre et la diversité orientés dans un but commun. C'est une stratégie de coopération et de partage des risques qui, en multipliant les échanges et les interdépendances, permet aussi **la décentralisation : la distribution des fonctions et des compétences chez les partenaires.** Autrement dit, l'antithèse parfaite de la centralisation et de la concentration monopolistique.

Choisir la multitude, c'est aussi se protéger en multipliant les points de sécurisation. Cette multiplication des acteurs chargés de protéger et sécuriser est une stratégie de résilience. En effet en cas de crise, si un agent ou une partie fait défaut, l'effectif restant peut prendre le relai et permettre la continuité des activités. La stratégie de la multitude permet ainsi de se relever et d'atténuer les effets d'une crise en diversifiant les réseaux de distribution ou les sources d'approvisionnement, c'est-à-dire les fonctions indispensables d'une organisation. C'est le paradigme du réseau, souple, robuste et adaptable.

Les oligopoles, résultant dans l'économie numérique de l'adage *« the winner takes all »*, vont précisément à l'encontre de l'approche de la multitude. La stratégie de centralisation et de déverrouillage (*lock-in*) qu'ils mettent en œuvre est aussi particulièrement vulnérable à la rupture de confiance.

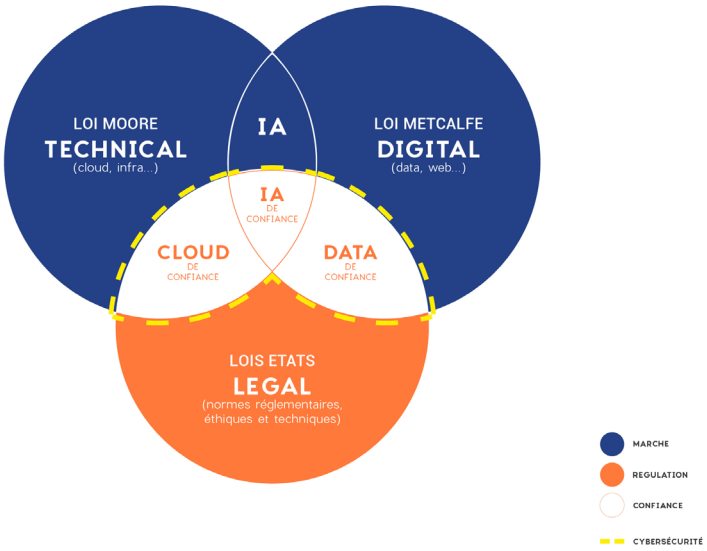
Lutter contre l'impérialisme des géants du net n'est donc pas une simple réaction protectionniste ou un désir d'autarcie, mais une lutte contre la centralisation des pouvoirs, pour créer un contre-modèle démocratique, souverain et digne de confiance.

2.1.2. Protéger, condition sous-jacente à la confiance

Souveraineté, une capacité à influencer et contrôler les menaces

« Une infrastructure de confiance, c'est une capacité à maîtriser l'indépendance d'une architecture, et la capacité à la défendre.[67] » Guillaume Poupard, Digital New Deal, 2021

ÉCOSYSTÈME DE CONFIANCE (CLOUD + DATA + IA) = AUTONOMIE STRATÉGIQUE



Protéger, c'est d'abord nous sécuriser par la cybersécurité et améliorer notre résilience. La question des cybermenaces devient de plus en plus prégnante, et son pendant, le pouvoir de coercition, est rarement évoqué. Or il n'y a pas de sécurité sans menace de sanction applicable suffisamment dissuasive. Un certain niveau de contrôle technologique, permettant de garder la main sur les points de vulnérabilité, est une nécessité. Pour exercer sa souveraineté dans l'espace numérique, il faut être en mesure de se protéger et de riposter aux menaces cyber, et garder le contrôle de l'infrastructure comme des données, aussi bien dans le monde physique que dans l'espace numérique. Ces menaces sont protéiformes, et peuvent cibler des éléments du cyberespace, des organismes d'intérêt vital (OIV), des exploitants des moyens numériques, des individus ou encore des organisations.

Un écosystème de confiance doit par ailleurs faire preuve d'anticipation stratégique et de résilience en démontrant son aptitude à faire face à un événement, à une perturbation, en réagissant ou en se réorganisant de manière à maintenir ses fonctions, son identité et ses structures essentielles, tout en conservant la capacité d'analyser, de décider ou d'agir.

Protéger, c'est ensuite sanctuariser nos intérêts par la régulation. Car c'est aussi une question de volonté politique de ne pas se cantonner à une stratégie uniquement défensive et sécuritaire, mais également mener une campagne offensive visant à influencer le marché par le droit. L'Union européenne doit continuer de jouer à plein son influence politique à travers un *soft power* assumé, mêlant poids géopolitique, capacité à imposer ses valeurs, à orienter la régulation, et à peser sur les normes et standards qui structurent le marché. Cette puissance normative européenne doit aussi créer les conditions d'un écosystème de confiance basé sur le droit en pesant dans sa gouvernance. *« L'Union européenne ne doit pas hésiter à remettre en question le cadre politique et juridique actuel si elle veut atteindre une totale autonomie dans sa capacité d'appréciation et de gestion du risque cyber[68]. »*

[67] Déclaration de Guillaume Poupard, Directeur Général de l'ANSSI, auditionné dans le cadre de notre publication *Infrastructures du numérique de confiance. Un enjeu stratégique pour les territoires*, Digital New Deal, novembre 2021
[68] *Cybersécurité, vigile de notre autonomie stratégique*, Arnaud Martin et Didier Gras, Digital New Deal, 2022

Le « marché » américain vs la « régulation » européenne

L'Europe et les États-Unis divergent justement dans leur approche de la régulation des technologies : **l'Europe privilégie une gestion des risques induits par la technologie *ex ante*, là où les États-Unis préfèrent un contrôle *ex post*.** On peut interpréter ces divergences par des visions de la confiance divergentes :

- **Pour l'Europe la confiance passe par la mise en place d'un écosystème de confiance reposant notamment sur des lois : la confiance est donc d'abord l'affaire des États.** C'est généralement la régulation qui donne le cadre permettant aux acteurs économiques d'innover et de se développer, en s'appuyant sur des exigences ou des valeurs de haut niveau (éthique, droits fondamentaux, etc.). L'Europe tente d'anticiper les risques, tant pour les individus que pour les entreprises, puis met en place des processus d'encadrement et de contrôle forts des marchés, afin de rassurer l'ensemble des acteurs économiques. **Cette approche permet à l'Europe de se positionner comme une référence mondiale sur les enjeux de régulation, mais elle est cependant vécue par certains acteurs comme un frein à l'innovation contribuant à la relative absence de champions du numérique sur certains secteurs.**

- **Pour les États-Unis la confiance est associée à une relation entre plusieurs entités (individus, organisations, systèmes) encadrée par des contrats. Cette vision offre une plus grande liberté d'innover pour les acteurs du marché.** Le poids plus faible de la régulation est contrebalancé par un système judiciaire offrant des capacités fortes aux acteurs de se défendre (ex : le mécanisme de *class action* hérité du système judiciaire anglais, qui inspire aujourd'hui l'Europe), et des mécanismes antitrust qui permettent à l'État de lutter contre les monopoles (ex : démantèlement d'*American Telephone and Telegraph, AT&T*, sur la seconde moitié du XX[e] siècle[69]). **Cette approche « par le marché » est certainement l'un des multiples facteurs clés du succès commercial des géants américains. L'approche des États-Unis permet à ses acteurs économiques d'apprendre et de progresser plus vite (*test and learn* ou *fail fast*).**

Ces dernières années, **l'impossibilité pour l'Europe de s'appuyer sur des champions numériques locaux a conduit à de nombreuses confrontations directes entre les institutions européennes et les géants du numérique américains**, en particulier sur les questions relatives aux données personnelles. **À la domination économique américaine presque totale sur les enjeux B2C (*Business to Consumer*[70]), l'Europe répond la plupart du temps par la régulation :**

- **Le RGPD** (*Règlement Général sur la Protection des Données*) propose un cadre réglementaire pour la protection des données personnelles.
- **Le DSA** (*Digital Services Act*) crée un marché plus sûr, plus équilibré et respectueux des droits fondamentaux.
- **Le DMA (***Digital Markets Act*) limite les pratiques anticoncurrentielles de verrouillage des grandes plateformes.
- **Le DGA** (*Data Governance Act*) et le **DA** (*Data Act*) proposent un cadre pour la création d'écosystèmes de partage des données.

Ces réglementations ont pour vocation à inspirer des normes mondiales. Le RGPD dispose désormais par exemple de son équivalent californien avec le California Consumer Privacy Act (CCPA), et de nombreux États dans le monde ont emboîté le pas de l'Europe en adoptant

[69] A.T.T est accusée par la justice américaine de violer la loi antitrust, Le Monde, 22.11.1974,
[70] Désigne l'ensemble des architectures techniques et logiciels informatiques permettant de mettre en relation des entreprises directement avec les consommateurs - Wikipedia

L'ASYMÉTRIE ÉCONOMIQUE ENTRE AMÉRICAINS ET EUROPÉENS EST SANS CESSE AMPLIFIÉE PAR UNE ASYNCHRONIE JURIDIQUE.

des réglementations similaires. L'influence de la réglementation est également décuplée par le principe d'extraterritorialité, qui impose aux acteurs internationaux souhaitant accéder au marché européen, toujours attractif pour eux, de se conformer.

« **Si ces dispositions [RGPD] avaient existé il y a 20 ans, il est probable que Facebook, Amazon ou Google n'auraient pas pénétré le marché européen aussi facilement et que la concurrence aurait pu démarrer sur des bases plus saines.**[71] » Cependant les leviers de souveraineté de l'Europe sont souvent réduits à des sanctions, parfois difficilement applicables, comme en témoigne le décevant bilan de l'effectivité des amendes : « Apple condamné à payer 13 milliards € a vu sa condamnation annulée en appel, Google condamné à 8,2 milliards € d'amendes n'en a vu qu'une seule confirmée soit 2,4 milliards €, condamnation d'Amazon annulée, etc… Pour l'ensemble des GAFAM c'est au final 3,4 milliards € effectivement payées en 20 ans, à comparer au 300 milliards € de profits pour la seule année 2021 »[72].

Il est enfin presque illusoire pour l'Europe de faire peser par exemple la menace d'un démantèlement sur des acteurs américains qui abuseraient de leur position dominante, ou encore de sanctuariser certains aspects de son développement technologique lorsqu'elle ne maîtrise pas les technologies requises.

2.1.3. Atteindre l'autonomie stratégique par l'écosystème de confiance

La régulation, condition nécessaire mais non suffisante

Le travail de transposition conceptuelle de la souveraineté appliqué au numérique et à l'espace européen jette une lumière crue sur nos dépendances à des géants économiques et géopolitiques dont les principes et les pratiques s'éloignent des nôtres. Or les **technologies numériques traduisent dans leurs architectures mêmes des conceptions de la société,** de la manière dont un pays perçoit son rôle dans le monde.

La Chine voit par exemple dans l'IA un « pilier de la société harmonieuse de demain[73] **», avec l'objectif de devenir le leader mondial d'ici 2030.** Grâce à sa puissance de calcul, sa masse de données disponibles et ses compétences, la Chine se hisse progressivement au niveau technologique des États-Unis. **Mais elle mobilise par exemple ces technologies pour mettre en œuvre un système de « crédit social » jugée inacceptable dans le règlement européen sur l'IA (AI Act).**

[71] *Donner un sens à l'intelligence artificielle, Mission parlementaire Villani*, 2018
[72] *Puissance européenne face aux GAFAM : Mythe ou réalité ?*, André Loesekrug-Pietri, Les Echos, mai 2022
[73] *Le système de crédit social en Chine. La discipline et la morale*, Séverine Arsène, revue Réseaux, 2021/1 n°225, pp. 55-86 ; *Le système de crédit social*, Wikipédia ; Xinyong. *L'expression de la confiance en Chine*, Thierry Pairault, *Rapport moral sur l'argent dans le monde*, Finance et société, 1996

LE CRÉDIT SOCIAL CHINOIS
L'AUTRE VISION DE LA CONFIANCE

L'instauration partielle du crédit social depuis 2018 a donné une visibilité inédite au positionnement chinois sur le numérique. Le crédit social est un système de notation des citoyens sur leur réputation personnelle, mais aussi des entreprises, avec un système de récompenses et de pénalités. **Le crédit social repose sur des outils de surveillance de masse et utilise le *Big Data* pour établir un « réseau de confiance »** (*shouxìn*) et réduire les possibilités de fraude.

L'idée prend forme dans les années 2000 dans une volonté d'améliorer la solvabilité des entreprises, puis des citoyens. Au-delà de la gestion de risque crédit, l'objectif était d'inciter « à l'intégrité et à la crédibilité au sein de la société », et améliorer l'économie socialiste de marché chinoise. Quatre objectifs structurent le crédit social : « l'honnêteté dans les affaires du gouvernement », « l'intégrité commerciale », « l'intégrité sociétale » et la « crédibilité de la justice ». Si le contrôle social a fait scandale à l'étranger, **l'objet principal de la réforme était de « fournir une réponse au problème du manque de confiance dans le marché chinois ».** Le crédit social servirait pour l'économie socialiste de « mécanisme de régulation du marché ».

La mise en œuvre du crédit social en Chine montre que l'invocation de la confiance dans le numérique n'est pas une garantie en soi. Au contraire, **cet usage de la confiance, considéré comme un risque inacceptable dans le projet d'AI Act européen, nous incite à définir rigoureusement nos principes, valeurs et objectifs, aussi bien politiquement que technologiquement.**

Sources : *Le système de crédit social en Chine. La discipline et la morale*, Séverine Arsène, revue Réseaux, 2021/1 n°225, pp. 55-86, *Le système de crédit social en Chine* | Cairn.info ; *Xinyong : l'expression de la confiance en Chine*, Thierry Pairault, *Rapport moral sur l'argent dans le monde*, Finance et société, 1996

L'Europe est donc face à un impératif politique fondamental : construire son autonomie stratégique autour d'un écosystème de confiance, basé sur ses valeurs, ceci en activant tous les leviers d'investissement et de souveraineté à sa disposition.

Une souveraineté industrielle nécessaire à l'écosystème de confiance

Protéger c'est faire de la confiance une force de l'Europe sur le marché global du numérique, mais aussi sa clé d'entrée sur le marché mondial de l'IA.

Face à la concurrence d'acteurs américains et chinois, souvent mieux armés financièrement, il est impératif pour l'Europe d'imposer ses valeurs non seulement par la réglementation (AI Act), mais aussi via une stratégie industrielle « par le marché », par les usages et par les technologies. L'Europe doit analyser attentivement les marchés prioritaires, et les chaînes de valeur de l'IA, dans tous les secteurs et y introduire, ou y représenter, la notion de confiance, ce que nous nous proposons de faire sur l'IA dans les chapitres suivants.

Cette approche favorisera ainsi le développement de champions européens de l'IA, et surtout, d'écosystèmes technologiques dynamiques et multiples comprenant l'ensemble du tissu économique : acteurs publics et privés, collectivités, grands groupes, startups, PME, universités, associations, etc.

FAIRE DE LA CULTURE DES SYSTÈMES CRITIQUES, LA POINTE DE LA FLÈCHE DE LA STRATÉGIE IA DE CONFIANCE.

2.2. LA CONFIANCE, UN ENJEU DE COMPÉTITIVITÉ POUR L'EUROPE

2.2.1. Faire de la confiance la valeur étalon du marché

Définir nos propres règles

Il existe deux approches culturelles de la confiance : la première davantage organisée par la sphère politique (culture européenne), la seconde par la sphère privée (culture anglo-saxonne). Dans ce deuxième cas de figure, la définition s'avère alors plus flexible, voire élastique, en fonction du contexte et des cas particuliers. Nous devons jouer suivant nos règles, et ainsi permettre à nos acteurs économiques d'être compétitifs **en changeant la valeur étalon du marché**.

Comment ? **En imposant nos propres agences de notation, en indexant la conformité sur nos critères et en promulguant nos outils de mesure pour ne pas avoir à subir l'opacité de services privés étrangers**. Puis en dépassant le simple *« ethics washing »* consistant à créer une forme d'ambiguïté entre le droit et la morale.

L'Union européenne doit poursuivre sa stratégie initiée avec le RGPD consistant à imposer de facto les critères de pénétration du marché, en faisant de la confiance le sujet principal. Sinon les *Big Tech* le réduiront un simple argument commercial. En effet, si nous n'imposons pas nos propres critères de compétitivité, alors les géants américains n'auront aucun mal à préempter la confiance en la réduisant au rang de commodité. Leurs incommensurables moyens financiers et leurs indiscutables talents marketing faisant le reste.

Mais y a-t-il une demande d'IA de confiance justifiant le développement de l'offre européenne ?

Pour nous aider à répondre à cette question, le Secrétariat Général Pour l'Investissement (SGPI) à EY-Parthenon de réaliser une étude de marché de l'IA de confiance (*Sans confiance, quel avenir pour l'Intelligence Artificielle dans l'industrie ?*, Novembre 2021) dans une série de secteurs industriels à hauts risques stratégiques (automobile, aéronautique, ferroviaire, banque, assurance, santé, énergie et réseau énergétique, etc.) en Europe, en Amérique du Nord et en Asie-Pacifique. Cette analyse a pour but d'identifier et d'évaluer les cas d'usages principaux, à ce jour, où un « degré variable de confiance » est nécessaire pour pouvoir déployer l'IA dans des environnements industriels. La partie suivante s'appuie donc principalement sur cette étude de marché, complémentaire du présent rapport.

2.2.2. Un déploiement prudent de l'IA expliqué par des difficultés d'industrialisation et un manque de confiance

Taille de marché globale de l'IA

Le marché de l'intelligence artificielle est estimé à **231 milliards d'euros en 2020, prévu en forte croissance de 18% par an d'ici à 2024**[74] et porté par l'adoption progressive de solutions d'IA dans tous les secteurs de l'industrie.

[74] Ibid (étude EY-Parthenon)

La prudence, voire un manque de maturité, des industriels

« Si en moyenne les industriels (tous secteurs confondus) investissent entre 0,4 % et 1 % de leur chiffre d'affaires (avec en tête les acteurs des technologies, etc.) dans des projets impliquant de l'IA, **l'industrie reste prudente (ou peu dotée) pour intégrer à plus large échelle des composants d'IA dans des processus, produits ou services industriels.** », rappelle Ile EY-Parthenon.

D'après leurs estimations :

- **Seules 10 à 15 % des entreprises interrogées ont réussi à industrialiser** des solutions à base d'IA,
- **30 à 40 % d'entre elles se limitent à des expérimentations** sur des périmètres ou processus bornés.

Plusieurs raisons expliquent ce constat, selon cette étude :

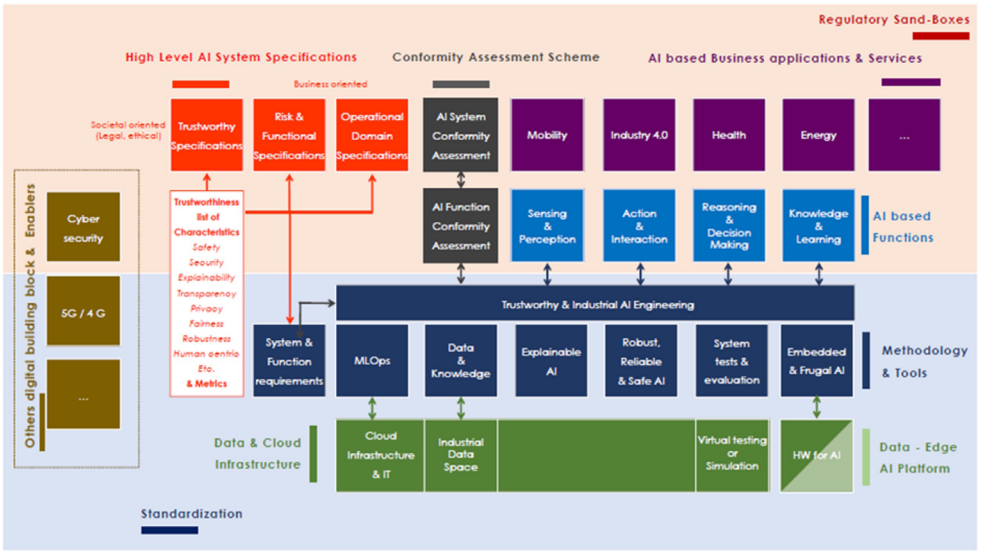
- **La disponibilité de données, à la fois en quantité et en qualité suffisante** pour entraîner, puis déployer en production des modèles d'IA et suivre leur « bon » fonctionnement. Il existe donc un enjeu de confiance et de complétude des données relatives à l'application visée.
- **L'évaluation du retour sur investissement (ROI) en IA s'avère complexe** et représente une incertitude lors de la prise de décision. L'IA est en effet une « technologie d'usage général » et, à ce titre, sert l'ensemble des business des industriels, mais ne constitue pas forcément leur cœur *business*.
- **La difficulté d'intégration à des systèmes plus globaux :** l'IA est un composant ou une fonction d'un système plus complexe. **Son industrialisation requiert une ingénierie spécifique, mais aussi des interopérabilités avec les autres chaînes d'ingénierie. La démonstration de la conformité aux exigences** ou spécifications fonctionnelles et non fonctionnelles constitue aussi un point bloquant. La confiance est nécessaire pour répondre à l'enjeu de responsabilité de l'entreprise.

Les conditions nécessaires à l'adoption par les industriels

Ce constat est également partagé par d'autres analyses, comme celle menée par Quantmetry en 2021 auprès de 25 entreprises clés du secteur. Celle-ci précise que **« la future réglementation est perçue comme une nécessité », afin notamment de sécuriser les relations entre industriels et garantir un cadre juridique pour encadrer la responsabilité.** Elle souligne le manque de mesures organisationnelles et stratégiques mises en œuvre pour garantir l'application de la future réglementation.

Le *position paper* franco-allemand « *speeding up industrial AI and trustworthiness*[75] », fruit du partenariat entre le Secrétariat Général pour l'Investissement (SGPI) et la *Big Data Value Association* (BDVA), en collaboration avec une cinquantaine d'experts académiques et industriels, illustre également **le lien étroit entre confiance et industrialisation de l'IA. Il souligne le besoin industriel de solutions logicielles pour implémenter la confiance dans les processus et applications à base d'IA, couplé à la disponibilité de normes et de standards**, ceci en cohérence avec les initiatives européennes sur le cloud (Gaia-X).

[75] *Speeding up industrial AI and trustworthiness*, Secrétariat Général pour l'Investissement, Big Data Value Association et al., 2021



Source : *Speeding up industrial AI and trustworthiness*, SGPI, Big Data Value Association, 2021

La bataille des outils et méthodes de conception d'IA

L'IA est un ensemble de technologies complexes pour lesquelles les compétences sont rares, variées, et le débat entre internalisation et externalisation incessant. En conséquence, outre le besoin de données et de connaissances, **de nombreuses industries reposent aujourd'hui principalement sur des plateformes d'outils logiciels et des méthodes de développement ou d'« ingénierie » fournis par des prestataires spécialisés**. Ces outils et méthodes permettent de développer des algorithmes d'IA à partir de modèles, de données et de connaissances.

Le segment des solutions d'intelligence artificielle tire d'ailleurs la croissance du marché de l'IA dans son ensemble, accompagné chez les industriels d'un besoin de plus en plus important en infrastructures et en accompagnement pour entraîner et déployer des modèles en production.

Les géants du numérique prennent ici position grâce à leurs offres intégrées, associant des infrastructures de stockage (*cloud*) et de calcul, sur lesquelles ils dominent très largement le marché, ainsi que des solutions logicielles pour gérer les données et accompagner les développements d'IA (quelle que soit la typologie de données et des modèles). **Toutefois, la confiance requiert une infrastructure, encore à développer, d'outils et de méthodes spécifiques, pour être déployée.** Celles-ci doivent être interopérables avec les autres chaînes de conception (MLOps[76],ModelOps et Systèmes), et permettre d'analyser les risques et de concevoir, valider et contrôler, y compris en fonctionnement, le système et ses attributs de confiance pour une application donnée.

L'IA de confiance sera en grande partie développée via ces outils et méthodes, que nous regroupons dans ce rapport sous le terme « InfraTech de la confiance dans l'IA », véritable *enabler* de la vision politique européenne (portée par sa réglementation), du marché de l'IA de confiance (compétitivité économique) et de sa diffusion à l'ensemble du tissu économique, y compris les PME et startups.

[76] Ensemble de pratiques qui vise à déployer et maintenir des modèles de machine learning en production de manière fiable et efficace

2.2.3. Une analyse du marché de l'IA de confiance sur des filières industrielles stratégiques pour l'Europe

Taille de marché de l'IA et de l'IA de confiance pour 9 industries à haut risque

L'étude EY-Parthenon permet de proposer une estimation « basse » de la taille réelle du marché adressable pour l'IA de confiance sur la base de 9 segments industriels analysés de manière non holistique (plus de 50 entretiens menés, plus une analyse sur la base des 25 à 35 principaux acteurs industriels par marché) : santé, transport routier, transport aérien, transport ferroviaire, pétrole et gaz, énergie, banque et assurance.

Elle fonde son analyse sur les principaux cas d'usage actuels par secteurs, tout en soulignant l'émergence de futurs cas d'usage innovants qui seront accompagnés par la monté en maturité de l'IA de confiance et pour lesquels une évaluation de marché s'avère complexe du fait de leur fort impact sur la transformation des *business* industriels.

Illustration 4 : principaux cas d'usage par secteur industriel

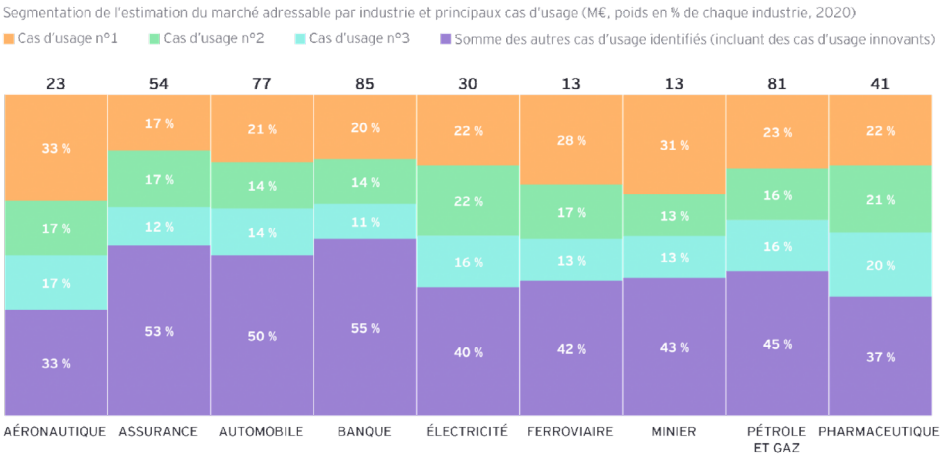
Segmentation de l'estimation du marché adressable par industrie et principaux cas d'usage (M€, poids en % de chaque industrie, 2020)

Industrie / segment du marché indirect	Cas d'usage n°1	Cas d'usage n°2	Cas d'usage n°3	Cas d'usage innovant
AÉRONAUTIQUE	Maintenance prédictive des avions / équipements	Contrôle qualité dans le processus de production	Industrie 4.0	Gestion du trafic aérien (IA embarquée et avion autonome)
ASSURANCE	Détection de fraude	Gestion et optimisation des flux de trésorerie	Cybersécurité	Conseillé augmenté (gestion des sinistres)
AUTOMOBILE	Industrie 4.0	Connaissance client via IA embarqué dans le véhicule	Services connectés (maintenance gestion flotte)	Interactions du véhicule avec son écosystème (véhicule autonome)
BANQUE	Cybersécurité	Lutte anti-blanchiment	Détection de fraude	Conseillé augmenté (octroi de crédit)
ÉLECTRICITÉ	Surveillance des infrastructures (sites / réseau)	Trading et tenue de marché	Maintenance prédictive des équipements	Pilotage smart home et smart grid
FERROVIAIRE	Maintenance prédictive des trains et des infrastructures	Surveillance du réseau et des infrastructures	Automatisation du trafic	Gestion des offres de nouvelles mobilités et pilotage énergétique
MINIER	Optimisation logistique (routes commerciales)	Maintenance prédictive des équipements de production	Smart mining (automatisation des processus de production)	Exploration et détection de gisements
PÉTROLE ET GAZ	Pilotage de la production (état du gisement et de l'extraction)	Optimisation des forages (design et exploitation du puit)	Optimisation logistique (routes commerciales)	Exploration et détection de gisements
PHARMACEUTIQUE	Optimisation des tests de molécules (phases cliniques	Optimisation de la découverte de molécules (biomarqueurs)	Suivi de la performance des ventes de molécules	Pharmaco vigilance

Source : EY Parthenon, SGPI, 2021

L'analyse propose un poids relatif des cas d'usage actuels par secteur industriel, permettant ainsi de déterminer les premières applications d'IA de confiance. Il représente en moyenne entre 40 et 60 % des budgets IA consacrés par ces secteurs. Toutefois, soulignons à nouveau que d'autres cas d'usage, dont le caractère transformant est non négligeable voire déterminant pour ces filières, ne sont pas évalués en termes de marché. Cela implique potentiellement une croissance très forte, tirée par des cas d'usage de plus en plus innovants, justifiant la qualification de « basse » pour cette évaluation de marché.

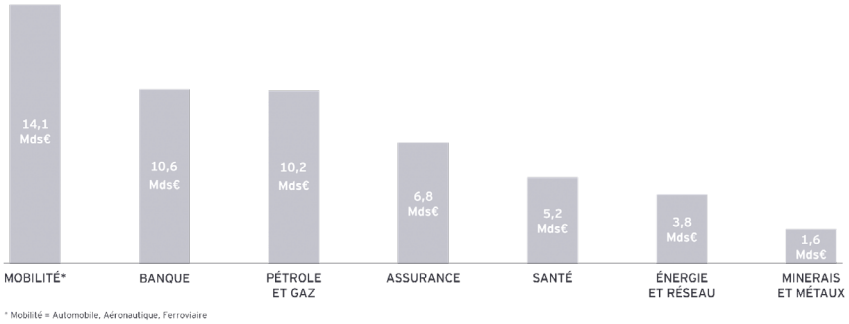
Illustration 3 : répartition des budgets IA entre les principaux cas d'usage par secteur industriel



Source : EY Parthenon, SGPI, 2021

L'étude EY estime ainsi une **taille de marché de l'IA de 80 milliards € pour ces 9 filières industrielles, pour un marché de l'IA de confiance de 53 milliards €, soulignant à nouveau l'importance de la confiance pour l'industrialisation de solutions à base d'IA.**

Illustration 2 : estimation du budget IA de confiance par secteur industriel étudié (limitée aux entreprises incluses dans le périmètre d'analyse en 2020)



Source : EY Parthenon, SGPI, 2021

Taille de marché pour une « InfraTech » de solutions d'IA de confiance

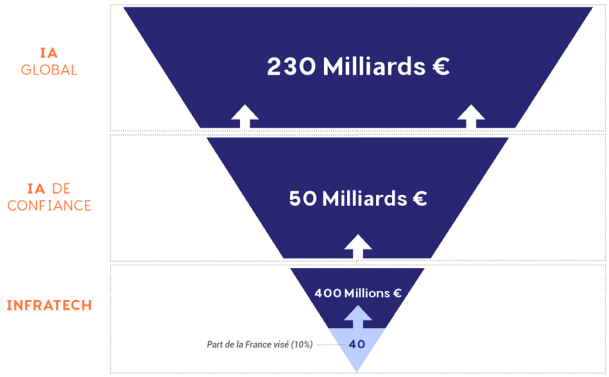
S'agissant de l'écosystème de solutions d'IA de confiance, EY-Parthenon souligne aujourd'hui une fragmentation importante entre :

- **des fournisseurs de plateformes** davantage généralistes (en particulier les *hyperscalers* américains) de développement d'algorithmes IA qui intègrent progressivement des solutions spécifiques,
- **des fournisseurs plus spécialisés de technologies pour un des attributs de confiance** (explicabilité, robustesse, etc.),
- **des fournisseurs de solutions** pour évaluer la qualité de logiciels,
- **des auditeurs**, voire certificateurs, de systèmes à base d'IA de confiance.

Le **« montant des investissements dans l'achat de telles solutions s'élève à environ 420 millions (€)** répartis entre l'Europe (180 M€ incluant 40 M€ en France), l'Amérique du Nord (110 M€) et l'Asie (130 M€). Parmi les secteurs les plus porteurs se trouvent la mobilité (113 M€ de marché identifié),

UNE INFRATECH QUI,
EN BAISSANT LES COÛTS
D'ACCÈS À L'IA DE
CONFIANCE, PERMET
DE CONQUÉRIR
LES MARCHÉS DE L'IA
EN CASCADE.

la banque (85 M€), l'assurance (54 M€) ou encore le pétrole et gaz (81 M€). **Les perspectives de croissance sont importantes, tirées par la commercialisation de futurs produits ou services innovants et la mise en application, à venir, de cadres réglementaires ou normatifs** comme l'illustre les travaux en cours de la Commission européenne sur l'AI Act ».



Nous considérons que les applications qualifiées « à haut risque » par la Commission européenne couvrent un périmètre bien plus large que les seuls systèmes critiques[77] (*safety critical*), englobant également des application considérées comme *business critical*[78] (tourisme, culture, alimentaire, etc.) **et surtout *society critical*[79]**(éducation, emploi, etc.) impliquant notamment des considérations éthiques ; ceci d'autant plus que le Parlement européen incite à l'élaboration d'un « code de conduite »[80] pour toutes les applications d'IA, y compris celles qui ne sont pas considérées comme « à haut risque ». Cela impliquerait à terme que l'ensemble des produits et services à base d'IA constituerait un marché pour l'InfraTech de l'IA de confiance.

Les États-Unis comptent dans leurs rangs des acteurs majeurs dans les domaines des systèmes critiques à haut risque (exemple : Boeing dans l'aéronautique). Cependant les développements de l'IA sont aujourd'hui fortement « tirés et influencés » par **les acteurs du B2C comme les GAFAM. Or leur culture est plutôt basée sur les usages, une approche IT *(Information Technology)* et *fail fast* qui rejette la peur de l'échec et valorise l'expérimentation continue**.

L'Europe peut adopter à contrario une approche fondée sur l'apport de preuve, de performances et pas uniquement processus, relativement à une analyse des risques et des exigences fortes qui caractérisent les systèmes critiques ou à « haut risque ».

Nous recommandons à l'Europe de s'appuyer sur la culture des systèmes critiques (*safety critical*) et de l'OT (*Operational Technology*), **véritable pointe de la flèche du développement d'ingénierie des systèmes à base d'IA de confiance** et donc de l'écosystème de confiance dans son ensemble. Ce constat est d'ailleurs partagé par des acteurs industriels issus de plusieurs filières (énergie, défense, santé, etc.) et se retrouve dans le *Manifeste IA*[81] produit par un collectif d'industriels français.

L'Europe doit aujourd'hui définir une stratégie industrielle ambitieuse s'appuyant sur les premières analyses du marché de l'IA de confiance. Pour ce faire, elle doit donc **promouvoir d'une part l'offre, par la normalisation, et d'autre part la demande par le développement de solutions d'IA de confiance (InfraTech)** ainsi que l'adoption, par son tissu économique. Enfin, elle doit également se doter de nouvelles institutions et d'organes de gouvernance, composantes essentielles de son écosystème de confiance, pour porter cette stratégie. C'est ce que nous allons étudier dans cette dernière grande partie.

---

77 Système dont la panne ou le dysfonctionnement peut avoir des conséquences dramatiques (morts, blessés graves), des dégâts matériels importants, ou des conséquences graves pour l'environnement

78 Système dont la panne ou le dysfonctionnement peut avoir un impact notable sur l'entreprise, et plus largement l'économie d'un territoire

79 Système dont la panne ou le dysfonctionnement peut avoir un impact notable sur la vie d'individus ou sur la société dans son ensemble

80 Proposition de règlement du parlement européen et du conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'IA et modifiant certains actes législatifs de l'Union)

81 https://www.economie.gouv.fr/intelligence-artificielle-au-service-des-entreprises-1

L'IA DE CONFIANCE, VÉRITABLE ÉCLAIREUR DE NOTRE SOUVERAINETÉ NUMÉRIQUE ET DE NOTRE COMPÉTITIVITÉ INDUSTRIELLE.

III. BÂTIR UNE STRATÉGIE INDUSTRIELLE PAR L'IA DE CONFIANCE

RAPPEL DES DEUX PRÉCÉDENTES PARTIES

La notion de confiance dans l'IA est centrale. Elle est au cœur des défis humanistes et donc des préoccupations d'un grand nombre d'institutions publiques et privées. A commencer par les citoyens qui s'interrogent face à des technologies amenées à se déployer partout. Mais aussi dans le monde de l'industrie pour qui l'IA est un défi majeur tout autant qu'une opportunité historique. L'Europe, malgré sa grande culture industrielle, a connu ces dernières années un phénomène de désindustrialisation que l'IA peut potentiellement contribuer à rattraper.

L'Europe peut s'appuyer, dans cette stratégie industrielle de la confiance, sur sa légitimité reconnue au niveau mondial sur les enjeux de régulation associés aux technologies. Dans le prolongement des règlements sur le numérique et les données (RGPD, DSA, DMA, DGA, DA) adoptés par la Commission européenne, la future réglementation sur l'IA (AI Act), basée sur une approche horizontale et fondée sur le niveau de risque, pourrait devenir la première pierre à l'édifice d'un écosystème de confiance mondial de l'IA.

ENJEU DE CETTE TROISIÈME PARTIE

Pour garantir cette double souveraineté industrielle et numérique, **l'Europe doit dépasser le seul cadre de la régulation, en déployant dans le même temps une stratégie industrielle volontariste suivant une approche horizontale** dans tous les secteurs. **Une coordination forte entre les États membres doit faire de cette stratégie industrielle IA de confiance l'un des piliers du « marché unique du numérique » en mettant en cohérence l'ensemble de la chaîne de valeur**, de la définition des risques acceptables au service de nos principes et valeurs, jusqu'à son implémentation industrielle.

3.1. UNE STRATÉGIE OFFENSIVE PAR LA RÉGULATION

3.1.1. Faire de la confiance un véritable avantage compétitif pour les européens

Les grands principes de la régulation européenne

En décembre 2019, Ursula Von der Leyen, fraîchement élue présidente de la nouvelle Commission européenne, annonce lors d'un discours devant le Parlement européen la volonté de réglementer l'IA dans un délai de 100 jours, en vue d'encadrer les secteurs clés de la révolution numérique. Un an et demi plus tard, la Commission européenne publie son *Livre Blanc sur l'intelligence artificielle*[82], qui expose le point de vue général de l'exécutif européen. En avril 2021 la Commission publie son projet de règlement établissant des règles harmonisées concernant l'IA, sur laquelle nous appuyons notre travail avec une réserve s'agissant de travaux toujours en cours.

Dans ce texte, **la Commission propose trois principes structurants comme fondations du projet de règlement IA :**

- **Une approche des systèmes d'IA par le risque**, avec une typologie allant des systèmes faisant courir un risque inacceptable aux systèmes faisant courir un risque moindre, voire aucun risque.

[82] https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_fr.pdf

- **L'application du règlement à des acteurs situés en dehors de l'Union européenne (extraterritorialité)**, à l'image du RGPD.
- **La neutralité du texte du point de vue technologiqu**e, qui complète sans les remplacer les législations sectorielles.

L'approche choisie par l'Europe s'enracine dans une méthode plus générale à destination des entreprises, et formalisée par l'ISO en 2015. La norme ISO 9001: 2015 codifie l'approche par le risque à l'échelle mondiale. **L'analyse des risques consiste à hiérarchiser les priorités en fonction du type de risque et de sa probabilité d'occurrence.**

L'approche par le risque a l'avantage de hiérarchiser les choses en différenciant les types de risques, et sert à adapter la réponse avec davantage de précision. Cette différenciation des usages permet aussi de mieux correspondre aux besoins du marché, à la variété des cas de figure et aux attentes des citoyens.

L'Union européenne devra veiller à ne pas suralimenter cette complexité, d'autant qu'elle doit aussi trouver un consensus entre 27 pays avec leurs législations respectives. **L'un des enjeux fondamentaux est donc la création d'un marché unique et lisible de l'IA de confiance à l'échelle de l'Union européenne.**

Les impacts attendus, autant négatifs que positifs, de la régulation

Nous avons besoin d'une régulation efficace. Elle doit soutenir l'innovation et le développement économique sans les brider. Plusieurs éléments sont à prendre en compte : d'abord, il est impératif de **réduire l'incertitude réglementaire** autour du marché de l'IA. D'autre part, **le futur règlement sur l'IA a le potentiel pour devenir un standard de *facto***, à l'instar du RGPD, en Europe et hors de ses frontières, et nous devons faire en sorte que ce soit effectivement le cas. Ensuite, **l'approche par les risques est certes complexe, mais elle est compatible avec la singularité des cas d'usage** et des expériences. Enfin, **cette régulation doit se montrer incitative**, et accompagner les acteurs voulant exploiter des systèmes à base d'IA, y compris ceux jugés « à haut risque ».

Dans sa version initiale d'avril 2021, **le projet de règlement IA (AI act) considère que la sécurité juridique mise en place par l'acte législatif sera le principal impact du texte** : *« Les fournisseurs de systèmes d'IA devraient bénéficier d'un ensemble minimal mais clair d'exigences, créant une sécurité légale et assurant l'accès au marché unique dans son entièreté. Les utilisateurs de systèmes d'IA devraient bénéficier d'une sécurité juridique sur les systèmes d'IA à haut-risque qu'ils achètent qui garantisse la conformité avec les lois et les valeurs européennes.*[83] »

En revanche, le *Center for Data Innovation*[84] propose une étude critique de l'impact financier sur les petites entreprises. En effet, **une PME qui déploie un système IA jugé « à haut risque » fera face à un coût de conformité pouvant aller, suivant la complexité et l'application, jusqu'à 400 000€.**

La Commission a émis le souhait de voir 75% des entreprises européennes utiliser des technologies et solutions à base d'IA d'ici 2030 ; mais l'étude du *Center for Data Innovation* évalue l'effet dissuasif des coûts de conformité à près de 20% d'investissements en moins. L'étude estime que les charges de conformité qui vont en découler coûteront aux entreprises européennes environ 10,9 milliards d'euros par an d'ici 2025, et 31 milliards d'ici 5 ans, ***« sans inclure les pertes d'opportunités et une probable fuite des cerveaux puisque les startups innovantes trouvent qu'il est plus facile de s'établir ailleurs.***[85] **»**

[83] Artificial Intelligence Act, COM (2021) 206 Final 2021/0106, Brussels 21.4.2021, p.94
[84] https://datainnovation.org/
[85] La Législation sur l'Intelligence Artificielle coûterait à l'économie européenne 31 milliards d'euros sur 5 ans, réduirait l'investissement dans l'IA de près de 20 pourcents, d'après un nouveau rapport,Benjamin Mueller, 2 août 2021, datainnovation.org

Pour France Digitale[86], ***"la nouvelle charge réglementaire ne doit pas décourager les fondateurs et les investisseurs en IA de s'engager en Europe"*, y compris sur les sujets « à haut risque »**. L'association déplore le manque de précisions sur la typologie des risques associés aux systèmes d'IA, mais aussi la difficulté à mettre en œuvre ces obligations pour les startups, en particulier celles opérant sur des domaines à risque (comme l'éducation ou le droit).

D'autres voient dans le futur règlement sur l'IA l'opportunité de créer de nouveaux métiers, comme celui de « responsable des risques IA » dans les structures de plus de 50 salariés utilisant une IA à haut risque. Les exigences légales augmenteront effectivement le nombre d'audits internes et externes. D'autre part, **« un nouveau profil d'ingénieur qualité, spécialiste en IA, va probablement voir le jour »**[87] pour s'assurer de la mise en conformité des produits et services avec les exigences de l'entreprise et celles du futur règlement IA, à l'instar des DPD ou DPO (Délégués à la Protection des Données ou *Data Protection Officer*) créés par le RGPD.

Concernant la réception du texte par la communauté juridique, le Dalloz[88] insiste sur la complexité de l'approche transversale, tout en se félicitant du choix : ***« la volonté de réguler non pas globalement les systèmes d'IA mais bel et bien les produits et services embarquant des systèmes d'IA s'avère être un véritable défi en termes d'articulation des normes. »***[89]

De son côté, la Stanford Law School - l'une des universités pionnières sur l'IA - salue avec emphase et enthousiasme l'initiative européenne : *"Il faut du courage et de la créativité pour légiférer dans cette matière orageuse et interdisciplinaire, forçant les entreprises américaines et chinoises à se conformer à des standards fondés sur les valeurs européennes avant de pouvoir accéder à ce marché de 450 millions de consommateurs. Par conséquent, la proposition possède un effet extraterritorial. En rédigeant l'Acte sur l'Intelligence Artificielle et en l'articulant à des normes et des valeurs humanistes dans l'architecture et l'infrastructure de notre technologie,* ***l'Union européenne fournit une direction et mène le monde vers une destination qui fait sens.*** *La Commission européenne, avec le RGPD, a déjà orchestré la création d'un standard international pour la protection de la vie privée, la protection et la souveraineté des données […]."*[90]

Face à l'enthousiasme de l'école de droit de Stanford, deux réserves peuvent néanmoins être émises sur le choix de cette approche législative. D'une part, **l'Union européenne ne doit pas se limiter à la stricte réglementation, car celle-ci pourrait apparaître comme "l'arme du pauvre"** pour entraver les succès économiques des géants américains et chinois. ***"Un second inconvénient, plus dommageable pour l'Europe, est le risque de voir l'Union européenne s'imposer et imposer à ses entreprises et à ses citoyens des règles vertueuses contraignantes sans être suivie par ses principaux concurrents, avec pour conséquence des distorsions de concurrence à son propre détriment."***[91]

3.1.2. Le règlement IA comme première pierre à l'édifice d'une UE ambitieuse

Une approche horizontale ambitieuse

Le projet de règlement sur l'IA s'inscrit dans les priorités stratégiques de la Commission européenne pour le mandat 2019-2024, résumées dans l'axe « Une Union plus ambitieuse ». Déjà en 2017, le Conseil européen enjoignait le législateur à faire preuve *« d'un sens de l'urgence face aux tendances émergentes, notamment en ce qui concerne des questions telles*

[86] https://francedigitale.org/
[87] *Responsable des risques IA, un nouveau métier au service du machine learning*, Antoine Crochet-Damais, 29 septembre 2021, journaldunet.com
[88] https://www.dalloz.fr/
[89] *Artificial Intelligence Act : avis conjoint des CEPD*, Cécile Crichton, 2 juillet 2021, Dalloz Acutalités
[90] *EU Artificial Intelligence Act: The European Approach to AI*, Mauritz Kop, Transatlantic Antitrust and IPR Developments (2021)
[91] *L'Europe comme puissance normative internationale : état des lieux et perspectives*, Laurent Cohen-Tanugi, Revue Européenne du Droit n°3

que l'intelligence artificielle. »[92]. Deux ans plus tard, le Conseil insistait sur l'importance de garantir le respect « intégral » des droits des citoyens européens en demandant une législation pertinente adaptée aux défis et possibilités qu'offre l'intelligence artificielle.

Parmi les différentes options envisagées par la Commission, **le choix s'est porté sur l'instrument législatif européen horizontal, suivant une approche proportionnée par le risque**, ainsi que l'encouragement à adopter des codes de conduite sur une base volontaire pour les systèmes d'IA hors « risque élevé ». Autrement dit, **ce n'est pas le secteur industriel ou commercial mais le produit ou service qui va déterminer l'applicabilité ou non du règlement IA, nonobstant les législations sectorielles déjà existantes.**

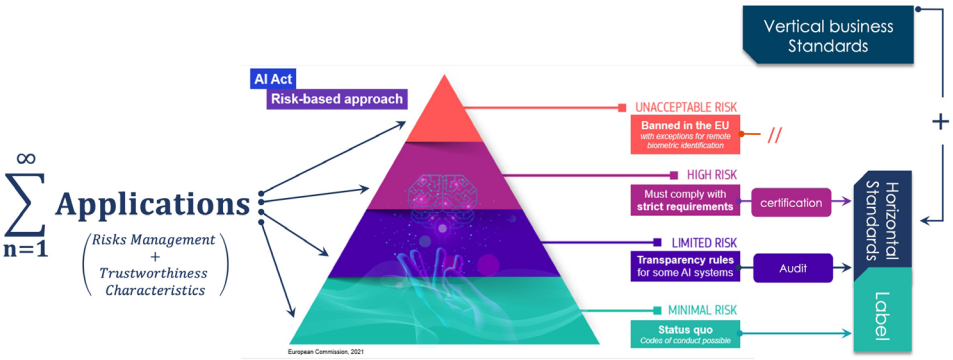
Des exigences fortes

Les grands principes de la régulation européenne sur l'IA ont d'abord été posés en 2020 dans le Livre Blanc *Intelligence Artificielle. Une approche européenne axée sur l'excellence et la confiance*[93] de la Commission européenne. **La Commission propose sept exigences pour bâtir une intelligence artificielle de confiance :**

- des contrôles humains sur le traitement des données,
- la robustesse technique et la sécurité du traitement des données,
- le respect de la vie privée et la bonne gouvernance des données,
- une attention à l'impact environnemental et sociétal,
- la diversité, la non-discrimination et l'équité,
- la transparence autour du traitement des données,
- les responsabilités associées au traitement des données.

Ces principes directeurs et les exigences techniques qui doivent en découler s'intègrent dans une approche par le risque dans le projet de règlement IA d'avril 2021[94]. Sont aussi qualifiés, les risques sur les technologies qui composent l'IA : opacité, complexité, imprévisibilité, comportement partiellement autonome. **Le législateur insiste sur les difficultés associées aux moyens de vérification technique de la conformité à la réglementation.**

La pyramide des risques



Source : Excellence et confiance en matière d'intelligence artificielle, Commission Européenne

[92] Règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législations sur l'intelligence artificielle - AI Act) et modifiant certains actes législatifs de l'union, 21.04.21
[93] COM (2020) 65 final, 19.2.2020
[94] *Proposal of a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM (2021) 206 Final 2021/0106, Brussels 21.4.2021

La version initiale d'avril 2021 du futur règlement propose une pyramide des risques[95] générés par les systèmes à base d'IA :

- **Risque inacceptable :** usages des systèmes d'IA considérés comme une menace évidente pour la sécurité ou les droits des personnes : interdit. Par exemple : les systèmes qui manipulent le comportement humain et privent les usagers de leur libre arbitre, et systèmes permettant la notation sociale par les États.
- **Risque élevé :** Le projet de règlement IA publié par la Commission propose de considérer comme à haut risque les systèmes d'IA constituant des composants de sécurité de produits tombant sous le coup d'une législation sectorielle harmonisée lorsqu'ils sont déjà soumis à une évaluation de conformité par un tiers, ainsi que les systèmes d'IA listés dans son annexe III dans divers domaines d'activité, comme par exemple :
  - Les systèmes d'IA utilisés comme composants de sécurité dans la gestion du trafic routier ou la fourniture d'énergies ;
  - Les systèmes d'IA utilisés pour déterminer l'accès ou l'affectation aux établissements d'enseignement ou de formation professionnelle, ou pour évaluer les étudiants ;
  - Les systèmes d'IA utilisés pour déterminer le recrutement, l'évaluation professionnelle, l'évolution de la carrière et le licenciement des personnes ;
  - Les systèmes d'IA destinés à évaluer l'éligibilité des personnes aux aides sociales, leur solvabilité, ou à leur attribuer une note pour l'obtention d'un crédit ;
  - Certains systèmes d'IA susceptibles d'être utilisés dans la gestion des enquêtes pénales (polygraphe, systèmes prédictifs notamment) ou des migrations, de l'asile et du contrôle aux frontières (ex : vérification de l'authenticité des documents de voyage) ;
  - Les systèmes d'IA aidant à la prise de décision des magistrats.
- **Risque limité :** systèmes d'IA avec obligations spécifiques de transparence. Exemple : robots conversationnels (*chatbots, deepfakes*[96])
- **Risque minime :** tous les autres systèmes d'IA, par exemple ceux utilisés dans les jeux vidéo, ou encore les filtres anti-spam : pas d'intervention législative, mais une incitation à l'élaboration d'un code de conduite.

Plus le système d'IA sera susceptible d'engendrer un risque (pour le consommateur, la société, l'État), plus les règles seront nombreuses, notamment au stade de la conception du système.

On retrouve certains usages de systèmes d'IA dits à haut risque dans les secteurs suivants : sécurité, finance, banque, assurance, emploi, éducation, soins de santé, transports, énergie, secteur public (asile, migration, contrôle aux frontières, système judiciaire, services de sécurité sociale).

Les systèmes d'intelligence artificielle dits « à haut risque » font l'objet d'exigences spécifiques à la hauteur de la criticité de leurs usages[97].

Par ailleurs, **le législateur européen prévoit la création d'un Comité européen pour l'IA (« the Board »), composé de représentants désignés par les États membres et de la Commission[98].**

[95] Artificial Intelligence Act : L'Union européenne invente la pyramide des risques de l'intelligence artificielle, Alexandra Bensamoun, Le Club des Juristes, 21 mai 2021

[96] Le *deepfake*, ou hypertrucage, est une technique de synthèse multimédia reposant sur l'intelligence artificielle. Elle peut servir à superposer des fichiers vidéo ou audio existants sur d'autres fichiers vidéo (par exemple changer le visage d'une personne sur une vidéo) ou audio (par exemple reproduire la voix d'une personne pour lui faire dire des choses inventées). Cette technique peut être utilisée pour créer des infox et des canulars malveillants. - Wikipedia

[97] COM (2020) 65 final, 19.2.2020, pages 21 à 26

[98] COM (2021) 206 Final 2021/0106, Brussels 21.4.2021, p.73

Tout en se félicitant de cette initiative législative, et bien que la désignation de ces représentants appartiennent à chaque État membre, **la CNIL française prend position sur quatre points jugés fondamentaux** dans un avis publié le 8 juillet 2021[99] :

- **la nécessité de tracer des lignes rouges par rapport aux futurs usages de l'IA,**
- **le défi de l'articulation avec le RGPD,**
- **l'importance d'une gouvernance harmonisée,**
- **et un accompagnement de l'innovation indispensable.**

Si les craintes d'étouffement de l'innovation et de l'expérimentation sont réelles, **« l'effet Bruxelles »[100] fait que l'Europe peut néanmoins utiliser cette régulation comme véhicule de sa puissance normative à l'international** en usant de ses capacités comme l'expose Laurent Cohen-Tanugi : *"la capacité à élaborer sa loi et à en imposer le respect sur son territoire, voire au-delà (extraterritorialité) ; la capacité à influer sur le contenu de normes (juridiques, techniques) résultant d'un processus de négociation internationale au sein de différentes enceintes multilatérales ; et la capacité à servir de modèle normatif volontaire au sein de la communauté internationale."*[101].

3.1.3. Une approche volontariste par la norme

"L'usage fait la norme et la norme fait l'usage. »[102]

RENFORCER L'INFLUENCE DE L'EUROPE SUR LES STANDARDS

Le 2 février 2022, l'Europe a présenté un plan pour renforcer son influence dans la création des standards internationaux, concernant aussi bien les batteries électriques que les services numériques. Ce plan s'inscrit dans la stratégie industrielle européenne présentée en mars 2020 dans lequel l'Europe montrait déjà une volonté de défendre ses entreprises et de prendre sa place dans la rivalité sino-américaine.

Certains observateurs redoutent une perte d'influence des entreprises européennes face aux géants industriels, ainsi qu'aux stratégies de Pékin et de Washington. Mais l'enjeu se situe plus en profondeur : le régime de Pékin pousse par exemple à la redéfinition des standards internationaux grâce à la puissance de ses acteurs numériques. On pense ici à l'action de Huawei pour modifier le protocole Internet avec sa *"New IP Initiative"*.

L'Europe souhaite donc impérativement renforcer son influence dans les organes de standardisation pour faire valoir ses intérêts et protéger les valeurs transcrites dans les normes technologiques internationales.

L'importance de la normalisation en IA

Il convient tout d'abord de clarifier la différence entre réglementation et normalisation parfois source d'incompréhension. Ainsi que le rappelle l'Association Française de Normalisation (AFNOR), « la normalisation volontaire s'efforce de détecter et d'accompagner en permanence les tendances technologiques et sociétales. Elle appréhende les besoins du marché en termes d'ouverture, d'appui à la compétitivité des entreprises françaises et de cadre de développement

[99] *Intelligence artificielle : l'avis de la CNIL et de ses homologues sur le futur règlement européen*, 8 juillet 2021, cnil.fr
[100] L'effet Bruxelles est le processus de mondialisation réglementaire unilatérale provoqué par l'extériorisation de facto (mais pas nécessairement de jure) de la législation de l`Union européenne au-delà de ses frontières par le biais des mécanismes de marché - encyclopedie.fr
[101] *L'Europe comme puissance normative internationale : état des lieux et perspectives*, Laurent Cohen-Tanugi, Revue Européenne du Droit n°3, Les chemins de la puissance européenne
[102] Olivier Dion, Digital New Deal, 2022

NOUS DEVONS
FIXER LES NORMES,
PAS LES SUBIR.[103]

[103] *We want to be a global standard-setter, not a standard-taker*, Ursula von der Leyenr

harmonieux pour les activités et les nouveaux emplois (…). Elle intervient dans un cadre largement européen et international, qu'elle contribue à bâtir. »

« La normalisation peut aussi venir en appui de la législation et de la réglementation, et contribuer ainsi à limiter l'inflation des textes législatifs et réglementaires. Tout ceci dans un rythme de changements du monde qui s'accélère. », **notamment au travers des « normes harmonisées »**[104]. **Celles-ci sont ainsi initiées par un mandat de la Commission européenne** afin de s'assurer que les produits et services respectent les prescriptions techniques de la législation correspondante. Elles représentent donc le point de rencontre entre exigences réglementaires et déclinaisons opérationnelles par le marché.

Ce choix, adopté par la Commission européenne dans le domaine de l'IA, confère ainsi aux normes une importance toute particulière à la fois pour le futur marché européen de l'IA, mais aussi pour la confiance des citoyens européens.

Prenons l'exemple des systèmes d'IA à haut risque. Ils seront soumis à des exigences réglementaires et à un examen de conformité par des organismes notifiés[105]. Ces exigences obligatoires, portant notamment sur des attributs de la confiance, ainsi que leur évaluation après analyse des risques, seront détaillées à travers des normes harmonisées, sur lesquelles les critères de certification s'appuieront. Pour tous ces systèmes, qui comprennent une grande partie des applications industrielles, les normes harmonisées constitueront donc les exigences techniques nécessaires pour être conforme à la réglementation (art. 43 AI act). Et, dans le même temps, la prise en compte des normes harmonisées peut également engendrer une présomption de conformité (art. 42 AI Act).

LE MODÈLE INSPIRANT DE L'AÉRONAUTIQUE

Plusieurs secteurs à haut risque, comprenant des systèmes critiques, sont en train d'ouvrir la voie sur l'intégration de l'IA de confiance à leur plan de marche industriel.

Il apparaît souvent que plus le niveau de risque et de responsabilité des acteurs économiques est grand, plus la coopération de filière est importante. **L'écosystème de confiance de l'aéronautique**, de manière générale (y compris hors IA), **pourrait servir de modèle aux autres filières. Néanmoins, le coût d'un tel cadre est particulièrement élevé**, les acteurs économiques ont un intérêt limité à le proposer concernant l'IA, si le risque ne l'impose pas. **C'est au monde politique d'aider à mettre en place les écosystèmes de confiance dans l'IA, en concertation avec les industriels.** Des modes différenciés d'intégration de l'écosystème d'IA de confiance, en fonction du niveau de risque, pourraient être mis en place, en adéquation avec les spécificités de chaque filière.

Certification et régulation coordonnées

Dans l'aéronautique c'est la *Federal Aviation Administration* (FAA) qui fait autorité pour la certification aux États-Unis. L'*European Union Aviation Safety Agency* (EASA) remplit un rôle similaire pour l'Europe. La définition des normes se fait en concertation entre les industriels du secteur et les autorités de certification. Les deux agences proposent, séparément, les mêmes normes certifiantes en régulations auprès du Congrès américain pour la FAA et du Parlement européen pour l'EASA. Les deux agences s'accordent au préalable, de manière réciproque, afin de coordonner les différents textes. **La coordination de filière permet à un constructeur certifié en Europe de pouvoir faire voler son avion aux États-Unis.** Les autres agences mondiales sont également impliquées dans le processus.

[104] [Les normes harmonisées] servent à prouver que les produits ou services respectent les prescriptions techniques de la législation européenne correspondante. Elles sont en général facultatives mais leur application laisse présager une forte présomption de conformité aux exigences techniques réglementaires. Cependant, un organisme peut choisir de ne pas les appliquer en préférant une autre solution technique que celle préconisée dans une norme harmonisée. https://8m-management.com/

[105] *Un organisme notifié est une organisation désignée par un État membre de l'UE (ou par d'autres pays dans le cadre d'accords spécifiques) pour évaluer la conformité de certains produits avant leur mise sur le marché*, Commission Européenne, Organismes notifiés (europa.eu)

Normalisation coordonnée

La construction d'un avion repose sur des disciplines d'ingénierie très variées. Par exemple, en dehors du domaine de l'IA, le document DO-178C permet d'approuver un logiciel dans un avion. Ce document est écrit conjointement par la Radio Technical Commission for Aeronautics (RTCA) pour les États-Unis et par l'*European Organisation for Civil Aviation Equipment* (EUROCAE). Il est ensuite repris par la FAA et par l'EASA pour servir de base à la certification. **La FAA et l'EASA font confiance à l'ensemble du processus.**

Intégration de l'IA

En 2019, le groupe de travail EUROCAE (WG-114) a été mis en place en Europe pour préparer la future intégration de l'IA dans les avions, avec le SAE (G34) comme équivalent outre-Atlantique. L'objectif des groupes conjoints de travail est de produire un texte de standardisation homogène pouvant être utilisé par l'EASA et la FAA, puis au niveau mondial. **En 2024 ou 2025, l'aéronautique devrait disposer des premières normes industrielles mondiales de filière pour l'IA de confiance.** L'IA est traitée par le secteur de l'aéronautique comme un sujet classique de normalisation et de certification, elle ne fait pas l'objet d'un traitement spécifique.

La difficile articulation entre normes horizontales et sectorielles

Le débat entre la normalisation horizontale et la normalisation sectorielle (ou métier) n'est pas nouveau. Historiquement les différents secteurs d'activités n'ont d'ailleurs que peu d'interactions : les spécificités sectorielles l'emportant largement, une approche en silo est préférée par défaut.

Néanmoins, une fois posé le postulat suivant sur l'IA : ***« une technologie à vocation transversale ne peut être efficacement réglementée que par des règles horizontales qui fournissent des solutions aux défis communs. »*** *(Thierry Breton)*. La question de l'articulation entre normalisation horizontale et sectorielle devient incontournable, tout comme la prise en compte des problématiques de positions dominantes de certains acteurs du numérique, présentant l'avantage concurrentiel d'être par nature multi-sectoriels.

Dans ce contexte, la coopération entre différentes filières industrielles et différentes typologies d'acteurs (Grands groupes, PME, startups, laboratoires de recherche, universités, institutionnels) **devient cruciale, ainsi que le partage d'un socle horizontal fédérateur pour la norme**, comme proposé dans la stratégie de normalisation française en IA (AFNOR).

L'effort devra donc être autant sectoriel (santé, aéronautique, éducation, etc.) pour refléter les spécificités et les besoins des filières et de leurs cas d'usage, que trans-sectoriel pour faciliter l'émergence d'un écosystème de confiance complet (outils d'ingénierie, régulation, certification, contrôle, enquête, etc.).

Plusieurs filières industrielles travaillent d'ores et déjà à normer ou standardiser leurs usages d'une IA responsable, fiable et sûre. Par exemple, le secteur de l'aéronautique a démarré des travaux dès 2019 avec l'EUROCAE/SAE, puis publié une première feuille de route en 2021. En 2019, la *National Medical Products Administrations* chinoise publie son guide d'intégration de l'IA en santé. En 2020, les assureurs américains ont adopté les principes de l'OCDE sur l'IA, et les banques américaines ont commencé leurs discussions de formalisation début 2021. En octobre 2021, c'est au tour de l'Office of Science and Technology Policy de la Maison Blanche de lancer un appel à la société civile pour encadrer les usages d'IA potentiellement nuisibles pour la société[106].

[106] Ibid EY-Parthenon, p.7

Tableau : organismes de standardisation

INDUSTRIE	EUROPE	ÉTATS-UNIS	CHINE
Automobile	ACEA, EEA, autorités nationales	NHTSA (FMVSS), FTA, AV Comprehensive Plan, DOT, EPA	CCC (GB standards)
Ferroviaire	ERA, autorités nationales	RSIA, FRA, FTA	CRCC (GB standards)
Aéronautique	EASA, autorités nationales	FAA, EPA	CAAC
Banque	EBA, autorités nationales	FED, FDIC	CBRC
Assurance	AEAPP	NAIC	CIRC
Pétrole & gaz	EUOAG, CEER, autorités nationales	PHMSA, EPA, FERC	NEA, SERC, CAEA
Minier	CEER	EPA	NEA, National Mining Law
Power & utilities	CEER, autorités nationales	PHMSA, EPA	NEA
Santé	AEM, autorités nationales	HSS, FDA	NMPA

Dans le même temps, les travaux s'intensifient au niveau des instances de normalisation européennes (CEN / CENELEC) ou internationales (ISO, IEEE). Plusieurs pays, comme la France et l'Allemagne, ont également publié des stratégies pour parvenir à répondre à cet enjeu, dans des délais contraints.

Toutefois, depuis ses débuts, l'Europe a su se positionner dans les organes de normalisation des grandes filières industrielles, comme l'aéronautique ou la santé. **Étant donné la nature transverse et donc trans-sectorielle de l'IA, une « technologie d'usage général » , nous devons nous donner les moyens de peser au sein des instances normatives concernées** et éviter de laisser d'autres pays (Chine, États-Unis) ou les grands acteurs du numérique décider seuls des standards. A ce sujet, un rapport de l'Ecole de guerre économique (EGE) à paraître s'intéresse aux stratégies d'influence dans le domaine de la normalisation, avec un focus sur le cas de l'IA.

Nous recommandons donc de soutenir la politique engagée par la Commission européenne faisant de la normalisation une priorité de souveraineté industrielle et numérique, ainsi que la mise en œuvre d'un label européen pour les applications non catégorisées comme à « haut risque ».

LA STRATÉGIE FRANÇAISE DE NORMALISATION EN IA

à l'initiative du Grand Défi IA de confiance de la Stratégie Nationale en IA

L'État français a mandaté l'AFNOR, dans le cadre du Grand Défi sur l'IA de confiance, avec **une mission : « créer l'environnement normatif indispensable au déploiement d'une IA de confiance »**

Après une enquête menée à l'été 2021 auprès de plus de 250 acteurs de l'industrie de la recherche et de la société civile, AFNOR a publié en mars 2022 la feuille de route nationale pour doter ce secteur stratégique de nouvelles normes volontaires et diffuser les bonnes pratiques.

Elle s'articule autour de **6 priorités** :

- **Développer les normes portant sur la confiance :** les caractéristiques prioritaires à normaliser retenues sont la sécurité, la sûreté, l'explicabilité, la robustesse, la transparence, l'équité (dont non-discrimination). Chaque caractéristique devra faire l'objet d'une définition, d'une description du concept, des exigences techniques et des métriques et contrôles associés.
- **Développer les normes sur la gouvernance et le management de l'IA :** L'IA génère de nouvelles applications, qui comportent toutes des risques. Ces risques sont d'origine diverse : mauvaise qualité des données, mauvaise conception, mauvaise qualification, etc. Une analyse des risques pour les systèmes à base d'IA est donc essentielle, pour ensuite proposer un système de management des risques.
- **Développer des normes sur la supervision et le reporting des systèmes d'IA :** il s'agit de s'assurer que les systèmes d'IA sont contrôlables (capacité à reprendre la main), que l'humain pourra reprendre la main aux moments critiques où l'IA sortira de son domaine de fonctionnement nominal.
- **Développer des normes sur les compétences des organismes de certification :** il reviendra à ces organismes de s'assurer, non seulement que les entreprises ont mis en place des processus de développement et de qualification des systèmes d'IA, mais également que les produits sont bien conformes aux exigences, notamment réglementaires.
- **Développer la normalisation de certains outils numériques :** l'un des enjeux de l'IA consiste à disposer de simulations reposant sur des données synthétiques, et non plus sur des données réelles. Les normes devront rendre ces données fiables.
- **Simplifier l'accès et l'utilisation des normes :** afin de faire vivre cette stratégie et de l'ajuster en cours de route, une plateforme de concertation sera mise à disposition pour animer l'écosystème français.

3.1.4. Garantir l'équilibre entre réglementation, normes et innovation par les sandboxes

La définition du cadre juridique de la régulation : un défi pour les autorités publiques et les régulateurs

Premièrement, adopter une définition à des fins de régulation s'avère complexe. L'IA est un sous-ensemble de disciplines riches et évolutives dans le temps (cf. chapitre 1). Or il est nécessaire en droit de définir des bornes d'application claires, lisibles et respectueuses du principe d'égalité devant la loi. La négociation actuelle (cf. annexes I, II et III AI Act) semble ainsi osciller entre, d'une part, une définition extensive, presque « diplomatique », car intégratrice de toutes les sensibilités de l'IA, et, d'autre part, des demandes de réduction du champ de cette définition au motif de ne pas imposer des obligations à des logiciels existants, parfois depuis de nombreuses années, et pour lesquels le cadre juridique actuel est satisfaisant.

Deuxièmement, il n'est pas assuré que l'ensemble des principes d'encadrement unifiés demeurent pertinents à l'épreuve du temps. A titre d'exemple, la première version du texte, s'agissant de l'IA fondée sur les données, était principalement indexée sur le paradigme de l'IA supervisée (*supervised learning*). Or, au même moment, le paradigme des modèles auto-supervisés géants prenait son envol et connaissait ses premières applications commerciales. Depuis, diverses prises de position recommandent de mieux équilibrer les responsabilités notamment s'agissant du développement et de la mise à disposition des modèles génériques

qualifiés « d'usage général ». Mais, cela interroge sur le fond : comment encadrer à bon escient des classes de modèles aux caractéristiques et aux propriétés encore mal connues ? Doit-on et peut-on laisser le fournisseur / utilisateur du système à usage générique assurer seul la conformité avec les obligations du règlement alors qu'il n'a pas forcément toute la maîtrise du système conçu avant lui ?

Pour un texte en cours d'élaboration, ce type d'amendement ne pose pas de difficulté procédurale. Mais, cela interroge sur le fond : comment encadrer à bon escient des classes de modèles aux caractéristiques et aux propriétés encore mal connues ? **A contrario, une fois que le texte sera publié, qu'adviendra-t-il quand de nouvelles évolutions substantielles de l'IA surviendront ?**

Les bacs à sable réglementaires ou l'indispensable espace juridique adapté à l'innovation

Les délais de négociation, d'adoption ou d'amendement des règlements européens sont tels qu'il faut anticiper toute inadéquation juridique substantielle à venir. Le recours à des actes délégués, prévus par l'AI Act, peut introduire de la souplesse, mais de manière limitée. Dès lors, l'organisation d'espaces juridiques adaptés à l'innovation technologique s'avère pertinente pour concilier réglementation et innovation.

Les bacs à sable réglementaires d'innovation (sandboxes) en sont la meilleure incarnation possible. Ils possèdent cinq caractéristiques principales selon l'OCDE :

- **La démonstration du caractère véritablement innovant du bien ou du service souhaitant intégrer le programm**e, qu'il s'agisse d'évaluation en conditions réelles de technologies nouvelles ou d'évaluation de nouveaux usages de technologies existantes ;
- **L'identification d'un intérêt économique ou social** (amélioration des fonctionnalités des biens et services, renforcement de la qualité, baisse de prix, etc.) ;
- **La détermination de limites temporelles, géographiques ou sectorielles** évitant d'affaiblir la portée d'une réglementation en ouvrant une dérogation permanente et générale ;
- **Des mécanismes de sauvegarde**, qui peuvent porter sur la protection des consommateurs, la sécurité ou la protection des données personnelles, etc. afin d'éviter tout impact systémique, il s'agit d'imposer des conditions relatives au niveau de diffusion du bien ou du service ;
- **La justification de l'inadaptation de certains éléments de la réglementation en vigueur (ex : RGPD)**, qui peuvent faire obstacle à l'évaluation en conditions réelles.

La réglementation IA telle que proposée limite fortement l'intérêt des bacs à sable réglementaire

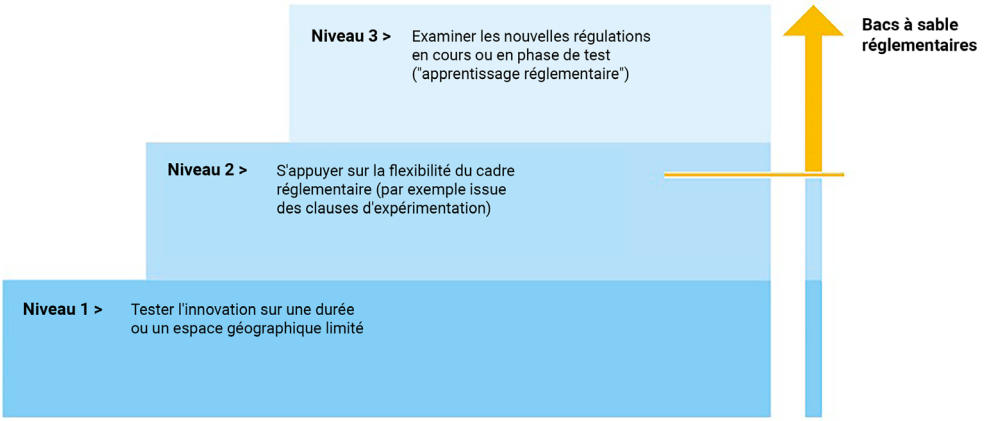
Or, la réglementation IA (AI Act), dans sa version initiale, n'offre qu'une version particulièrement prudente et contraignante des bacs à sables réglementaires, les vidant presque leur substance et limitant ainsi fortement leur impact :

- **Les tests en conditions réelles, même maîtrisées, sont proscrits** au bénéfice du seul environnement contrôlé ;
- **Le bac à sable ne peut porter que sur des démonstrateurs** et pas sur des biens ou services déployés en conditions réelles ;
- **Les bornes temporelles sont très contraintes et tout assouplissement réglementaire temporaire justifié et motivé est proscrit.**

Ce choix peut se justifier par la préservation du modèle « protecteur » de l'Europe. Néanmoins, à contrario, c'est se priver de dispositifs qui au contraire, par leur agilité, aident à le garantir.

Les bacs à sable réglementaires présentent en effet de nombreux atouts : aider les régulateurs à affiner leur doctrine, laisser le temps aux autorités publiques d'ajuster les textes, accompagner la montée en compétence de l'écosystème d'audit et de conformité, élaborer de nouvelles normes volontaires spécifiques et plus adaptées, garantir le développement en conformité à la réglementation.

Bacs à sable réglementaires



Source : *German Federal ministry for economic affairs and climate action*

Nous recommandons de promouvoir des bacs à sable réglementaires *(sandboxes)* suivant les critères promulgués par l'OCDE afin de ne pas se priver d'un dispositif qui par son agilité, à contrario des idées reçues, aide à garantir la future mise en œuvre de la réglementation, et donc la protection des citoyens européens.

3.2. UNE STRATÉGIE INDUSTRIELLE PAR LA COOPÉRATION

3.2.1. Accroître les efforts de RDI et de formation en IA de confiance[107]

L'UE bien positionnée en recherche, mais en retard dans l'innovation

L'Union européenne dispose d'une recherche fondamentale de qualité en IA avec des universités et des scientifiques réputés et classés parmi les meilleures du monde, mais affiche un retard notable dans la course à l'innovation, ainsi qu'en matière de soutien aux startups, ceci dans un contexte où l'IA s'avère l'un des domaines de recherche qui suscitent le plus d'intérêt de la part de la communauté scientifique (cf. édition 2021 du rapport sur l'état de la science[108] de l'UNESCO).

Les États-Unis sont en revanche plus prompts à valoriser les connaissances issues de la recherche en innovation et en création d'entreprises spécialisées en IA.

L'analyse des statistiques sur la période entre 2010 et 2021, issues du AI Index Report 2022[109], est en effet éclairante.

- **Publications :** l'Europe et UK, (env. 19 %) sont classés devant les États-Unis (env. 14 %) et derrière la Chine (env. 31 %) en nombre de publications dans les journaux scientifiques. Dans le même temps, son influence scientifique, calculée par un indicateur de citations

[107] RDI : Recherche & Développement et Innovation
[108] *UNESCO Science Report: the Race Against Time for Smarter Development*, 2021
[109] *Artificial Intelligence Index Report 2021*, Stanford University

CRÉER UN COUCHE INTERMÉDIAIRE INFRATECH DE CONFIANCE DANS L'IA, POUR ENDIGUER ET SURTOUT CONQUÉRIR LE MARCHÉ DES *BIG TECH*.

d'articles issus des journaux scientifiques les plus prestigieux, illustre une baisse notable sur la période entre 2010 et 2021 (21% contre 26%). Toutefois, cette tendance est encore plus marquée pour les États-Unis (25% en 2010 contre 17.5 % en 2021) ; la Chine bénéficiant au contraire d'une hausse importante de 21 % à 28%.

- **Brevets :** Dans un contexte où le nombre de brevets déposés a cru de façon exponentielle depuis 2015, les États-Unis restent largement en tête dans la course aux brevets octroyés (39.5 % en 2021) ; ceci avec une baisse notable depuis 2010 (65 %). Puis, suivent loin derrière l'Europe et UK (7.5 % % en 2021 contre 11 % en 2010) puis la Chine (6% en 2021 contre moins de 1 % en 2010). Cette tendance, si elle se confirme, placera l'Europe et UK en 3e position sous peu.

Fait notable : Les États-Unis dominent aussi largement les librairies logiciels open source en IA. Notons que la librairie française *Scikit-Learn* développée par des équipes Inria Paris-Saclay se classe aujourd'hui en 5e position, après avoir longtemps occupé la 2nde place du classement.

Ce panorama du continuum entre recherche et innovation est complété par un rapport de la Banque Européenne d'Investissement (BEI)[110] publié en 2021 sur l'IA et la Blockchain. Les États-Unis représentent en 2019 environ 20 milliards ($) d'investissement, soit approximativement 65 % du volume mondial, contre 5 milliards ($) pour la Chine et 2 milliards ($) pour l'Europe. Dans le même temps, les États-Unis disposent d'une avance indéniable en nombre d'entreprises spécialisées, plus de deux fois supérieur à la Chine en seconde position. L'Union européenne n'abrite "que" trois fois plus de petites entreprises par rapport au Royaume-Uni seul, alors qu'elle comporte 27 pays.

La RDI en IA de confiance s'organise, et reste une véritable opportunité pour l'Europe

Depuis quelques années, de nombreuses initiatives ont été lancées dans le monde pour soutenir la RDI sur une IA digne de confiance, responsable, éthique et sûre.

Sans être exhaustif, nous pouvons l'illustrer en citant certaines d'entre elles :

- Aux **États-Unis**, la DARPA *(Defense Advanced Research Projects Agency*) a initié des programmes sur l'explicabilité de l'IA - un des attributs de la confiance nécessaire à l'interaction entre humain et machine – ou sur l'autonomie des systèmes à base d'IA, notamment la conformité à un domaine d'emploi.
- Dans le même temps, le **Canada** s'est mobilisé, autour de la déclaration de Montréal, et de programmes comme DEEL (*DEpendable and Explainable Learning*) en partenariat avec la France.
- La **France** a également décidé en 2018 dans le cadre de sa Stratégie Nationale en IA et de sa 2e phase lancée en 2021, de soutenir la recherche et l'innovation en IA de confiance autour d'un programme de recherche industrielle comme *"le grand défi sur la sécurisation, la fiabilisation et la certification des systèmes à base d'IA, ou grand défi sur l'IA de confiance pour l'industrie"*, (dont nous préciserons le contenu ultérieurement dans ce rapport) ou de l'Institut Interdisciplinaire en IA (3IA) ANITI basée à Toulouse.
- En **Allemagne**, l'institut de recherche technologique Fraunhofer IAIS[111] pilote un programme KI-Zertifizierung sur la certification de systèmes à base d'IA. Le DFKI[112] conduit par exemple des recherches dans le cadre du *« Certlab »*. Enfin, les instituts de normalisation (DIN[113] et VDE[114]) ont publié une feuille de route en normalisation.

[109] *Artificial intelligence, blockchain and the future of Europe: How disruptive technologies create opportunities for a green and digital economy*, European Commission/European Investment Bank, https://www.eib.org/attachments/thematic/artificial_intelligence_blockchain_and_the_future_of_europe_report_en.pdf
[110] Institute for Intelligent Analysis and Information Systems
[111] Deutsches Forschungszentrum für Künstliche Intelligenz
[112] Deutsches Institut für Normung
[113] Fédération allemande des industries de l'électrotechnique, de l'électronique et de l'ingénierie de l'information

- Au **niveau européen**, l'initiative Etami supporte le développement d'une IA éthique, le réseau d'excellence européen TAILOR regroupe plus de cinquante partenaires autour de recherches sur les fondements de l'IA de confiance ou les futures *Testing and Experimentation Facilities* (TEF) ambitionnent, dans le cadre de Digital Europe, de faire émerger des sites de références pour tester et expérimenter des solutions d'IA, notamment leurs conformités à la future réglementation européenne.

Avec près de 75 % des publications à la conférence FACCT (fairness, accountability and transparency) en 2021, **l'Amérique du Nord semble toutefois la plus dynamique en recherche** en comparaison avec l'Europe et l'Asie Centrale (17 % En 2021) et la zone Asie Pacifique (moins de 5 %), surtout relativement à sa 3e position précédemment mentionnée en nombre de publication dans les journaux scientifiques.

La réglementation IA porte tout son sens pour protéger les valeurs et principes européens. Toutefois, le risque d'accroître notre retard en matière d'innovation n'est pas à exclure. Un équilibre entre régulation et innovation est donc indispensable. Pour ce faire, **nous recommandons d'accroître les efforts d'investissements européens en recherche fondamentale et appliquée, en formation (notamment dans les domaines de l'ingénierie de la donnée et de la connaissance, de l'ingénierie algorithmique) et en innovation sur l'IA de confiance pour soutenir la vision politique de l'Europe en IA.**

3.2.2. Créer une InfraTech de l'IA de confiance

« Créer un couche intermédiaire InfraTech de confiance dans l'IA, pour endiguer et surtout conquérir le marché des *Big Tech*. »

Le marché du « Numérique de confiance » en est à ses débuts. La stratégie de « Cloud de confiance » consiste pour l'instant principalement à encadrer juridiquement les relations contractuelles avec les *hyperscalers*[114] afin de limiter l'extraterritorialité du droit américain. Cette stratégie défensive, parfois décriée par les militants de la souveraineté numérique, est une approche pragmatique consistant à favoriser le multi-cloud comme opportunité commerciale pour pénétrer le marché, le temps que les offres européennes soient perçues comme pleinement compétitives par nos entreprises.

Pour que la stratégie « Cloud de confiance » s'accomplisse, nous devons faire levier sur ses deux corollaires à savoir la « Data de confiance » et l'« IA de confiance » qui peuvent selon nous créer une faille de marché par la réglementation et la coopération. C'est tout l'objet des notes dédiées du think-tank Digital New Deal, dont la vocation est de faire du « Numérique de confiance » non plus une approche défensive, mais bien une stratégie industrielle offensive face aux oligopoles actuels (ces derniers ayant trusté le marché via leurs offres cloud « packagées »[115]).

L'avance technologique et commerciale des *hyperscalers* est telle, et leur capacité financière est si grande, que nous n'avons d'autre choix que de « contenir » dans un premier temps leur expansionnisme en créant une couche intermédiaire entre le cloud et les applications, pour ensuite se servir de ce « smart middleware » ou « InfraTech » comme base de conquête du marché. A travers cette logique de « *containment* », nous pourrions alors « réintermédier » le tissu d'offres technologiques souveraines avec la demande marché. Évidemment tout cela exige de la coordination et surtout la mutualisation des investissements pour bénéficier d'un socle commun et compétitif d'outils et services capables de capter la valeur socio-économique. **C'est ce que nous nommons l'InfraTech, qui constitue le ciment des enjeux entre Cloud, Data**

[114] *Hyperscaler* : l'hyperscale est le procédé qui consiste à mutualiser les ressources serveurs (cloud computing). Un *hyperscaler* est une entité qui propose ce service comme Scaleway, Switch, Alibaba, IBM, QTS, Digital Realty Trust, Equinix, Oracle, Facebook, Amazon Web Services, SAP, Microsoft ou Google, etc. - Augustareeves.fr

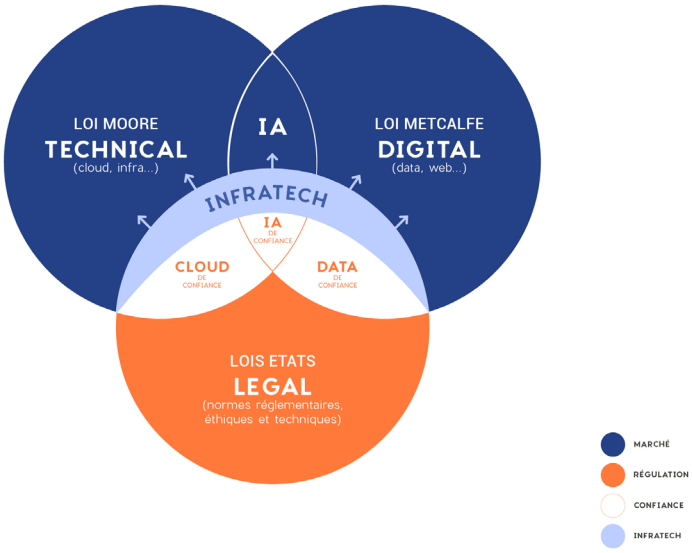
[115] *Antitrust : OVH a porté plainte contre Microsoft devant la Commission européenne*, Siècle Digital, avril 2022

CRÉER L'INFRATECH,
SOCLE DE LA CONFIANCE,
ET LEVIER DE SCALABILITÉ
POUR LA FRENCHTECH
ET L'INDUSTRIE
EUROPÉENNE.

et IA. Sans cet investissement d'infrastructure, nous nous condamnons telles les Danaïdes à « jeter l'argent dans le sable[116] » pour reprendre l'expression de Bruno Le Maire.

Pour une InfraTech de l'IA de confiance

ÉCOSYSTÈME DE CONFIANCE (CLOUD + DATA + IA) = AUTONOMIE STRATÉGIQUE



Comme nous l'avons expliqué précédemment, l'IA est un ensemble de technologies complexes pour lesquelles les compétences sont rares, variées, et le débat entre internalisation et externalisation incessant. En conséquence, outre le besoin de données et de connaissances (par exemple issues des *data spaces*, comme nous le verrons plus loin), **de nombreuses industries reposent aujourd'hui principalement sur des plateformes d'outils logiciels et des méthodes de développement, ou d'ingénierie, fournis par des prestataires spécialisés.** Ces outils et méthodes permettent une scalabilité à moindre coût pour les startups, PME et grands groupes souhaitant développer des produits et services à base d'IA.

Toutefois, **la confiance requiert un ensemble d'outils et de méthodes spécifiques, ou InfraTech, pour être déployée, et réduire le coût de conformité pouvant aller, suivant la complexité et l'application, jusqu'à 400 000€, ce qui constitue une impasse économique pour les startups et les PME.**

Cette InfraTech doit être interopérable avec les autres chaînes de conception (MLOps[117], ModelOps et Systèmes), et permettre d'analyser les risques et de concevoir, valider et contrôler, y compris en fonctionnement, le système et ses attributs de confiance pour une application donnée. **L'IA de confiance sera en grande partie développée sur ces outils et méthodes, qui deviendront les constituants de base de l'InfraTech de l'IA de confiance.**

Cette InfraTech repose sur une approche de l'ensemble de la chaîne de valeur « donnée, connaissances, algorithme et système », en cohérence et en complément des initiatives existantes au niveau européen (Gaia-X, Simpl, etc.), et doit donc s'attacher à solutionner les enjeux de chacune de ces composantes, en adressant :

- **Les méthodes et processus** (vis-à-vis de spécifications et d'exigences fonctionnelles et non fonctionnelles des composants et systèmes) : il s'agit d'un prérequis aux développements

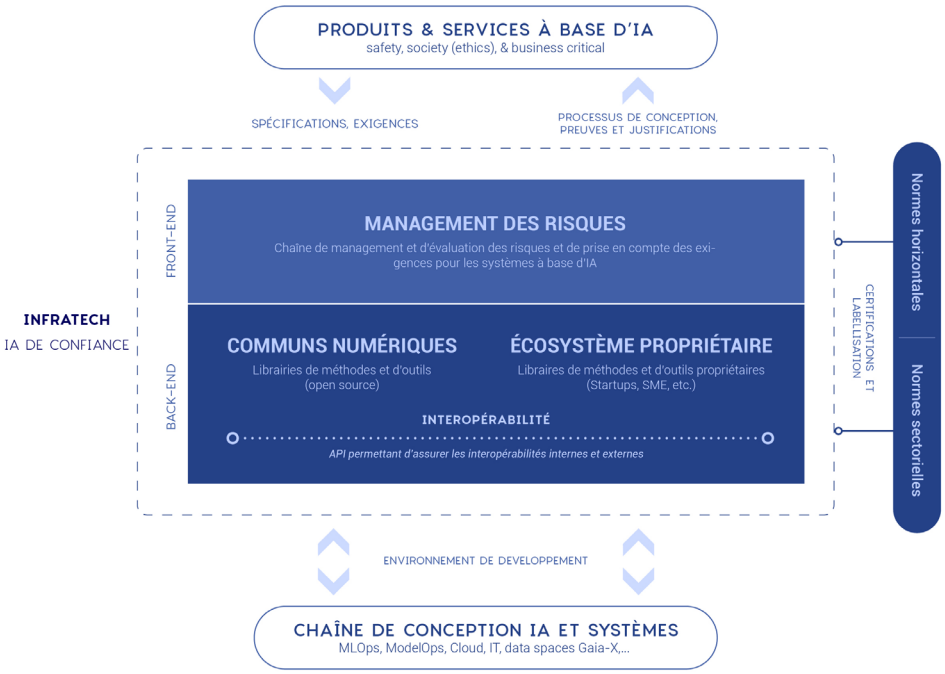
[116] *Déclaration du Ministre de l'économie et des finances, sur les efforts du gouvernement en faveur de l'innovation*, à Paris le 19 novembre 2019.
[117] MLOps ou ML Ops est un ensemble de pratiques qui vise à déployer et maintenir des modèles de machine learning en production de manière fiable et efficace.

d'applicatifs à base d'IA de confiance. Cette action doit donc être réalisée dans une phase préliminaire, avec une évaluation des risques spécifique au produit ou au service. Elles permettent également de porter une vision construite vis-à-vis des normes et standards horizontaux et verticaux et donc d'assurer la conformité.

- **Les environnements pour la constitution de bases de données et connaissances qualifiées :** la fonction réalisée par le système est très fortement dépendante de la base de données d'apprentissage et des connaissances « métier ». **Il est donc essentiel de disposer d'environnements d'« ingénierie des données et des connaissances »**, produisant des bases qualifiées en fonction des exigences et du domaine d'emploi. Cette approche doit être couplée à des technologies de validation et vérification (liste non exhaustive) de biais, de qualification d'annotation, de représentativité par rapport aux domaines d'emploi ou de détection d'attaques par empoisonnement[118], etc.

- **Les environnements pour la conception, la validation, la caractérisation et la vérification de composants d'IA et de systèmes à base de ces mêmes composants** : d'une part au travers d'outils d' « ingénierie algorithmiques » afin d'accompagner la démarche de conception et d'intégration systèmes, d'autre part au travers d'outils de validation et de vérification de propriétés (modèles, apprentissage, dynamique) comme la robustesse ou la stabilité[119] numérique (liste non exhaustive : détection de faux positifs, altération de données, réseaux adverses, satisfiabilité, régime d'incertitude, etc.). La conception de modèle de confiance par construction ou « design » doit également être prise en compte.

Le schéma ci-dessous décrit les principales composantes de l'InfraTech, ainsi que ses interfaces avec les normes, les autres chaînes de conception, et les applications métier :



[118] Les attaques par empoisonnement visent à modifier le comportement du système d'IA en introduisant des données corrompues en phase d'entraînement (ou d'apprentissage). Elles supposent que l'attaquant soit en mesures de soumettre des données à utiliser lors de l'entraînement du système d'IA. - CNIL

[119] Les attaquants utilisent une instabilité de l'algorithme entraînant une mauvaise classification ou une perception erronée de l'image à identifier (par exemple, un panneau de signalisation « limitation à 50 km/h » au lieu d'un « stop »). Ainsi, les outils de validation et de vérification des propriétés de stabilité sont d'importances majeures dans le domaine de la cybersécurité des IA.

InfraTech IA de confiance

Si l'Europe souhaite développer ses propres solutions d'IA de confiance et être compétitive, les outils et méthodes d'ingénierie, d'inspection et d'évaluation seront un élément déterminant de la chaîne de valeur de la confiance, qu'il faut absolument maîtriser, ainsi qu'un vecteur intrinsèque de diffusion des valeurs européennes.

LE PROGRAMME CONFIANCE.AI

Cette initiative du Grand Défi IA de confiance de la Stratégie nationale française en IA, piloté par l'IRT SystemX, associe des partenaires académiques et industriels (Air Liquide, Airbus, Atos, CEA, Inria, Naval Group, Renault, Safran, IRT Saint-Exupéry, IRT SystemX, Sopra Steria, Thales, Valeo), tous fédérés autour d'une même ambition : **concevoir un environnement de développement pour l'IA de confiance.**

Cet environnement sera composé de briques technologiques interopérables avec les propres ateliers d'ingénierie des partenaires industriels permettant ainsi de composer une chaîne outillée, des méthodes et des guides de bonnes pratiques répondant aux fonctionnalités nécessaires pour concevoir, valider, déployer et maintenir des systèmes critiques à base d'IA dans l'ensemble des filières, tout en tenant compte des contraintes réglementaires et normatives.

La réalisation de systèmes critiques à base d'IA nécessite en effet de revisiter les ingénieries classiques (ingénierie des données et des connaissances, ingénierie algorithmique, ingénierie logiciel et ingénierie système) et de les enrichir. **Cela requiert de s'assurer de la conformité du système aux besoins et contraintes du client, définir des méthodes et des outils pour sécuriser l'ensemble des phases de conception, mais aussi garantir des propriétés de confiance** *(cf. attributs exposés en chapitre 1.2)* tout au long du cycle de vie.

Le programme Confiance.ai s'articule autour de 7 projets indépendants et rassemble aujourd'hui plus de 40 partenaires (grand-groupes, startups, PME, laboratoires de recherche à proportion égale).

Nous recommandons de financer la création d'une plateforme logicielle européenne d'outils et de méthodes pour le développement d'IA de confiance (InfraTech), et de soutenir l'écosystème des fournisseurs de solutions. L'InfraTech de l'IA de confiance devra être développée en complément de celle en cours de lancement dans le partage de donnée (« InfraTech de la Data de confiance ou *Smart Middleware for Data-Sharing* ») associé au développement de Gaia-X, permettant la conformité à la future réglementation européenne, ainsi qu'aux futurs normes et standards associés.

Afin d'accélérer les développements, d'amplifier les dynamiques de coopération et de favoriser l'émergence d'un marché unique du digital en Europe, **cette réalisation doit s'appuyer sur nos « forces » industrielles, notamment en tirant les développements par leurs cas d'usage, et sur les initiatives déjà en cours.**

Développer les « Communs Numériques » de l'IA de confiance

Un **« commun numérique » est un patrimoine numérique appartenant à une communauté, et dont l'usage peut être libre. Ce commun peut être composé d'infrastructures, de données, de librairies, de logiciels, d'outils, de méthodes, etc., gouvernés et protégés par la communauté.**

Si l'Europe souhaite investir dans l'InfraTech, deux approches sont envisageables :

1. **Créer un champion européen unique de l'infrastructure de confiance** capable de concurrencer les GAFAM
2. **Créer un écosystème pour l'infrastructure de l'IA et de la Data de confiance (InfraTech)** avec le concours de multiples acteurs publics et privés, dont les startups européennes.

L'approche du champion européen unique a déjà été tentée sur d'autres sujets technologiques, sans succès probant. Les moyens dont disposent les géants et les fonds américains et chinois sont conséquents, et il est pour le moment difficile de rivaliser frontalement. Les récentes confrontations entre les institutions publiques américaines et chinoises et leurs propres géants numériques montrent également que cette approche pose d'autres problèmes de souveraineté pour les États concernés. « Les communs numériques offrent une occasion unique de créer une souveraineté numérique européenne non prédatrice ».[120]

LES COMMUNS NUMÉRIQUES

Les communs numériques s'inscrivent dans la réflexion plus générale sur les biens communs. Au croisement de la théorie économique et de la science politique, le thème a été popularisé par le travail d'Elinor Ostrom en 1990, puis par son prix Nobel d'économie en 2009. **Les biens communs désignent au départ des ressources physiques (rivière, forêt, pêcherie) pouvant être administrées collectivement.** Ces biens communs se caractérisent par un ensemble de droits et d'obligations sur une ressource donnée, et par un système de gouvernance au plus près de la singularité du cas. L'enjeu est de protéger la ressource, et si possible œuvrer à son épanouissement.

Les communs numériques ont pour spécificité un coût de copie et de diffusion proche de zéro. Ce sont par exemple des codes sources, des données ou bases de données, des logiciels ou du contenu numérique (image/vidéo/son). À cela s'ajoutent les licences libres, qui organisent juridiquement l'accès, l'utilisation, la production, la modification, la diffusion et la gestion de ces ressources numériques.

Les communs numériques se distinguent des communs physiques par deux caractères principaux : **ils sont non-exclusifs** (pas de limite d'accès si quelqu'un utilise la ressource), et ils sont **non-rivaux** (l'usage de la ressource ne prive pas les autres usagers, elle reste disponible).

Sources : *Governing the Commons. The Evolution of Institutions for Collective Action, Political Economy of Institutions and Decisions*, Elinor Ostrom, Cambridge University Press, 1990

Les Communs, une brèche politique à l'ère du numérique, Valérie Peugeot, Les Débats du Numérique, pp.77-78, Presses des Mines, 2013

Des barbelés sur la prairie Internet : contre les nouvelles enclosures, les communs numériques comme leviers de souveraineté, Benjamin Pajot, note du Ministère des Affaires Étrangères, août 2020

Nous recommandons d'adopter « une stratégie des communs numériques » afin de :

- **mutualiser les coûts de développement de l'InfraTech** (en faisant baisser le coût d'accès au marché *« level the playing field »*)
- **préserver notre souveraineté** en remplaçant les solutions fournies par des géants monopolistiques par une multitude d'acteurs privés (qui s'appuieront sur ces communs pour construire ensemble une offre « packagée » cohérente).

L'Europe pourrait initier cette démarche, avec des pays tiers alignés, et s'attacher à orienter les développements dans le respect de ses valeurs.

[120] *Pour que les communs numériques deviennent un pilier de la souveraineté numérique européenne*, Tribune du 22 juin 2022, Mediapart

3.2.3. Favoriser l'adoption via les data spaces et des cas d'application industriels

Pour un développement de l'IA de confiance dans les *data spaces* européens

Dans sa stratégie Data, l'UE favorise l'émergence du concept de *data spaces* (qui rassemblent des acteurs publics et privés désireux de partager leurs données, personnelles et non personnelles, au moyen d'une infrastructure décentralisée et d'une gouvernance commune) **pour faciliter la circulation et la mutualisation des données**, sur la base de normes communes[121] (développées dans le cadre de Gaia-X). Dix secteurs prioritaires ont été annoncés à ce jour dont la mobilité, la finance, la santé, les compétences, l'énergie, le *Green Deal*, l'administration. **La Commission estime que le marché mondial généré par cette circulation de données s'élèvera à 530 milliards d'euros par an.**

Dans tous les secteurs, l'IA sera au cœur du fonctionnement des *data spaces* pour l'exploitation des données. La promesse fondamentale du *data space* étant de favoriser la confiance entre des acteurs, l'IA de confiance n'est donc pas une possibilité, c'est une nécessité. Ce positionnement prend appui sur les objectifs de l'UE sur le mandat 2019-2024[122].

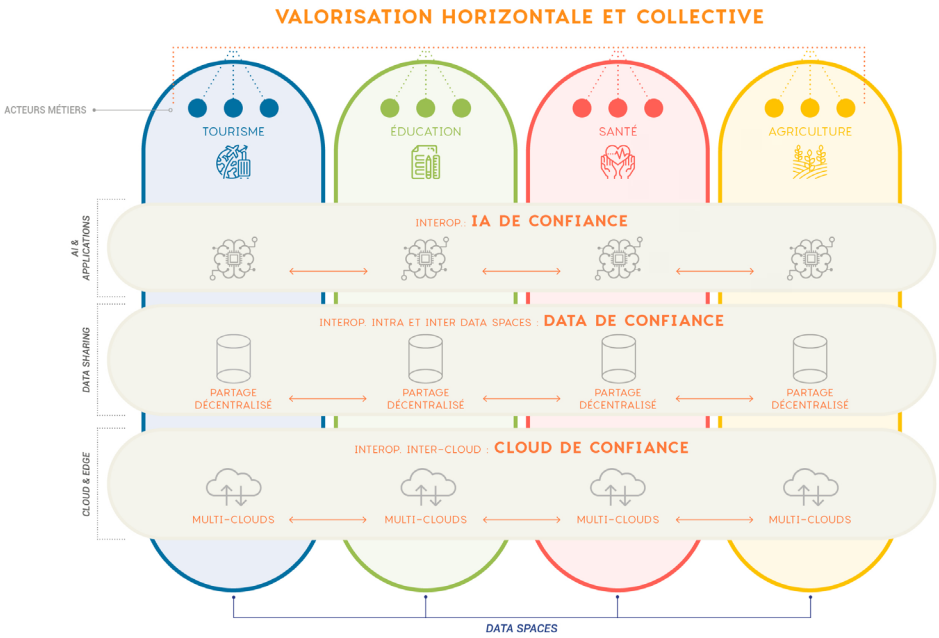
« Il n'y aura pas de *data space* de confiance sans IA de confiance »

La mise en commun des données dans les *data spaces* s'appuiera sur un environnement interopérable, une InfraTech digne de confiance, pour partager les données (« *Smart Middleware for Data-Sharing* » ou InfraTech de la Data de confiance), **et les exploiter pour la réalisation de produits et services, via notamment une InfraTech de l'IA de confiance**. *Green Deal*, Santé, Agriculture, Éducation, Mobilité, sont des exemples de *data spaces* communs européens en construction[123].

Par cette mise à disposition massive de données, la Commission européenne tente d'imposer ses standards et ses valeurs. En effet, quiconque souhaitera accéder aux données de ces *data spaces* devra adhérer à leur gouvernance et en respecter les standards. L'infrastructure s'assurera de la souveraineté des organisations et des personnes sur leurs données et évitera le « *vendor lock-in*[124] » des *Big Tech*, grâce à des normes d'interopérabilité (Gaia-X).

Les *data spaces*, par leurs objectifs, leur traction et leurs valeurs, constituent donc un excellent terrain de diffusion pour tous les acteurs de l'IA de confiance. Ils vont déterminer le développement des systèmes d'IA grâce à la masse de données rassemblées, aux connaissances métier des industriels, et peser sur la gouvernance de l'infrastructure et sur le respect des règles européennes par les acteurs extérieurs. **Les *data spaces* seront un moyen stratégique de s'assurer que les systèmes d'IA de confiance sont à la source des développements et se diffusent.**

[121] *Data de confiance : le partage de données, clé de notre autonomie stratégique*, Olivier Dion et Arno Pons, Digital New Deal, 2022
[122] *European Union Priorities 2019-2024*
[123] *Information session on a preparatory action for the common European Green Deal Data Space under the Digital Europe Programme (DIGITAL), Shaping Europe's digital future*, 15 décembre 2021,
[124] L'enfermement propriétaire consiste à créer une particularité, volontairement non standard, dans la machine, ou le logiciel vendu, pour empêcher le client de l'utiliser avec des produits d'un autre fournisseur, l'empêchant également de le modifier ou d'accéder aux caractéristiques de sa machine pour la modifier. Par exemple, de nombreux éditeurs de logiciels utilisent des formats qui ne peuvent fonctionner qu'avec leur logiciel. Ces stratégies aboutissent à la création de monopoles.



Data Spaces et infrastructures Cloud-Data-IA

CRITICITÉ SOCIALE : L'EXEMPLE DU *DATA SPACE* «EDUCATION & SKILLS»

Le *data space* européen «*Education & Skills*», nommé Prometheus-X, vise à fournir aux acteurs du domaine une infrastructure leur permettant de mutualiser et d'échanger des données ainsi que le développement de solutions, notamment à base d'IA.

En ayant accès à l'ensemble des exercices, notes, intérêts, formations et expériences professionnelles d'un individu, un service d'IA pourra analyser ses forces et faiblesses et lui recommander des choix personnalisés en lien avec les opportunités et besoins d'un territoire donné. Les bénéfices reviennent également aux organisations qui pourront recruter plus facilement.

Cependant cette approche comporte des risques considérables – outre celle de la sécurité des données personnelles – notamment associés au déterminisme induit par les algorithmes (biais systémique) **sur les individus.**

L'IA de confiance est le seul moyen d'adresser à la fois la nécessité de faire usage de l'IA pour exploiter une grande quantité de données, et celle d'encadrer son usage par des règles communes. Ces règles d'ordre général doivent d'un part pouvoir être spécifiques au secteur de l'éducation et des compétences en incluant les principes métiers du domaine (l'orientation par exemple dans les systèmes de recommandation), et pouvoir également, d'autre part, être édictées par la personne concernée elle-même (**«je dois pouvoir décider de ce que je veux ou pas, et l'IA augmentera la précision et la force de ce choix»**).

Une approche globale des marchés Data et IA de confiance

Les enjeux d'un *data space* critique d'un point de vue sociétal (*society critical*), comme l'éducation ou la santé, démontrent la nécessité de ***« considérer ces standards élevés comme des opportunités stratégiques, voire comme des éléments différenciants, dans la course mondiale à l'intelligence artificielle »***[125]. L'avènement de ces standards élevés et sûrs passe nécessairement par une régulation, à la fois audacieuse et sécurisante.

De manière générale, les *data spaces*, qui portent plus sur des sujets *business critical* et *society critical*, que *safety critical*, sont pour le moment en construction. Ils sont donc peu matures sur les sujets d'IA et d'IA de confiance. **En revanche, ils devraient à terme représenter un marché très important pour l'IA de confiance, dopés par les techniques de *federated learning*[126], probablement supérieur au marché des systèmes *safety critical*.**

Les *data spaces* reposent dès le départ sur la confiance entre les acteurs, pour faciliter le partage de données. L'IA de confiance apparaît comme une prochaine étape naturelle pour leur développement. Grâce aux avancées réalisées dans le domaine *safety critical*, où les risques sont les plus importants, et où les acteurs sont acculturés à leur prise en compte, **l'Europe pourra ouvrir de multiples nouveaux marchés pour l'IA de confiance, en s'appuyant sur ses *data spaces* qui symbolisent les enjeux de la « Data de confiance »[127].**

Favoriser la réalisation de cas d'application (systèmes, produits ou services) à base d'IA de confiance

Afin de diffuser les bonnes pratiques et de créer de la valeur dans les secteurs industriels européens, **nous recommandons d'accompagner l'ensemble des filières au travers un soutien à la réalisation de systèmes, produits ou services à base d'IA de confiance.**

Cette approche permet :

- **de favoriser l'adoption de l'InfraTech d'IA de confiance auprès des utilisateurs, qu'ils soient industriels, organismes d'audit ou de certification.** Un accompagnement à la montée en compétences des filières, notamment des PME et startups, est indispensable à la diffusion des pratiques. Il doit donc être proposé au travers d'un Hub de développement, de qualification et de tests en IA de confiance. Connectant de fait l'écosystème des fournisseurs de solutions d'IA de confiance et l'écosystème des développeurs de futurs produits et services.
- **de démontrer l'intégration de l'ensemble des technologies matérielles et logicielles** (électronique, connectivité, IA) **dans une même fonction ou système complexe**, s'appuyant sur l'interopérabilité des différentes chaînes outillées.
- **de structurer l'écosystème industriel, soit l'ensemble des acteurs de la chaîne de valeur,** autour du développement de systèmes complexes à base d'IA de confiance. Favorisant la coopération des fournisseurs de technologies, des équipementiers, des intégrateurs de systèmes, des opérateurs de services, etc.
- **de valider la maintenabilité, mais aussi les évolutions souhaitées des systèmes,** que ce soit dans le temps (par exemple lié à la dégradation, aux conditions d'emploi) ou au travers de mise à jour, par exemple *over the air*[128].

[125] *Rapport Villani*, p.28
[126] En intelligence artificielle et en apprentissage machine, l'apprentissage fédéré (en anglais : federated learning) est une méthode ou un paradigme qui consiste à entraîner un algorithme sur la machine des utilisateurs d'une application et à partager les apprentissages réalisés sur la machine de chaque utilisateur. Cette méthode s'oppose à l'apprentissage centralisé où l'apprentissage se fait sur les serveurs du fournisseur de service. Elle permet notamment un meilleur respect de la vie privée des utilisateurs. - Wikipedia
[127] *Data de confiance*, Digital New Deal, 2022
[128] L' *over-the-air* (ou OTA) est une technologie de communication permettant de transférer des données à distance. - Wikipedia

- **de démontrer les propriétés de confiance** de l'ensemble en fonctionnement et à l'échelle, donc la conformité aux règlements, normes et standards et ceci tout le long du cycle de vie et d'usage.
- **de valider l'adéquation avec les besoins marchés** et les usages.

3.2.4. Une gouvernance partagée pour le numérique de confiance

Un noyau d'alliance industrielle pour l'écosystème de confiance européen

Nous recommandons la création d'une alliance industrielle pour l'IA de confiance en Europe, autour des industriels des filières stratégiques, qui s'appuiera sur les écosystèmes initiés d'ores et déjà au niveau des États membres.

S'INSPIRER DU GRAND DÉFI IA FRANCE (SGPI) DE LA STRATÉGIE NATIONALE EN IA

L'investissement de la France pour l'IA de confiance est porté par le SGPI (Secrétariat Général Pour l'Investissement) dans le cadre de la Stratégie Nationale en IA (SNIA) et à travers le Grand Défi « Sécurisation, fiabilisation et certification des systèmes à base d'intelligence artificielle« ou « IA de confiance ».

Il s'articule autour de 3 piliers stratégiques et complémentaires :

- **Pilier #1 Infrastructure** (cf. encart sur le programme Confiance. ai §3.2.2.) qui ambitionne de développer un environnement de conception de systèmes à base d'IA de confiance. Il s'articule autour de 7 projets indépendants et rassemble aujourd'hui plus de 40 partenaires (grand-groupes, startups, PME, laboratoires de recherche).
- **Pilier #2 Évaluation de Conformité** qui permettra d'assurer la bonne conduite opérationnelle de l'exploitation des systèmes reposant sur l'IA de confiance, puis de définir le rôle et les compétences des certificateurs ou tiers de confiance.
- **Pilier #3 Normes** (cf. encart sur la stratégie française de normalisation en IA §3.1.4.) qui permettra d'établir, en concertation avec les différents acteurs de l'industrie, les normes, standards, environnement réglementaire et certifications.

Le Grand Défi ambitionne de créer un écosystème national de l'IA de confiance, mais, plus largement, au travers de partenariats avec d'autres acteurs européens (Allemagne, etc.) et extra-européens (Québec, etc.), **vise à inspirer la stratégie européenne par des réalisations concrètes**, accompagnant de fait la mise en œuvre opérationnelle de la future réglementation européenne sur l'IA.

A l'échelle européenne, il contribue à faire émerger de futures solutions pour le développement industriel de l'IA de confiance, à renforcer les moyens et l'expertise des acteurs de l'évaluation de conformité et à répondre aux besoins de normes.

Enfin, il ambitionne de préfigurer une future alliance de l'IA de confiance, véritable fédérateur et catalyseur de la vision politique de l'Europe en IA.

Créer une agence européenne d'évaluation de l'IA, voire de la Data et de la robotique

Afin de veiller au respect de la réglementation (et des normes ou standards) au sein de l'espace européen et de conseiller, sous un angle technique et normatif, les politiques publiques vis-à-vis des évolutions de plus en plus rapide des technologies, **nous recommandons de créer une agence européenne d'évaluation de l'IA de confiance.**

Pour ce faire, nous proposons deux options : la création d'un nouvel acteur spécifique, ou la mise en réseau d'acteurs européens existants (et la gouvernance associée) disposant

NOUS DEVONS CRÉER UNE GOUVERNANCE INDUSTRIELLE EUROPÉENNE UNIFIÉE IA ET DATA DE CONFIANCE.

d'expertises et des plateformes techniques de caractérisation. La deuxième option présente l'avantage de s'appuyer sur un écosystème existant disposant de ressources humaines et matérielles, qu'il s'agira sans doute de compléter.

L'agence aura la responsabilité de développer des approches métrologiques d'évaluation de de l'IA au meilleur niveau de l'état de l'art, de mettre en œuvre les plateformes techniques de caractérisation pour y parvenir, et d'apporter une expertise technique dans le champ de la normalisation et des standards. Sur ce dernier point, elle facilitera aussi la « bonne » articulation entre l'approche horizontale et sectorielle, point délicat pour les filières industrielles comme nous avons pu le voir précédemment. Un soutien sera nécessaire pour lui permettre de répondre à ses missions.

Plusieurs composantes, à la fois compétences et moyens matériels, sont ainsi indispensables :

- **les méthodologies d'évaluation et d'essais,** en lien avec les problématiques de réglementation ou d'homologation (mais aussi des normes et standard permettant de démocratiser et diffuser les méthodes et résultats), y compris des facteurs humains ;
- **des plateformes de caractérisation physique et d'expérimentation,** des plateformes numériques ou environnements de simulation, afin de réduire les coûts d'évaluation et d'expérimentation ; les données réelles issues des expérimentations permettant également d'en améliorer les performances ;
- **des plateformes d'évaluation de cybersécurité** de ces systèmes (notamment pour permettre des tests d'intrusion et des scans de vulnérabilité répétables et reproductibles) afin de garantir les meilleures pratiques et technologies pour limiter les attaques malveillantes, notamment celles rendues possibles par l'introduction d'IA ;
- **les protocoles ou méthodologies de spécification des domaines d'emploi et des scenarii de test ;**
- **les usages et l'acceptabilité sociale.**

Cette agence accompagnera la montée en compétences des autorités notifiantes et organismes notifiées des différents secteurs industriels, ainsi que de l'écosystème d'évaluation de la conformité (audits, essais, etc.).

LE NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST)

Le NIST est une agence du département du Commerce des États-Unis. Sa mission est de soutenir l'économie en développant des technologies, la métrologie et des normes en collaboration avec les industriels. Fondé en 1901, cet institut est l'un des plus anciens laboratoires de sciences appliquées des États-Unis. **Il dispose d'un budget notable : 1 milliard ($) en 2010, 1.34 milliards ($) en 2020.** Son activité est supervisée par le Comité des sciences, de l'espace et des technologies de la Chambre des Représentants des États-Unis. Cette agence s'est aussi vue confier l'étude de l'effondrement de Twin Towers lors de l'attaque du 11 septembre 2001, afin de déterminer les raisons techniques des effondrements et des incendies. Par ailleurs, le NIST compte pas moins de quatre prix Nobel parmi les chercheurs qu'il a employés. Les homologues européens du NIST sont notamment le LNE en France, le PTB en Allemagne, l'INRIM en Italie et le CEM en Espagne.

Sources : NIST, Office of the Director, Congressional and Legislative Affairs, National Institute of Standards and Technology, US Department of Commerce, National Institute of Standards and Technology, Wikipedia

Une gouvernance unifiée et industrielle Data & IA

Le 19 février 2020 Margrethe Vestager (*Vice-Présidente exécutive de la Commission européenne*), et Thierry Breton (*Commissaire européen, chargé du marché intérieur, de la politique industrielle, du tourisme, du numérique, de l'audiovisuel, de la défense et de l'espace*) ont présenté **la nouvelle « feuille de route » digitale pour l'Europe : un cadre législatif cohérent entre IA et Data** a été posé pour servir l'objectif commun d'une Europe adaptée à l'ère du numérique et humaniste. **Les données y sont envisagées comme *« le carburant de l'intelligence artificielle »***.

Les textes associés à la stratégie Data et à la stratégie IA de l'UE, prévoient la mise en place de nouveaux organes de gouvernance, dénommés Comités ou *« Boards »* Data et IA. **Cependant on remarque que leurs constructions sont à ce stade envisagées de manière différentes**, la structure semble similaire au premier abord, mais la composition, et surtout le rôle des parties prenantes, sont différents.

Pour le *Data Innovation Board*[129] tout est pensé en termes trans-sectoriels avec l'interopérabilité au cœur de l'action. **De nombreuses parties prenantes sont impliquées, y compris du côté des industriels**, qui sont associés à la démarche dès le départ. L'ensemble des acteurs est aussi incité au dialogue et à l'action commune.

À l'inverse **pour l'*Artificial Intelligence Board*, la proposition de la Commission** (version d'avril 2021) **articule la composition du *Board* et son action autour de la coopération de « représentants nationaux », autour de la coopération avec les autorités nationales de supervision**, que la Commission supervise à son tour. Le rôle des parties prenantes de l'IA est seulement évoqué et les industriels ne sont pas directement associés. **Or, nous ne parviendrons pas à construire des organes clés pour la régulation numérique au niveau européen, équilibrant les missions de contrôle et la préservation de l'innovation, sans organes profondément multipartites, capables de favoriser l'interdisciplinarité et donc de mixer différentes cultures techniques, juridiques, sciences humaines et sociale, voire de la société civile.** La présence d'experts techniques, même minoritaires, mais capables de suivre et de comprendre l'évolution de l'état de l'art technologique et d'en exposer les enjeux au reste du *Board* semble indispensable. **A ce titre, le projet de règlement sur l'IA paraît moins abouti que son équivalent côté Data, duquel nous recommandons de nous inspirer.**

[129] *Regulation of the European Parliament and of the Council on European data governance* (Data Governance Act) 2020/0340 (COD)

Comparaison des Boards envisagés pour la Data et pour l'IA (version avril 2021)

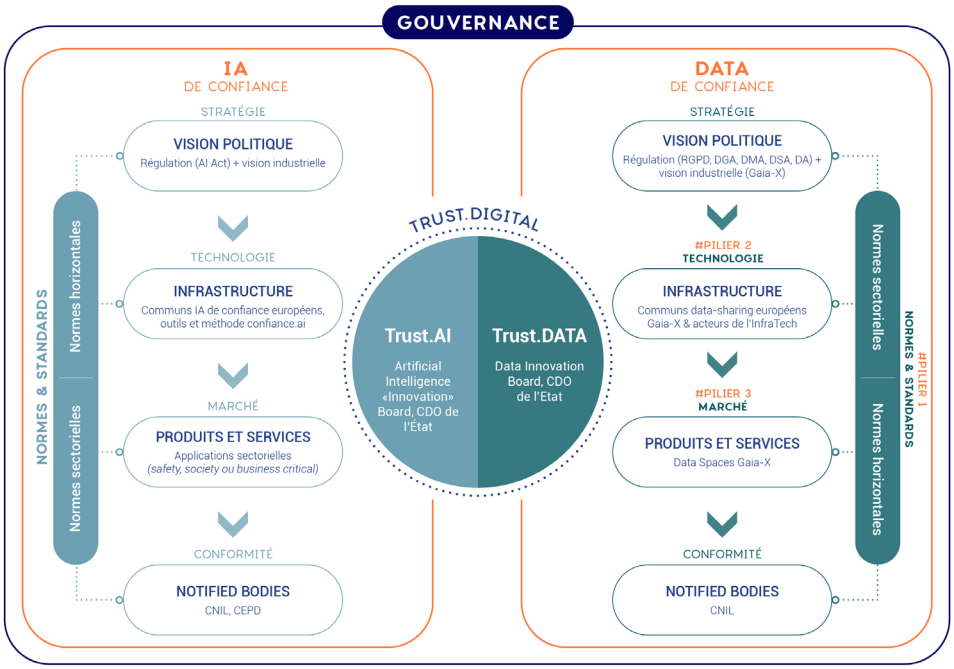
ACTE LÉGISLATIF	GOUVERNANCE	TÂCHES
Data Governance Act : « Data Innovation Board	« un groupe d'experts, qui se compose des représentants des autorités compétentes de tous les États membres, du comité européen de la protection des données, de la Commission, des espaces de données (*data spaces*) pertinents et d'autres représentants d'autorités compétentes dans des secteurs particuliers. »	• « conseiller et assister la Commission dans l'élaboration d'une pratique cohérente des organismes du secteur public » • « conseiller et assister la Commission dans l'élaboration d'une pratique cohérente des autorités compétentes quant à l'application des exigences auxquelles sont soumis les prestataires de services de partage de données » • « conseiller la Commission sur la hiérarchisation des normes transsectorielles à utiliser et à mettre au point pour l'utilisation de données et le partage de données entre différents secteurs (...) » • « aider la Commission à améliorer l'interopérabilité des données ainsi que les services de partage de données (...) » • « faciliter la coopération entre les autorités compétentes »
Artificial Intelligence Act : « Artificial Intelligence Board »	« composé de représentants des États membres et de la Commission. »	« Le Comité facilitera la mise en œuvre fluide, efficace et harmonisée du présent règlement en contribuant à la coopération efficace entre les autorités de contrôle nationales et la Commission et en fournissant des conseils et une expertise à la Commission. Il recensera également les meilleures pratiques et les diffusera dans les États membres. »

Prenons l'exemple du RGPD. Le texte est bousculé par le rythme effréné de l'innovation technologique en IA. Ayant été négocié et adopté avant l'explosion des succès applicatifs de l'IA contemporaine, il n'intègre pas la notion d'apprentissage automatique par les données, ni la dichotomie qui existe entre l'usage des données personnelles pour l'entraînement et celui pour l'inférence. Pourtant, elles sont porteuses de risques de natures différentes pour les personnes. Elles appellent donc des équilibres différents pour l'application de nombreux principes comme la durée de conservation, la minimisation, le besoin de données hors finalité pour construire des classifieurs, le besoin renforcé de données d'apprentissage sur les catégories protégées pour lutter contre les biais à leur encontre, etc. Or, cinq ans après la survenance de la « vague de l'IA », le Comité européen de la protection des données (CEPD), sans doute à cause de sa composition et d'un manque de moyen, n'a toujours pas intégré dans sa doctrine ces enjeux de manière substantielle et suffisamment claire et lisible pour les acteurs industriels européens.

Nous recommandons à l'Europe de favoriser une réflexion pionnière et fertile permettant de concilier l'indispensable protection des données personnelles avec l'innovation technologique et économique, sans quoi le risque est réel d'inadéquation des réglementations européennes au développement technologique. La réflexion autour d'une révision de la composition du CEDP, en lien avec les sujets de données et d'IA, couplée à une réflexion sur les moyens qui y sont consacrés, devrait être entreprise.

Comme nous l'avons vu tout au long de ce rapport, les sujets Data et IA sont fortement liés. Il apparaît également qu'**une stratégie de bout-en-bout (régulation, norme, industrialisation) est la condition *sine qua non* de la défense de nos intérêts.** En conséquence, nous recommandons également d'**unifier la gouvernance de l'ensemble des composantes de l'écosystème de confiance (Cloud-Data-IA) dans un même organisme, qui serait porté par une vision politique industrielle d'ensemble, conciliant innovation et protection sociétale, au niveau européen. Cette gouvernance européenne unifiée pourrait par exemple être accompagnée de la nomination d'un CDO (Chief Data Officer), ou CTO (Chief Technical Officer), par État membre,** sur le modèle espagnol[130], **et d'un CDO/CTO pour l'Europe.**

Gouvernance de l'écosystème de confiance IA et Data



[130] Alberto Palomo-Lozano premier CDO de l'Espagne

RECAPITULATIF DES RECOMMANDATIONS

RECOMMANDATION 1 :

FAIRE DU NUMÉRIQUE DE CONFIANCE UN PROJET POLITIQUE EUROPÉEN

1.a Imposer la confiance comme doctrine de la troisième voie numérique	Porter à un niveau sociétal le message politique de l'IA de confiance comme seul chemin raisonnable et envisageable.
1.b Assumer notre stratégie de régulation extra-territoriale.	Faire du paquet réglementaire (RGPD, DSA, DMA, DGA, DA, AI Act), et de l'approche par le risque, la première pierre à l'édifice d'un écosystème de confiance mondial.
1.c Promouvoir une vision matricielle de l'IA de confiance.	Défendre une approche industrielle horizontale transfilières, et verticale de l'écosystème de confiance (de la régulation à l'industrialisation, en passant par la norme).

RECOMMANDATION 2 :

ÉTABLIR LA NORMALISATION EN PRIORITÉ STRATÉGIQUE

2.a Promouvoir la confiance dans l'IA comme "marque" européenne.	Développer des normes et labels européens génériques reposant sur la réglementation, pour conquérir le marché.
2.b Devenir un standard-setter en mobilisant tous les acteurs.	Coordonner et amplifier les efforts de l'UE, des États membres, des grands groupes et des startups/PME.
2.c Garantir un équilibre entre régulation et innovation de l'IA par les *sandboxes*.	Promouvoir des *sandboxes* réglementaires, suivant les critères proposés par l'OCDE, afin de conserver une agilité indispensable à la future mise en œuvre de la réglementation.

RECOMMANDATION 3 :

BÂTIR UNE STRATÉGIE INDUSTRIELLE DE L'IA DE CONFIANCE

3.a Miser sur la confiance comme clé d'entrée sur le marché mondial de l'IA.	Conquérir le marché global de l'IA (231 Milliards €), en pénétrant celui de l'IA de confiance via le marché "accessible" des solutions d'ingénierie et méthodes.
3.b Faire de la culture des systèmes critiques la pointe de la flèche de la stratégie	Capitaliser sur la culture européenne des systèmes critiques, où les exigences sont les plus fortes, pour en faire un avantage compétitif.
3.c Développer une approche orientée industrie et innovation.	Capitaliser sur notre expertise en recherche pour accroître nos capacités en RDI, et financer davantage de formations notamment dans les domaines de l'ingénierie..

RECOMMANDATION 4 :

CRÉER L'INFRATECH DE L'IA DE CONFIANCE

4.a Faire émerger un large écosystème d'acteurs de l'InfraTech de l'IA de confiance.	Créer et animer une large communauté européenne innovante (en particulier startups) pour contenir et surtout conquérir le marché de l'IA trusté par les *Big Tech*.
4.b Développer une plateforme logicielle InfraTech de l'IA de confiance.	Fédérer une InfraTech pour une offre packagée, bout en bout, couvrant toute la chaîne de management des risques IA, afin de permettre aux acteurs métier de passer à l'échelle.
4.c Accompagner la création de communs numériques de l'IA de confiance.	Abaisser les barrières à l'entrée en mutualisant les efforts de développement via des communs numériques open source gouvernés par une multitude d'acteurs de l'InfraTech.

RECOMMANDATION 5 :

PASSER À L'ÉCHELLE EUROPÉENNE

5.a Construire une alliance industrielle européenne d'IA de confiance.	Capitaliser sur les projets en cours de construction dans les États membres, et créer un noyau d'alliance à partir des premiers partenariats (France et Allemagne).
5.b Imaginer un centre européen d'évaluation des IA.	Favoriser la montée en compétence d'un écosystème d'audit et d'évaluation de conformité européen, le tout en partenariat avec les organes sectoriels.
5.c Soutenir la réalisation de démonstrateurs industriels.	Lancer des projets phares européens (*lighthouse projects*) de démonstrateurs industriels à base d'IA de confiance dans des filières stratégiques pour l'Europe.

RECOMMANDATION 6 :

PROPOSER UNE STRATÉGIE COMMUNE DATA ET IA

6.a Développer l'IA de confiance pour les data spaces européens.	Appliquer les solutions d'IA de confiance aux data spaces multi-sectoriels "*society et business critical*" de Gaia-X (éducation, santé, tourisme, *Green Deal*,...).
6.b Incarner les stratégies Data et IA en Europe.	Promouvoir la nomination de Chief Technical Officer (CTO) / Chief Digital Officer (CDO) pour chaque État membre et pour l'Europe.
6.c Créer une gouvernance industrielle européenne unifiée IA et Data.	Concevoir un nouvel organe de gouvernance européen unifié de l'IA et de la data de confiance, garantissant un équilibre entre protection et innovation, comprenant l'ensemble des parties prenantes (régulation, industrie, société, etc.).

CONCLUSION

Maîtriser les technologies numériques souveraines et sûres est une impérieuse nécessité. France 2030, qui mobilise pour ce faire 3 Md€, incarne la volonté française renouvelée de faire de notre indépendance numérique européenne une réalité tangible. Cette ambition, portée au plus haut niveau depuis plusieurs années, est devenue une exigence depuis que la crise Covid a révélé nos dépendances à certaines technologies critiques. Elle repose sur deux fondamentaux. Le premier consiste à protéger nos organisations, publics et privés, et nos concitoyens et à faire respecter nos valeurs en s'appuyant sur la régulation (RGPD, DSA, DMA, IA act, Cyber act). La seconde entend faire émerger des offres souveraines, alternatives crédibles aux géants extra-européens, en investissant et accompagnant des écosystèmes innovants de confiance.

Pour atteindre cet objectif, la France est à l'initiative de projets industriels communs, notamment dans le domaine de l'intelligence artificielle. France 2030 a vocation à renforcer cette ambition en donnant aux acteurs français les moyens d'investir dans l'IA mais également de former aux emplois de demain dans le domaine.

Pour ce faire, un éclairage scientifique, véritable état de l'art, était nécessaire. Le rapport présenté ici en constitue une synthèse fidèle: Qu'entend-on pas IA de confiance ? L'Union Européenne peut-elle faire de l'IA de confiance une valeur étalon ? Comment la France peut-elle faire de l'IA de confiance une opportunité pour son autonomie industrielle ? Répondre à ces questions, et bien d'autres encore, doit nous permettre collectivement de constituer une feuille de route qui permette à la France de prendre en main son destin numérique. La France de 2030 sera celle du mieux vivre, du mieux produire et du mieux comprendre. Prendre des risques, développer des solutions nationales, faire naître et grandir des champions français ou encore accompagner les entreprises et l'ensemble de la société dans les grandes transitions, notamment numériques, en les formant, , comptent parmi les raisons d'exister de France 2030.

Cette note constitue une contribution importante à cette démarche. Elle rappelle la nécessaire coopération entre les acteurs émergents et les industriels pour bâtir une véritable indépendance numérique, prérequis à la compétitivité de nos entreprises que nous appelons tous de nos vœux. Le numérique de confiance (Cloud-Data-IA) est essentiel pour accompagner ces objectifs. Ce rapport de grande qualité permet d'éclairer l'État sur la stratégie à mettre en œuvre pour répondre aux besoins des chaînes de valeur. Les auteurs, aidés de prestigieux contributeurs, valident ainsi les investissements auxquels nous avons procédé dans le cadre de la stratégie IA financée par France 2030, et offrent une mise en perspective intéressante pour la suite.

Les propositions phares de ce rapport consistant d'une part à maîtriser des infrastructures stratégiques du numérique de confiance pour garantir notre souveraineté, et d'autre part à consolider une stratégie commune IA et Data par une gouvernance tournée vers l'Europe, vont dans le bon sens. Elles nous confortent dans nos choix et constituent un cap pour aller plus loin, en nous munissant de notre « Boussole numérique », qui devra nécessairement indiquer la même direction que notre cap écologique. France 2030 répondra présent en allant plus vite et plus fort.

Bruno Bonnell

Secrétaire Général pour l'Investissement, France 2030.

NOS REMERCIEMENTS

Julien Chiaroni, directeur Grand Défi IA, Secrétariat Général Pour l'Investissement (SGPI) et **Arno Pons**, Délégué Général, Digital New Deal, **tiennent à remercier pour leur contribution :**

ANIMATION ET ÉDITORIALISATION

Olivier Dion — Coordination technique et rédaction, Digital New Deal. CEO Onecub, co-fondateur aNewGovernance

Prune Zammarchi — Chargée de mission, Digital New Deal

CONSEIL ET CONTRIBUTIONS

Guillaume Avrin — Responsable du département Evaluation de l'IA du LNE

Patrick Bezombes — CEN CENELEC, co-président atelier "Souveraineté numérique", Vice-président du JTC 21 (IA), AFNOR, Président du comité de normalisation IA et Big Data

Matthias de Bièvre — CEO Visions, co-fondateur aNewGovernance, CEO Prometheus-X

Yannick Bonhomme — Responsable valorisation IRT SystemX, Confiance.ai

Bertrand Braunschweig — Coordonnateur scientifique du programme "Confiance.ai"

Loic Cantat — R&D Manager AI and Data Science - IRT SystemX

Agnès Delaborde — LNE Research Engineer

Emmanuelle Legrand — Chargée de mission IA, DGE ministère des Finances

Juliette Mattioli — Experte senior en IA Thales, Présidente du Hub "Data Sciences & Artificial Intelligence" du pôle Systematic Paris Région

Eric Pol — Chairman aNewGovernance

Benoît Rottembourg — Responsable projet REGALIA, INRIA

Renaud Vedel — Préfet, Coordonnateur pour le Gouvernement de la Stratégie Nationale en IA, Co-chair, Steering Committee Global Partnership on AI (GPAI)

INTERVIEWS ET PARTICIPATIONS COMPLÉMENTAIRES

Christine Balagué — Professeure, Good in Tech Chair (Institut Mines-Télécom)

Julien Chasserieau — AI Policy Manager, DIGITALEUROPE

Thierry Collette — Directeur du groupe Information Science et Technologie, Thales

Andreas Dengel — Prof. Dr. Prof. h.c. Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

Laurence Devillers — Professeure de sciences informatiques, Université Paris-Sorbonne, co-chair CNRS du groupe IA ad hoc "Éthique/Nudging" pour l'AFNOR

Emmanuelle Escorihuela — AI Transformation Leader for Systems, Airbus

Jean-Baptiste Fantun — CEO NukkAI

Kilian Gross — Chef d'Unité AI, DG Connect, Commission Européenne

Miapetra Kumpula-Natri — Député Européen, rapporteur Data Act (Finlande)

Guillaume Leboucher — DG Délégué Openvalue, Membre COMEX Docaposte

Marc Leobet — Directeur de projet "IA et Transition écologique", Ministère de la Transition écologique

Dragos Tudorache — Député Européen, rapporteur AI Act (Roumanie)

Axel Voss — Député Européen, shadow rapporteur AI Act (Allemagne)

JULIEN CHIARONI

Directeur des Grands Défis sur "l'IA de confiance pour l'industrie" au sein du Secrétariat Général à l'Investissement (SGPI), expert auprès du Conseil européen de l'innovation. Il a occupé auparavant des postes à l'institut de microélectronique et des technologies de l'information (Leti), puis à l'Institut des technologies du numérique et de l'intelligence artificielle (List) qui compte plus de 700 chercheurs répartis entre Paris-Saclay et Grenoble. En tant que Directeur de la stratégie et des programmes, il a mis en œuvre la stratégie de l'institut dans un large éventail de technologies numériques, et a établi des partenariats entre la recherche et l'industrie. De 2008 à 2010, il avait coordonné le programme nanoscience et nanotechnologie à l'Agence nationale de la recherche. Il a été également attaché aux affaires académiques et universitaires au Consulat général de France à Hong Kong et Macao.

ARNO PONS

Délégué général du « Think-tank Digital New Deal », co-auteur de quatre notes sur le numérique de confiance (Cloud de confiance, Infrastructures du numérique de confiance, IA de confiance, et Data de confiance). Il a fondé en 2021 l'activité « Do Tank - Digital New Deal », dédiée aux enjeux de coopération pour accompagner les entreprises et collectivités à structurer les filières en écosystèmes digitaux via des alliances technologiques. Il lance la première initiative de place du Do Tank en co-créant en 2022 Themis (*data space* tourisme comprenant soixante entités partenaires). Il avait créé auparavant plusieurs start-up (Checkfood-gaspillage alimentaire, Medicimo au Canada,...), et également enseigné à SciencesPo sur les enjeux de souveraineté numérique liés à la centralisation des pouvoirs par les *Big Tech*.

A PROPOS DE FRANCE 2030

LE PLAN D'INVESTISSEMENT FRANCE 2030

- Traduit une double ambition : transformer durablement des secteurs clefs de notre économie (énergie, automobile, aéronautique ou encore espace) par l'innovation technologique, et positionner la France non pas seulement en acteur, mais bien en leader du monde de demain. De la recherche fondamentale, à l'émergence d'une idée jusqu'à la production d'un produit ou service nouveau, France 2030 soutient tout le cycle de vie de l'innovation jusqu'à son industrialisation.
- Est inédit par son ampleur : 54 Md€ seront investis pour que nos entreprises, nos universités, nos organismes de recherche, réussissent pleinement leurs transitions dans ces filières stratégiques. L'enjeu : leur permettre de répondre de manière compétitive aux défis écologiques et d'attractivité du monde qui vient, et faire émerger les futurs champions de nos filières d'excellence. France 2030 est défini par deux objectifs transversaux consistant à consacrer 50 % de ses dépenses à la décarbonation de l'économie, et 50% à des acteurs émergents, porteurs d'innovation sans dépenses défavorables à l'environnement (au sens du principe Do No Significant Harm).
- Sera mis en œuvre collectivement : pensé et déployé en concertation avec les acteurs économiques, académiques, locaux et européens pour en déterminer les orientations stratégiques et les actions phares. Les porteurs de projets sont invités à déposer leur dossier via des procédures ouvertes, exigeantes et sélectives pour bénéficier de l'accompagnement de l'Etat.
- Est piloté par le Secrétariat général pour l'investissement pour le compte du Premier ministre.

Plus d'informations sur : https://www.gouvernement.fr/france-2030

DIGITAL NEW DEAL
LE THINK-TANK DE LA NOUVELLE DONNE

Digital New Deal accompagne les décideurs privés et publics dans la création d'un Internet des Lumières, Européen et Humaniste. Notre conviction est que nous pouvons offrir une 3e voie numérique en visant un double objectif : défendre nos valeurs en proposant une nouvelle régulation contre la centralisation des pouvoirs ; et défendre nos intérêts en créant les conditions de la coopération face à la captation de la valeur par les « *Big Tech* ».

Notre activité de publication a pour vocation d'éclairer de manière la plus complète possible les évolutions à l'œuvre au sein de enjeux de « souveraineté numérique », dans l'acception la plus large du terme, et d'élaborer des pistes d'actions concrètes, voire opérantes via le Do tank, à destination des organisations économiques et politiques.

LE CONSEIL D'ADMINISTRATION

Olivier Sichel (président fondateur) et Arno Pons (délégué général), pilotent les orientations stratégiques du think-tank sous le contrôle régulier du conseil d'administration.

Forts de leur intérêt commun pour les questions numériques, les membres du Conseil d'administration ont décidé d'approfondir leurs débats en formalisant un cadre de production et de publication au sein duquel la complémentarité de leurs expériences pourra être mise au service du débat public et politique. Ils s'impliquent personnellement dans la vie de Digital New Deal, notamment dans le choix des rapports et de leurs rédacteurs. Il sont les garants de notre indépendance, académique et économique.

SÉBASTIEN BAZIN
PDG AccorHotels

NATHALIE COLLIN
DG branche Grand Public et Numérique Groupe La Poste

NICOLAS DUFOURCQ
DG de Bpifrance

AXELLE LEMAIRE
Ex-Secrétaire d'Etat du Numérique et de l'Innovation

ALAIN MINC
Président AM Conseil

DENIS OLIVENNES
DG Libération

YVES POILANE
DG Ionis Education Group

ARNO PONS
Délégué général du think tank Digital New Deal

JUDITH ROCHFELD
Professeur agrégée de Droit, Panthéon Sorbonne

OLIVIER SICHEL
President Digital New Deal
DGA Caisse des Dépôts

BRUNO SPORTISSE
PDG Inria

ROBERT ZARADER
PDG Bona fidé

NOS PUBLICATIONS

Cybersécurité, vigile de notre autonomie stratégique | Arnaud Martin, Didier Gras - *juin 2022*

RGPD, acte II : la maîtrise collective de nos données comme impératif | Julia Roussoulières, Jean Rérolle – *mai 2022*

Fiscalité numérique, le match retour | Vincent Renoux - *septembre 2021*

Défendre l'état de droit à l'ère des plateformes | Denis Olivennes et Gilles Le Chatelier - *juin 2021*

Cloud de confiance : un enjeu d'autonomie stratégique pour l'Europe | Laurence Houdeville et Arno Pons - *mai 2021*

Livres blancs : Partage des données & tourisme | Fabernovel et Digital New Deal - *avril 2021*

Partage de données personnelles : changer la donne par la gouvernance | Matthias de Bièvre et Olivier Dion - *septembre 2020*

Réflexions dans la perspective du Digital Services Act européen | Liza Bellulo - *mars 2020*

Préserver notre souveraineté éducative : soutenir l'EdTech française | Marie-Christine Levet - *novembre 2019*

Briser le monopole des *Big Tech* : réguler pour libérer la multitude | Sébastien Soriano - *septembre 2019*

Sortir du syndrome de Stockholm numérique | Jean-Romain Lhomme - *octobre 2018*

Le Service Public Citoyen | Paul Duan - *juin 2018*

L'âge du web décentralisé | Clément Jeanneau - *avril 2018*

Fiscalité réelle pour un monde virtuel | Vincent Renoux - *septembre 2017*

Réguler le « numérique » | Joëlle Toledano - *mai 2017*

Appel aux candidats à l'élection présidentielle pour un #PacteNumérique | *janvier 2017*

La santé face au tsunami des NBIC et aux plateformistes | Laurent Alexandre - *juin 2016*

Quelle politique en matière de données personnelles ? | Judith Rochfeld - *septembre 2015*

Etat des lieux du numérique en Europe | Olivier Sichel - *juillet 2015*

contact@thedigitalnewdeal.org

THINK-TANK
DIGITAL
NEW DEAL

juin 2022

www.thedigitalnewdeal.org