IA GÉNÉRATIVE : S'UNIR OU SUBIR

Le partage des données privées, atout stratégique pour un IA agentique européenne

Olivier Dion, Michel-Marie Maudet, Arno Pons

Mouvement des Entreprises de France

THINK-DO-TANK DIGITAL NEW DEAL

Novembre 2024

GEN AI = WEB 2^2

SOMMAIRE

PRÉFACE

Si l'Intelligence Artificielle n'est pas une nouvelle technologie, l'émergence de l'IA générative marque un tournant décisif : ses usages et ses impacts immédiats et futurs sont majeurs. L'engouement suscité par les premiers tests et l'adoption croissante dans les entreprises en est la preuve. Les nombreuses possibilités qu'elle offre ouvrent des perspectives considérables pour la compétitivité. Toutefois, l'IA est, comme beaucoup de technologies, un « pharmakon » : une solution autant qu'un problème.
Pour que l'Europe tire pleinement profit de cette transformation, il est essentiel de bâtir une IA sur les fondations pérennes que sont la transparence, la sécurité et la souveraineté. Ces fondations sont un socle incontournable mais elles ne seraient rien sans la matière première que sont les données.
Cela peut sembler une évidence mais le véritable levier d'innovation est là, pour nous européens qui avons et générons parmi les données à plus hautes valeurs : celles des entreprises.

Nous sommes tous collectivement créateurs de données qui nécessitent de l'extraction, du raffinage et de l'enrichissement comme toute autre matière. Or, ce traitement de la donnée est plus facile à dire qu'à faire car les infrastructures et la puissance de calcul ne sont pas toujours entre nos mains. Dans ce cas l'union fera notre force : le partage de données, ou data sharing, est un mécanisme clé pour tirer de nos données l'enrichissement des modèles d'IA générative qui sauront répondre de manière pertinente aux besoins spécifiques des secteurs dont elles sont issues.

Bien entendu, ce partage doit se faire sur la base du volontariat, dans des espaces de données sectoriels sécurisés (Data Spaces). C'est cette approche volontaire et collaborative qui fait la force de notre stratégie, puisqu'elle offre à chaque secteur, et à chaque entreprise, l'opportunité de rester maître de ses données, tout en bénéficiant des avantages de l'intelligence collective.
En misant sur le partage sécurisé des données à l'échelle européenne via les Data Spaces, nous donnons aux entreprises les moyens de construire des IA adaptées à leurs besoins. C'est en mutualisant les données crée des écosystèmes puissants et souverains, capables de rivaliser avec les grandes plateformes technologiques.

Cependant, la réussite de cette transformation ne repose pas uniquement sur la technologie. L'autre défi de taille réside dans les compétences puisque la maîtrise des outils numériques et de l'IA reste la clé de voûte de cette transition. Pour que nos entreprises puissent exploiter pleinement le potentiel de l'IA générative, nous devons investir massivement dans la formation continue et le développement des compétences.

Enfin, il ne s'agit pas seulement de construire des IA performantes et compétitives : elles devront aussi s'inscrire dans une démarche responsable en relevant les défis éthiques, énergétiques, environnementaux et cyber. Ces enjeux sont cruciaux pour assurer une transition numérique durable, sécurisée, et respectueuse de nos ressources.

Il est de notre responsabilité collective de construire des environnements capables et de confiance, et qui serviront les intérêts de nos entreprises. En s'appuyant sur le partage volontaire des données et en renforçant les compétences de nos talents, nous pourrons faire de cette révolution technologique une opportunité pour toutes nos filières. Ensemble, nous ne subirons pas cette transformation, mais nous en serons les architectes.

Virginie Fauvel,
Coprésidente de la commission Numérique et Innovation du MEDEF.

INTRODUCTION

Après avoir publié la collection "Cloud de confiance[1]", "IA de confiance[2]" et "Data de confiance[3]" en 2022, notre think tank a décidé de consacrer un rapport à l'"IA générative de confiance", tant cette technologie révolutionnaire bouleverse l'écosystème numérique analysé dans nos précédents livres blancs. Transformation inédite, car contrairement aux révolutions industrielles du chemin de fer et de l'électricité qui avaient nécessité des décennies pour déployer leurs infrastructures physiques, cette rupture technologique repose sur un réseau déjà en place : Internet. Laissant peu de temps aux décideurs, qu'ils soient publics ou privés, pour s'adapter progressivement aux changements profonds qui se profilent.

L'IA générative est un game changer qui nous met face à une situation comparable à l'arrivée du Web 2.0 il y a 25 ans, avec les mêmes questions : resterons-nous passifs, simples consommateurs de technologies extra-européennes ? Répéterons-nous les erreurs du passé en laissant des acteurs étrangers capter la valeur de nos filières par la data ? Après le "Free Internet", allons-nous tomber dans le nouveau piège libertarien du "Free AI", c'est-à-dire un marché libre non régulé par les États, un espace où la liberté individuelle et la propriété privée restent sans contrôle externe ?

La réponse est NON. Nous ne pouvons pas courir le risque que l'IA générative (ou GenAI) amplifie davantage les phénomènes de centralisation du pouvoir et de captation de la valeur que notre think tank combat depuis huit ans. Menace que nous résumons en une formule : **GenAI = Web2**[2] (enfermement au carré).

Si on ne veut pas que les entreprises soient davantage prises en tenaille par les BigTechs, entre d'un côté leurs BigClouds (hyperscalers comme Azure, AWS, ou GoogleCloud) et de l'autre leurs BigAI (comme OpenAI, Copilot ou Gemini), il faut pour cela adopter une stratégie volontariste en s'appuyant sur les récentes initiatives françaises et européennes qui offrent un cadre plus propice à nos entreprises :

- Une politique de formation avancée générant des talents pour la recherche universitaire et les laboratoires privés de recherche appliquée.
- Un écosystème technologique dynamique illustré par la French Tech avec quelques remarquables réussites françaises, en particulier dans le domaine de l'IA.
- Un niveau de financement des programmes digitaux ambitieux avec par exemple France 2030 (54 Mrd€) ou Horizon Europe (93 Mrd€).
- Un cadre réglementaire européen robuste avec le DMA, le DSA, le RGPD, et l'AI Act, première régulation mondiale en la matière.
- Une stratégie européenne des données (2020) très ambitieuse, et ayant vocation à créer un marché unique de la donnée via le "Shared data", et les Espaces Communs de Données (les fameux "Data Spaces" que nous aborderons en détail dans ce rapport). Cette stratégie s'appuie sur une régulation offensive

[1] *Cloud de confiance, un enjeu d'autonomie stratégique pour l'Europe*, Laurence Houdeville et Arno Pons, Digital New Deal, 2021

[2] *IA de confiance : opportunité stratégique pour une souveraineté industrielle et numérique*, Julien Chiaroni et Arno Pons, Digital New Deal, 2022

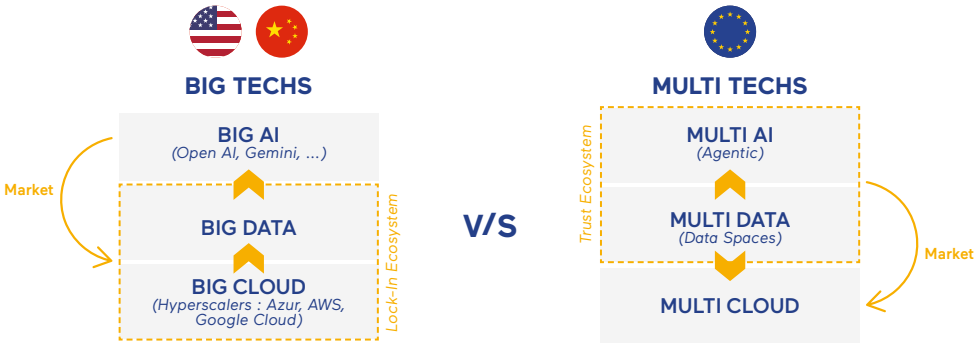
[3] Data de confiance : le partage des données, clé de notre autonomie stratégique, Olivier Dion et Arno Pons, Digital New Deal, 2022

et une gouvernance d'ensemble (DGA, DA, EDIB)[4] , un cadre opérationnel (GAIA-X[5], DSSC[6], IDSA[7]), et des financements dédiés tant au niveau européen qu'au niveau des États membres (Simpl[8], ATF[9]), etc.

Tout est là, dessiné, prêt à être activé. Un marché unique, une gouvernance harmonisée, et des Data Spaces comme nouveaux piliers. Soucieuse de la biodiversité technologique, l'Europe est désormais en mesure de devenir le terreau de cette alternative aux écosystèmes monopolistiques qui emprisonnent leurs utilisateurs dans des environnements cloud-data-IA verrouillés (lock-in).

Ce rapport détaille comment nos entreprises, et administrations, peuvent collectivement prendre leur destin en main en s'appuyant sur ces nouveaux avantages compétitifs. Bien exploités, ceux-ci peuvent non seulement combler l'écart, mais aussi ouvrir de nouvelles voies pour des innovations différenciantes et un potentiel leadership.

Les entreprises européennes doivent pour cela faire face à un choix : s'unir ou subir. Subir, c'est consentir à une dépendance accrue, à l'idée que nos technologies, nos talents, et nos perspectives soient capturés par d'autres, étrangers à nos valeurs. S'unir, c'est bâtir une innovation qui, loin d'être une simple réponse à la marche inexorable d'une homogénéité imposée, serait un modèle européen de confiance. Notre conviction c'est qu'en unissant nos forces, nos données et nos ambitions, nous pouvons relever ce défi, et co-construire une IA générative performante digne de confiance.



> *D'un côté, les **BigTechs** exploitant **la position dominante des hyperscalers pour verrouiller l'innovation** en matière de données et d'IA, créant ainsi un marché fermé, dominé par des solutions intégrées (cloud, data, IA) reposant sur une utilisation opaque et souvent excessive des données.*
> *De l'autre, des écosystèmes **Multi Tech** ouverts et collaboratifs, fondés sur des données fiables partagées entre les acteurs **(Data Spaces),** intégrant dès leur conception les enjeux légaux, éthiques et de souveraineté, offrant ainsi un marché aux **offres de cloud de confiance avantagées par le multi-cloud.***

[4] Data Governance Act, Data Act, European Data Innovation Board

[5] https://gaia-x.eu/

[6] https://dssc.eu/

[7] https://internationaldataspaces.org/

[8] Smart middleware financé à hauteur de 150 M€ par la Commission européenne, développé par les consortiums Sovereign-X et InfrateX (initié par Digital New Deal Do Tank)

[9] "Accompagnement et transformation des filières », mutualisation de données pour les filières, Bpifrance

L'EXEMPLE FIL ROUGE DE LA FILIÈRE DU VOYAGE

Ce rapport offre aux entreprises des orientations stratégiques ainsi qu'une feuille de route opérationnelle concrète pour l'exploitation des technologies d'IA générative.

Les idées présentées dans ce rapport sont applicables et réplicables dans toutes les filières. Cependant, pour en faciliter l'appropriation, nous illustrerons nos recommandations avec un exemple fil rouge tiré de la filière du voyage. **Le projet Gen4Travel** regroupe divers acteurs de la filière qui collaborent à la création d'une boîte à outils commune d'IA générative. Ce commun numérique permettra à chaque acteur de la filière qui le souhaite de proposer un assistant IA (en anglais un "*AI Travel Assistant*" ou "*AI companion*", à ses clients pour simplifier l'expérience de voyage.

L'initiative de place technologique Gen4Travel est née et développée au sein du **Data Space tourisme et mobilité EONA-X**[10] (espace commun de partage des données regroupant de nombreux acteurs du voyage et des acteurs technologiques, avec notamment Accor, Aéroports de Paris, Aéroport Marseille Provence, Air France-KLM, Allianz, Amadeus, Anysolution, Apidae, Atout France, Atos, Capgemini, Compagnie des Alpes, Digital New Deal, Inria, Renault, SNCF,...).

[10] *Data space* labellisé France 2030 bénéficiant d'un financement de 20 M€

PRÉALABLE : DÉFINITION DE L'IA GÉNÉRATIVE

Comprendre l'IA générative en 4 étapes :

" L'IA générative est une branche de l'Intelligence Artificielle qui se concentre sur la création de contenu nouveau et original à partir de données existantes. Plutôt que de simplement analyser ou classer des données, les modèles d'IA générative sont capables de produire entre autre du texte, des images, des vidéos, du son ou encore du code informatique ."[11]

1. Collecte de données (Data collection) :
Les données sont la **matière première brute nécessaire pour entraîner l'IA**. Il s'agit de vastes ensembles de données, pouvant inclure du texte, des images, des vidéos, ou d'autres types d'informations (selon la ou les modalités supportées par le modèle), qui serviront de base pour que le modèle puisse apprendre.

2. Prétraitement des données (Processing) :
Cette étape implique la transformation et la **préparation des données brutes pour les rendre utilisables par l'algorithme.** Cela inclut le nettoyage des données, leur normalisation, et parfois l'extraction de caractéristiques spécifiques. Ce processus est crucial pour améliorer la qualité des données et maximiser l'efficacité des performances finales du modèle. c'est aussi la plus longue et délicate car elle doit permettre de vérifier la qualité des données et le respect des cadres réglementaires en vigueur (RGPD, droit d'auteur…)

3. Entraînement du Modèle (Model Training) :
Une fois les données préparées, elles sont utilisées pour pré-entraîner le modèle. Plusieurs architectures de modèle génératif existent.

La plupart des LLM (Large Language Models) s'appuient sur l'**architecture Transformer**[12] (ou "modèle transformateur" en français), qui a été dévoilée en 2017 dans l'article "Attention Is All You Need"[13] publié par les équipes de recherche de Google. Les transformers ont permis un bond d'efficacité sans précédent. Cette architecture apprend à partir des données, ajustant ses paramètres pour être capable de comprendre des motifs clés, des relations, et des structures particulières dans les données. Le résultat est un modèle capable de générer du contenu pertinent en se basant sur ce qu'il a appris. Après cette phase de pré-entraînement, le modèle passe par plusieurs étapes cruciales :

- **phase d'évaluation** : le modèle est testé sur un ensemble de données distinct de celui utilisé pour l'entraînement. Cette étape permet de mesurer ses performances et sa capacité à généraliser sur des données qu'il n'a jamais vues.

[11] Source ChatGPT 4o
[12] https://blogs.nvidia.com/blog/what-is-a-transformer-model/
[13] https://arxiv.org/abs/1706.03762

- **phase d'instruction (ou fine-instruct) :** le modèle est affiné sur des tâches spécifiques ou des domaines particuliers. Cette étape permet d'adapter le modèle à des usages précis et d'améliorer ses performances sur ces tâches.
- **phase de test :** une série de tests rigoureux est menée pour évaluer les capacités du modèle dans diverses situations. Cette phase permet de vérifier la robustesse du modèle, sa cohérence, et sa conformité aux objectifs fixés.

Ces étapes supplémentaires sont essentielles pour produire un modèle de langage performant, fiable et adapté aux besoins spécifiques des utilisateurs.

4. Inférence (Using the Model) :

Après l'entraînement, le modèle est mis en exploitation pour être utilisé pour faire des prédictions ou générer de nouveaux contenus à partir de nouvelles données ou de prompts[14]. Cette phase est celle où le modèle est appliqué dans des scénarios réels pour produire du texte, des images, du code, ou d'autres formes de sorties basées sur l'entrée fournie par l'utilisateur.

Rôle des LLM :

Lors de l'inférence, un LLM est utilisé pour générer du texte ou répondre à des prompts en temps réel. Grâce à l'entraînement préalable, le LLM peut produire des contenus cohérents, pertinents, et adaptés au contexte fourni par l'utilisateur.

Rôle des assistants et robots conversationnels :

L'exemple le plus connu d'assistant conversationnel est ChatGPT proposé par OpenAI. Il s'agit d'une application basée sur un LLM. Actuellement, ChatGPT 4o utilise le modèle GPT-4 pour sa version la plus récente, tandis que la version précédente reposait sur GPT-3.5. Alors que les LLM constituent le moteur central des traitements de l'IA générative, des assistants comme ChatGPT en sont l'interface visible, permettant aux utilisateurs d'interagir facilement avec l'IA à travers des conversations en langage naturelles

[14] Un prompt dans le contexte de l'IA générative est une entrée textuelle ou une consigne donnée au modèle pour générer une réponse. Il sert à orienter la production de contenu, qu'il s'agisse de texte, d'images ou d'autres formes de création générative.-

ÉVITONS LA
DEPENDANCE AU
« POWERED BY OPENAI »
COMME CELLE AU
« POWERED BY GOOGLE »
IL Y A 20 ANS

I. IA GÉNÉRATIVE : TOURNANT DÉCISIF POUR NOTRE SOUVERAINETÉ NUMÉRIQUE ?

Les LLM (Large Language Models), incluant par extension les modèles fondation qui sont plus généraux[15], forment la base technologique de l'IA générative, capable de produire de manière contextuelle des contenus textuels, visuels, multimédias et du code informatique. Ils s'appuient sur d'immenses volumes de données d'entraînement, ouvrant ainsi la voie à des innovations majeures dans tous les secteurs d'activités.

1.1. LLM : DOIT-ON S'ENGAGER DANS UNE QUÊTE À LA PUISSANCE ?

Les LLM représentent une avancée majeure dans le domaine de l'Intelligence Artificielle, mais ils posent déjà une question cruciale pour l'Europe : nos entreprises sont-elles prêtes à rivaliser ?

1.1.1. Les immenses défis des LLM

Le défi du gigantisme

La domination actuelle de quelques acteurs, principalement américains, menace non seulement la compétitivité européenne, mais aussi sa souveraineté numérique. Cette hégémonie s'explique en grande partie par les ressources financières considérables nécessaires pour :

- **la phase de collecte de données :** sachant que les plus grands acteurs disposent déjà de grandes masses de données sans avoir à les collecter chez d'autres, et disposent de tous les moyens nécessaires pour accéder à l'ensemble des données publiques (Web, Open Data, etc.) ;
- **les phases de préparation des données et d'entraînement :** très gourmandes en temps et en ressources machines ;
- **la phase d'inférence :** elle aussi particulièrement gourmande en ressources machines.

Selon Stanford[16], le **coût de développement d'un modèle fondation** de premier rang est colossal, se chiffrant **entre 10 et 100 millions d'euros**. Le coût estimé pour entraîner GPT-4 d'OpenAI serait d'environ 78 millions de dollars, et même 191 millions pour Google Gemini.

Ces montants impressionnants sont dus à la nécessité de monopoliser des milliers de GPU[17] pendant plusieurs semaines, ce qui exige une capacité de calcul que peu d'acteurs européens peuvent mobiliser à tout moment et de manière répétée, puisqu'il faut en effet recommencer pour l'entraînement de chaque nouveau modèle.

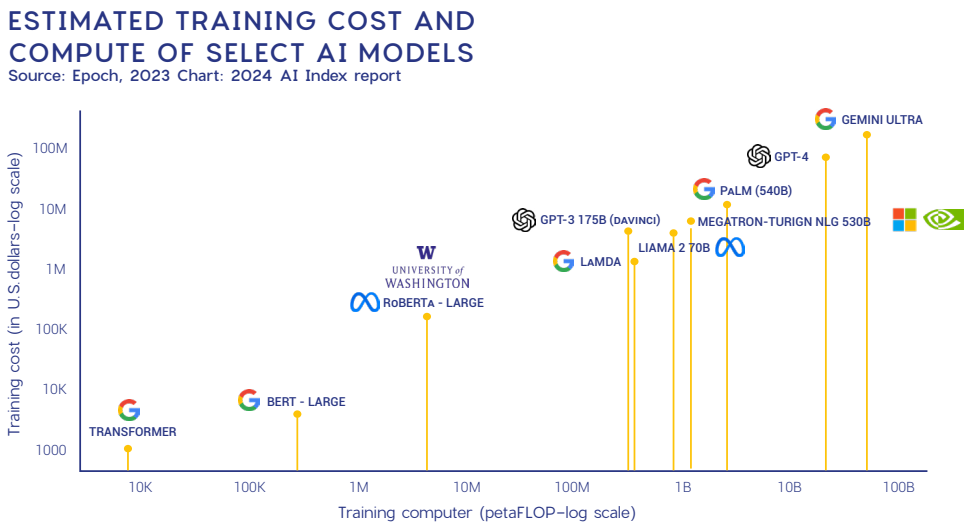
Cette situation impose une double peine : d'une part, les coûts conséquents liés à l'acquisition ou la location de ces ressources, et d'autre part, la nécessité d'atteindre une masse critique. Cette masse critique suppose soit la constitution d'une réserve importante, soit la possession d'un parc de GPU largement supérieur à la demande d'un seul acteur, pour atteindre une rentabilité

[15] Tous les LLM sont des modèles fondation, mais tous les modèles fondation ne sont pas des LLM. Les modèles fondation peuvent couvrir une gamme plus vaste de types de données (texte, image, etc.), tandis que les LLM se concentrent exclusivement sur la langue et la compréhension du langage. Par abus de langage, nous utilisons parfois le terme LLM pour qualifier les deux dans ce rapport.

[16] Institut HAI (Institute for Human-Centered Artificial Intelligence) de l'Université de Stanford dans son rapport 2024

[17] Un GPU (Graphics Processing Unit) est un composant électronique spécialisé, initialement conçu pour le traitement graphique, qui excelle dans les calculs parallèles. Cette capacité le rend particulièrement performant pour les tâches d'IA exigeantes en ressources, permettant une accélération significative des temps de traitement par rapport aux CPU traditionnels. En conséquence, le nombre de GPU est souvent utilisé comme indicateur de la puissance de calcul des infrastructures d'IA, reflétant la capacité d'une organisation à entraîner et déployer des modèles d'IA à grande échelle.

économique. Cela crée un effet de seuil critique, en particulier pour les opérateurs de cloud, rendant encore plus difficile pour l'Europe de rivaliser sur ce terrain.



Déjà un manque de données d'entraînement ?

Les LLM sont connus pour leur besoin considérable en données d'entraînement. Paradoxalement, ils font face à une pénurie croissante de données de qualité suffisante pour continuer d'améliorer leurs performances. Cette situation crée un défi majeur pour l'évolution de ces modèles, car leur efficacité dépend directement de la quantité et de la qualité des données utilisées pour leur entraînement.

Pour pallier ce problème, l'utilisation de données synthétiques émerge comme une solution envisageable. Ces données, générées artificiellement grâce à des algorithmes et des modèles avancés, offrent la possibilité de créer des ensembles de données vastes et variés sans être limité par la disponibilité des sources de données réelles.

Il est important de noter que la quantité de données nécessaire pour entraîner efficacement un LLM est considérable. L'article "Scaling Laws for Neural Language Models"[18] a établi une règle empirique cruciale à ce sujet. Selon cette étude, il faut à minima 20 fois plus de données que le nombre de paramètres visés par le modèle pour un entraînement optimal. Ainsi, pour un modèle comportant 70 milliards de paramètres nécessite au minimum 1,4 trillion (1 400 milliards) de tokens[19] de données d'entraînement. Si nous considérons qu'un livre moyen contient environ 50 000 mots (ou tokens), 1,4 trillion de tokens équivaudraient à environ 28 millions de livres. Pour mettre cela en perspective, c'est plus que le nombre total de livres dans la bibliothèque du Congrès américain, qui est la plus grande bibliothèque du monde avec environ 24 millions de livres catalogués.

Cette règle souligne l'ampleur du défi en termes de collecte et de gestion des données pour l'entraînement des LLM les plus avancés, et explique en partie pourquoi la génération de données synthétiques devient une option de plus en plus attrayante dans le domaine de l'Intelligence Artificielle.

[18] https://arxiv.org/abs/2001.08361

[19] Dans le domaine des LLM, un token est une unité de traitement du texte. Les modèles de langage ne traitent pas le texte entier mot par mot, mais le décomposent en tokens, qui sont des segments de texte. Ces tokens peuvent correspondre à des mots entiers, des sous-parties de mots, des groupes de mots (phrases courtes), ou même des caractères individuels, selon la complexité du texte et du modèle.

Néanmoins, il faut comparer les volumes de données nécessaires avec l'étendue de l'usage et la qualité du langage exigée pour cet usage. Ce qui ouvre la voie à des LLM spécialisés appuyés sur des jeux de données plus qualifiés, plus spécialisés s'adressant aux professionnels.

Le défi géopolitique

Les LLM sont actuellement entraînés et mis à disposition du grand public par un nombre restreint d'acteurs, principalement américains. Cette concentration soulève des préoccupations majeures de soft power pour l'Europe :

1. **Sous-représentation culturelle :**
   Les langues, valeurs et cultures européennes sont souvent marginalisées dans ces modèles, entraînant un biais anglo-saxon dans la génération de contenus et dans les résultats qui sont proposés à l'utilisateur.

2. **Risque de dépendance :**
   La centralisation des ressources (financières, computationnelles et de talents) crée un risque de dépendance technologique pour l'Europe, menaçant sa souveraineté numérique et sa compétitivité économique.

3. **Vulnérabilité mondiale :**
   La concentration excessive de ces technologies critiques les rend vulnérables aux cyberattaques, pouvant potentiellement paralyser des pans entiers de l'économie et de la société numériques à l'échelle mondiale. Sans compter les risques géopolitiques liés au fait par exemple que 90% des GPU sont fabriqués par TSMC à Taïwan[20].

Face à ces enjeux, il est crucial pour l'Europe de développer sa propre expertise en LLM, non seulement pour préserver son identité culturelle et son autonomie stratégique, mais aussi pourcontribuer à une (bio)diversité technologique mondiale, gage de résilience face aux risques cybernétiques.

Quelques "irréductibles gaulois" et pépites européennes font face à cette domination américaine avec des acteurs, tels que le français Mistral AI, qui montrent que le vieux continent représente toujours un potentiel important d'innovation. Ces entreprises innovantes (LightOn, Allonia, Giskard, Hugging Face,...) sont à l'avant-garde du développement de technologies d'IA sophistiquées et peuvent jouer un rôle crucial dans la création d'un écosystème technologique français et européen robuste.

Des initiatives communautaires, à l'image d'OpenLLM France[21] (initiée par Linagora), développent des LLM et d'autres communs numériques d'IA générative sous licence Open Source pour l'intérêt commun général. Il est essentiel que les gouvernements et les institutions européennes soutiennent cet écosystème dynamique à travers des politiques de financement, de collaboration et de régulation favorables aux entrepreneurs[22] pour permettre à l'Europe de devenir un leader mondial dans le domaine de l'Intelligence Artificielle.

Les LLM soulèvent aussi des questions sur la souveraineté linguistique et cognitive. Si l'Intelligence Artificielle devient le principal outil par lequel nous accédons à des connaissances et en générons de nouvelles, comment préserverons-nous la diversité linguistique et la richesse des cultures locales face à des écosystèmes technologiques dominés par une poignée d'acteurs non européens ? La centralisation des modèles de langage au sein de grandes entreprises pose également la question du contrôle de l'information et de la pluralité des points de vue.

[20] TSMC is making the best of a bad geopolitical situation, The Economist, janvier 2023

[21] Communauté lauréate de l'appel à projets France 2030 "Communs Numériques pour l'IA Générative" : https://www.openllm-france.fr/

[22] Open source software and global entrepreneurship, Science Direct, Elsevier, novembre 2023

Le défi énergétique et environnemental

Sommes-nous (déjà) au bout du modèle ? Selon Cédric Villani[23], nous avons atteint les limites du modèle actuel de développement de l'IA en général. Les approches basées sur des modèles gigantesques, extrêmement gourmands en énergie et en données, montrent leurs limites. Ces modèles nécessitent une quantité de ressources insoutenable à long terme, tant sur le plan écologique qu'économique : **l'entraînement d'un seul grand modèle de langage comme GPT-3 peut consommer environ** 1,287,000 kWh d'électricité[24], soit **l'équivalent de la consommation annuelle de 320 foyers européens** moyens et peut générer jusqu'à 552 tonnes de CO_2[25], soit l'équivalent des émissions annuelles de 112 véhicules thermiques standards ou 205 vols aller-retour entre Paris et New-York.

De plus, la compétition pour l'accès aux ressources énergétiques entre les data centers dédiés à l'IA et les besoins humains soulève des inquiétudes croissantes. Un rapport Forrester[26] cite deux exemples alarmants : lors d'une sécheresse au printemps 2021, le gouvernement taïwanais a donné la priorité à l'eau destinée à l'agriculture et à la consommation humaine plutôt qu'au refroidissement de l'usine TSMC[27], obligeant le plus grand fabricant de puces à réduire sa production de 85 %. En juillet 2022, les *hyperscalers*[28] ont fermé les centres de données pendant plusieurs heures en raison d'un pic de chaleur près de Londres.

Un autre exemple éclairant : dans son dernier rapport environnemental, Microsoft a révélé que sa consommation mondiale d'eau a grimpé de 34 % entre 2021 et 2022 pour atteindre près de 6,3 milliard de litres, soit l'équivalent de plus de 2500 piscines olympiques. Cela représente une forte augmentation par rapport aux années précédentes à laquelle des chercheurs extérieurs attribuent ses recherches en IA. Phénomène mis en avant par le procès entre Microsoft et la ville de Des Moines dans l'Iowa[29] pour la consommation massive d'eau pour refroidir les data centers, mettant ainsi en difficulté la population locale durant les périodes de sécheresse.

Par ailleurs, Shaolei Ren, chercheur à l'Université de Californie à Riverside, a tenté de calculer l'impact environnemental des produits d'IA générative tels que ChatGPT : il estime par exemple que ChatGPT "avale" 500 ml d'eau chaque fois que vous lui posez une série de 5 à 50 questions (phase d'inférence). Ces chiffres mettent en lumière la nécessité urgente de trouver un équilibre entre les avancées technologiques et les besoins énergétiques de base de la société.

L'Europe doit donc explorer des alternatives, comme des modèles plus petits et spécialisés, et/ou des approches hybrides et mutualisées combinant différentes technologies pour réduire la dépendance aux ressources tout en augmentant l'efficacité et l'adaptabilité des systèmes d'IA. C'est ce que nous explorons dans la suite de ce rapport (§2 et §3).

Le défi de cybersécurité

Les LLM peuvent présenter des failles de sécurité exploitables, avec la question particulière de la protection des informations sensibles (risque de fuites de données confidentielles via l'utilisation de LLM externes). Les LLM peuvent être des cibles potentielles pour des cyberattaques, notamment de la part d'États ou d'acteurs malveillants cherchant à compromettre la sécurité des données européennes. En utilisant ces modèles, les entreprises européennes pourraient involontairement exposer leurs données sensibles à des risques de piratage.

[23] *Thinkerview, 8 décembre 2023*

[24] Étude menée par des chercheurs de l'Université de Copenhague en 202"; Heidi News

[25] Étude "Carbon Emissions and Large Neural Network Training"de l'université de Californie à Berkeley

[26] Forrester's State of AI Report Suggests a Wave of Disruption Is Coming, HPC wire, février 2024

[27] Taiwan Semiconductor Manufacturing Company

[28] Un hyperscaler désigne une entreprise ou une infrastructure capable de scaler (c'est-à-dire augmenter ou réduire) rapidement et massivement ses ressources informatiques, notamment dans le domaine du cloud computing. Une utilise ce terme en particulier pour désigner les géants du cloud (AWS, Azure, Google Cloud, etc.)

[29] *https://www.lemondeinformatique.fr/actualites/lire-la-consommation-d-eau-liee-a-l-ia-generative-inquiete-91508.html*

Il est crucial d'insister sur la nécessité de disposer d'une diversité de modèles et de fournisseurs pour assurer la résilience du système. Une dépendance excessive envers un nombre limité de modèles détenus par quelques entreprises crée un risque systémique majeur. En cas d'attaque réussie ou de défaillance d'un de ces modèles dominants, l'impact pourrait être dévastateur à l'échelle mondiale, entraînant des interruptions de service généralisées (comme nous l'avons vu en juillet 2024 avec l'incident crowdstrike[30]) et compromettant la sécurité des données à grande échelle.

De plus, il est impératif de prévenir dès maintenant les risques émergents tels que les "prompt injections". À l'instar des attaques par injection SQL pour les bases de données, ces techniques pourraient permettre d'extraire des données sensibles contenues dans le modèle ou d'injecter des commandes malveillantes, conduisant à la génération de contenus ou d'actions non désirés. Cette menace souligne l'importance cruciale de maîtriser non seulement les modèles eux-mêmes, mais aussi les données utilisées pour leur apprentissage.

Ces défis en cybersécurité soulèvent la nécessité de renforcer les infrastructures de sécurité et d'établir des cadres réglementaires stricts afin de garantir que l'utilisation des LLM respecte les normes européennes en matière de sécurité et de protection des données.

1.1.2. Privilégier la confiance à la puissance

Révolutionnaire et inquiétant

Le lancement inattendu de ChatGPT en 2022 (basé sur le modèle GPT-3.5) a propulsé l'IA dans une ère "grand public", à la surprise même des experts les plus avisés. La technologie s'est rapidement répandue (1 million d'utilisateurs en 5 jours, 100 millions en 3 mois), générant un enthousiasme massif et bouleversant l'ensemble du secteur, et même au-delà.

Ce déploiement global, bien que réellement révolutionnaire (un terme souvent galvaudé, mais ici tout à fait pertinent), n'a pas été exempt de problèmes majeurs et de polémiques. Il reflète une approche caractéristique des États-Unis, où la priorité est donnée à la rapidité d'exécution, parfois au détriment de la prudence.

L'Europe, bien qu'elle puisse peiner à égaler une telle rapidité, a l'opportunité de se démarquer en adoptant une approche unique. En capitalisant sur ses points forts, plutôt qu'en tentant de pallier ses lacunes, L'Europe peut concevoir des solutions d'IA adaptées aux besoins spécifiques de son marché, tout en respectant les normes les plus strictes en matière de fiabilité et d'éthique. **En faisant de la "confiance" l'axe central du développement des Intelligences Artificielles génératives, l'Europe peut non seulement se distinguer, mais aussi s'affirmer comme un leader mondial dans ce domaine.**

Exigence d'éthique

Construire une IA européenne exige de concevoir des systèmes éthiquement responsables, qui respectent nos valeurs. Cela signifie intégrer dès le départ "by design" des principes essentiels tels que la transparence, l'équité, et la protection de la vie privée. Pour offrir une alternative crédible, une IA européenne doit développer un écosystème de confiance en axant ses travaux sur l'intégration des normes et réglementations spécifiques de l'Union Européenne, telles que le RGPD, l'IA Act, le DMA, le DSA, la protection de la propriété intellectuelle et les règles de concurrence en général.

[30] https://www.cert.ssi.gouv.fr/actualite/CERTFR-2024-ACT-032/

Exigence de fiabilité

La confiance repose également sur la fiabilité des solutions offertes, qu'elles soient destinées aux entreprises ou aux particuliers. Les nombreuses "hallucinations"[31] produites par les premières technologies d'IA générative compromettent sérieusement leur crédibilité dans des contextes professionnels, où l'exigence de précision et de qualité est primordiale.

Ces erreurs manifestes, qui consistent en la génération de réponses incorrectes ou incohérentes, sans que l'assistant IA ne s'en rende compte, posent un problème majeur de fiabilité. Cette difficulté est d'autant plus préoccupante que, par leur conception basée sur des réseaux de neurones profonds[32], ces IA sont encore complexes à ajuster et à rendre transparentes. Certains experts parlent même d'un phénomène de "boîte noire", soulignant le manque de compréhension et d'explicabilité des décisions prises par les modèles. Ces limitations techniques, difficilement maîtrisables à ce stade, disqualifient ces technologies pour de nombreuses applications d'entreprise nécessitant une rigueur absolue, et où la moindre erreur peut avoir des conséquences graves tant pour les organisations que pour les personnes (santé, finance, administration, commerce, etc.).

Exigence de construction d'écosystèmes de confiance Cloud-Data-IA

Au-delà des spécialistes de l'IA générative comme OpenAI, et des outils de data collaboration (ex: Snowflake, Databricks), les acteurs dominants du marché aujourd'hui sont les hyperscalers américains tels que Amazon, Microsoft et Google. Ces entreprises globalisées se distinguent par leur capacité à enfermer leurs clients "lock-in" dans des écosystèmes intégrés "tout-en-un", offrant une gamme complète, issue du cloud public, de services Cloud-Data-IA. Même les acteurs traditionnels de l'infrastructure physique de l'IA comme NVIDIA commencent à remonter la chaîne de valeur en rachetant des champions de l'IA générative pour les entreprises[33].

Le défi pour l'Europe ne réside donc pas uniquement dans le développement de technologies spécifiques d'IA et d'IA générative de confiance, mais dans la création d'écosystèmes complets, alliant Cloud, Data et IA, basés sur des chaînes de valeur cohérentes et alignées avec les principes que l'Europe souhaite promouvoir et défendre.

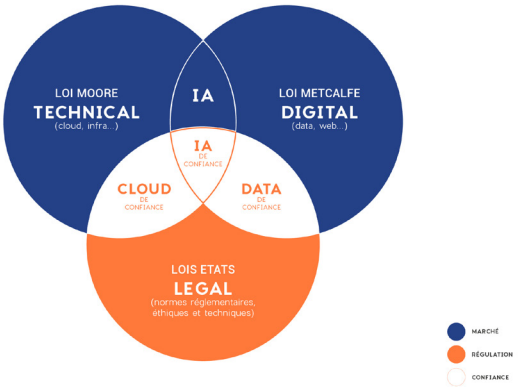


Schéma "Écosystème de confiance"extrait du rapport

"IA de confiance : opportunité stratégique pour une souveraineté numérique et industrielle[34]"

[31] Une hallucination en IA générative désigne un phénomène où le modèle produit des informations incorrectes, non factuelles ou totalement inventées. Bien qu'il génère du contenu qui semble cohérent, celui-ci peut être faux ou trompeur.

[32] Un réseau de neurones est un modèle informatique inspiré du cerveau humain, composé de multiples couches de "neurones" artificiels qui traitent des informations et ajustent leurs connexions (poids) en fonction des données d'entrée. Il est difficile de comprendre précisément comment il fonctionne parce que ces réseaux, en particulier les plus complexes, contiennent des milliers ou millions de paramètres qui interagissent de manière non linéaire. ChatGPT 4o

[33] https://www.forbes.fr/technologie/nvidia-rachete-octoai-et-domine-desormais-les-solutions-dia-generative-pour-les-entreprises/

[34] Julien Chiaroni et Arno Pons, Digital New Deal - France 2030, juin 2022

HUMAINS ET MACHINES NE PARTAGENT PLUS SEULEMENT DES DONNÉES, MAIS UN LANGAGE COMMUN

1.2. DE LA RÉVOLUTION DIGITALE À LA RÉVOLUTION LINGUALE ?

L'avènement des LLM constitue un tournant majeur dans l'évolution de l'Intelligence Artificielle, s'apparentant à une véritable révolution linguistique où le langage humain prend le pas sur le langage machine. Alors que l'ère du Big Data (années 2000 et 2010) a d'abord été marquée par des outils de visualisation de données (dataviz), permettant de rendre l'information accessible visuellement, elle a ensuite évolué avec des outils d'analyse qui ont permis aux data analysts et data scientists de rendre ces données plus compréhensibles et exploitables.

Aujourd'hui, **avec les LLM, nous entrons dans une nouvelle phase : ces modèles ne se contentent plus de rendre les données intelligibles, ils les rendent "intelligentes".** Les IA basées sur les LLM permettent une interaction fluide et naturelle avec les humains, simplifiant ainsi la manipulation de l'information, et ouvrant des perspectives inédites en matière d'automatisation et de prise de décision, sans nécessiter d'expertise technique avancée. Cela bouleverse non seulement la façon dont nous interagissons avec les données, mais également l'ensemble des processus métiers, en facilitant l'accès à une Intelligence Artificielle plus intuitive et accessible.

La véritable révolution réside dans le fait que les LLM rendent désormais possible une interaction humain-machine en langage naturel. La communication numérique, traditionnellement via le clavier et la souris, évolue vers une communication "linguale", proche de celle que les humains pratiquent entre eux depuis des millénaires. Cette avancée ouvre la voie **à une nouvelle forme d'intelligence collective, où humains et machines ne partagent plus seulement des données, mais un langage commun.** Cela promet de transformer en profondeur notre manière de collaborer avec les machines, en passant de simples interactions avec des interfaces utilisateur souvent rigides et lourdes, à des échanges fluides, riches et nuancés avec des entités "intelligentes", capables de saisir des subtilités linguistiques et contextuelles complexes.

1.2.1 Les assistants, potentielles nouvelles interfaces humain-machine

Les assistants conversationnels, tels que Siri, Google Assistant et Alexa, ont considérablement progressé depuis leur lancement, évoluant de simples dispositifs de commande vocale à des systèmes plus sophistiqués capables de traiter une large gamme de requêtes. Cependant, ils ont souvent déçu en ne parvenant pas à offrir des services véritablement utiles au quotidien. Demander l'heure ou la météo à la voix ne suffit visiblement pas à capter l'intérêt des foules.

À ce jour, en dehors des grandes plateformes, ces assistants restent peu répandus. On les retrouve parfois sur des sites web pour des fonctions de support client ou de service après-vente, mais leur utilisation se limite principalement à ces tâches spécifiques, avec des performances encore perfectibles. Conscient des limites de son modèle actuel et face au potentiel révolutionnaire de l'IA générative, Amazon a annoncé un plan social incluant des licenciements au sein de sa division Alexa, témoignant des difficultés rencontrées pour transformer l'assistant en un outil indispensable. Cette restructuration a pour objectif de moderniser son assistant vocal vieillissant et de l'adapter à l'ère de l'IA générative[35].

Cette évolution pose la question de savoir si les technologies d'assistants conversationnels constituent simplement une avancée progressive, restreinte à quelques usages spécifiques, ou bien une véritable révolution dans notre façon d'interagir avec les machines. Leur aptitude à simuler des conversations humaines, à apprendre des interactions passées et à intégrer les avancées de l'Intelligence Artificielle laisse toutefois entrevoir une transformation profonde de notre relation avec la technologie.

[35] L'usine Digitale, 20 novembre 2023

Les récents progrès de l'IA générative pourraient finalement concrétiser le potentiel, pour le moment non réalisé, de ces assistants, rendant leur utilité au quotidien plus palpable et significative.

Vers un futur sans écran

Bien que nos interfaces actuelles aient encore de beaux jours devant elles, les assistants conversationnels ont le potentiel de devenir l'interface dominante pour l'accès à l'information et aux services. Ils offrent une interaction plus fluide et naturelle que les écrans, claviers ou souris que nous utilisons actuellement. Les utilisateurs pourraient poser des questions, effectuer des achats, organiser des voyages, prendre des rendez-vous médicaux ou gérer leurs tâches quotidiennes simplement en dialoguant avec leurs appareils via le langage naturel, notamment par la voix, rendant ainsi l'expérience plus intuitive et accessible.

Les entreprises et administrations européennes doivent se préparer à cette transition en adaptant leurs services et leur présence numérique pour être compatibles avec ces interfaces conversationnelles. Cette évolution ne se limite pas à transformer la manière dont nous accédons aux informations et aux services, elle va redéfinir également la façon dont les organisations conçoivent et distribuent leurs produits et services, ouvrant la voie à de nouvelles opportunités dans l'économie numérique.

Désintermédiation et concentration du marché

Si les assistants conversationnels deviennent l'interface principale pour les utilisateurs, les entreprises risquent de perdre progressivement le contact direct avec leurs clients. Pourquoi passer du temps sur plusieurs sites web spécifiques, avec chacun une arborescence de pages complexe et des formulaires et clics sans fin, si mon assistant peut répondre à tous mes besoins directement et sans effort ? La désintermédiation qu'apportent ces assistants va bien au-delà de ce que nous avons connu avec les grandes plateformes monopolistiques de l'ère qui s'achève.

La question qui peut se poser est de **savoir si nous assisterons à l'émergence d'un assistant unique écrasant le marché** (comme ChatGPT, Apple Intelligence/Siri, Gemini, Copilot, ou autre), ou si les fonctions d'assistants seront plutôt distribuées et intégrées dans une grande partie des sites web et applications mobiles de notre quotidien.

La seconde option serait préférable, car elle éviterait une concentration des pouvoirs sans précédent. Aujourd'hui, le web attribue un rôle central, de "gatekeeper", aux moteurs de recherche, tout en laissant aux entreprises la possibilité de maintenir une présence en ligne via leur propre site web, ou application mobile. Cela pourrait devenir beaucoup plus difficile si les utilisateurs finaux, les individus, réalisaient toutes leurs actions via un assistant unique, **reléguant ainsi les entreprises à de simples fonctions "backend", invisibles du client final et entièrement dépendantes de l'assistant.** Quoi qu'il en soit, l'Internet que nous connaissons est probablement sur le point d'être profondément transformé.

DÉMOCRATISER LES ASSISTANTS IA POUR TOUTE LA FILIÈRE

Le projet **Gen4Travel** a pour principal objectif de permettre à chaque acteur de la filière du voyage d'accéder aux technologies d'IA génératives : compagnies aériennes et ferroviaires, gares, ports, aéroports, sociétés de taxi, de bus, transports en commun, musées, parcs d'attraction, lieux touristiques divers, offices de tourisme, etc. La filière, à l'échelle de l'Europe, est composée de quelques acteurs privés et publics structurants, et d'une multitude de petits acteurs (+95%[36]).

Depuis de nombreuses années, **la filière est largement désintermédiée** par quelques grandes plateformes (Booking, Google, Airbnb, Tripadvisor, Uber, etc.).

Si les assistants IA deviennent la norme d'interaction entre un voyageur et ses services de voyage, et qu'un acteur monopolistique devient le Travel Assistant unique de tous les voyageurs (et la probabilité pour que cet acteur ne soit pas européen est élevée), **la désintermédiation risque de devenir quasi totale,** coupant ainsi les acteurs du voyage de la relation directe avec leurs clients. Ces acteurs ne seront plus que des fournisseurs de données pour le Travel Assistant dominant.

L'objectif du projet Gen4Travel est de **rendre les technologies d'IA générative accessibles à l'ensemble de la filière,** dans le cadre d'une gouvernance équitable, assurée par le secteur lui-même via le Data Space EONA-X, et non par un acteur technologique monopolistique.

Il est essentiel que tous les acteurs de la filière puissent y participer, y compris les plus petits, souvent moins digitalisés et donc plus exposés à la désintermédiation. Même ceux qui n'ont pas les moyens financiers et humains de développer et opérer un assistant IA par eux-mêmes, doivent pouvoir proposer leurs offres dans ce nouvel environnement, sans subir les lourdes commissions imposées par les grandes plateformes actuelles[37].

[36] Tourism Data Space Blueprint - https://www.tourismdataspace-csa.eu/wp-content/uploads/2024/01/DRAFT-BLUEPRINT-Tourism-Data-Space-v3.3_final.pdf

[37] Les grandes plateformes d'intermédiation (OTA) comme Booking.com représentant environ 40 % des transactions globales, appliquent des commissions moyennes de 15 %, pouvant atteindre jusqu'à 30 % pour les hôtels. Source : Livres blancs *"Partage des données et tourisme"*, Bpifrance Le Lab et Digital New Deal, juin 2020

ARTIFICIAL CAPABLE INTELLIGENCE : L'ÈRE DES ACTIONS AUTONOMES

1.2.2 L'agentification, la nouvelle frontière de l'IA générative

De nombreuses figures, telles que l'historien Yuval Noah Harari, ou encore l'entrepreneur Elon Musk, avaient anticipé la transition de l'IA générative, la voyant évoluer d'un simple outil de génération de contenu, à un véritable agent autonome. Dans son dernier livre, The Coming Wave (2023), Mustafa Suleyman, cofondateur de Google DeepMind et aujourd'hui à la tête de Microsoft IA, dévoile sa vision de la prochaine grande étape de l'Intelligence Artificielle. Il y introduit le concept de l'ACI ou **"Artificial Capable Intelligence"**, une IA non seulement en mesure de dialoguer en langage naturel avec les humains, mais aussi **capable de réaliser de manière autonome des tâches complexes avec une supervision minimale.**

Selon Suleyman, cette ACI, dotée d'une mémoire, et hautement interactive, sera apte à comprendre le contexte spécifique d'un utilisateur, et à amplifier les capacités de celui-ci, en effectuant en son nom des actions concrètes dans le monde réel (réserver, organiser, réaliser des tâches, etc.). Les entreprises pourront également tirer parti de ces capacités pour automatiser des processus internes ou des opérations complexes (gestion de chaînes logistiques, de la relation client, des processus administratifs et financiers, etc.).

En dotant les assistants conversationnels actuels de capacités de mémoire, d'intégration avec les Systèmes d'Information des entreprises et administrations, et de compétences pour **enchaîner des raisonnements complexes ("chains of thoughts")**, ces outils pourraient se transformer en véritables agents intelligents. Ils seraient alors en mesure d'assister les individus dans toutes leurs interactions, qu'elles soient personnelles (citoyens, consommateurs, etc.) ou professionnelles (employés, dirigeants, etc.). Cette évolution promet des gains de productivité jamais vus auparavant, tout en nous invitant à repenser notre relation avec la technologie, à mesure que la frontière entre l'humain et la machine devient de plus en plus floue.

L'IA générative a déjà transformé la manière dont nous interagissons avec la technologie, permettant aux assistants conversationnels de communiquer de manière fluide et naturelle. Cependant, **une révolution encore plus importante se profile : le passage des Large Language Models (LLM) aux Large Action Models (LAM[38]).** Nous explorons cette évolution en détail plus loin dans ce rapport, notamment dans la section dédiée aux LAM (§2.3).

[38] Les Large Action Models (LAM) sont une évolution des modèles de langage actuels, conçus non seulement pour comprendre et générer du texte, mais aussi pour exécuter des actions concrètes de manière autonome.

UNE VÉRITABLE AGENCE DE VOYAGE DANS VOTRE POCHE

Un assistant de voyage IA peut offrir une multitude de fonctionnalités. Certaines se limitent à afficher du texte ou du contenu multimédia plus ou moins contextualisé, pour lesquelles les LLM actuels sont suffisants. tionnalités, en revanche, nécessitent que l'assistant agisse directement au nom du voyageur, avec son autorisation mais via une supervision minimale de ce dernier, ce qui relève de l'agentification.

Avec **Gen4Travel,** les assistants de voyage IA pourront offrir les fonctionnalités suivantes :

Planification et réservation de voyage :
Recherche de vols, trains, hôtels, activités touristiques diverses
Comparaison de prix
Itinéraires personnalisés en fonction des goûts et préférences
Réservation automatique en temps réel ou en différé

Gestion du voyage :
Mises à jour en temps réel en cas de perturbation, disruption, annulation
Notifications et rappels d'agenda et de formalités à effectuer
Suivi des bagages
Démarches administratives et documents (passeport, boarding pass, etc.)

Personnalisation et recommandations :
Suggestions personnalisées et recommandations
Guides locaux

Assistance pendant le voyage :
Réorganisation du voyage en cas de perturbation, disruption, annulation
Traduction en temps réel des contenus multimédias
Navigation multimodale et accès billetterie transport local

Support client et assistance immédiate :
Service client 24/7 multi-services
Règlement des litiges et liens avec les assurances
Sécurité, assistance et urgences (services médicaux, ambassades, etc.)

Gestion des préférences utilisateur :
Historique et préférences (budget, alimentation, PMR, goûts, etc.)
Comptes de fidélité et avantages

Expérience post-voyage :
Archivage et souvenirs
Avis client

1.3. QUELS SONT LES SCÉNARIOS SOUHAITABLES POUR NOS FILIÈRES ?

La course à la puissance dans le développement des LLM est un phénomène global, principalement mené par des entreprises disposant de vastes ressources. Ces modèles, tirant parti de quantités massives de données et d'infrastructures de calcul impressionnantes, atteignent des performances remarquables. Toutefois, pour les acteurs européens, une question se pose : faut-il s'engager dans cette course effrénée ? Rejoindre cette compétition nécessite des investissements conséquents en technologie, données et ressources humaines. Il devient alors essentiel de mettre en balance les bénéfices potentiels en matière de compétitivité et d'innovation, avec les coûts et risques inhérents à cette démarche.

1.3.1. La bataille des LLM, entre adoption et adaptation

Les LLM sont déjà des produits

Comme l'a souligné dans Yann Le Cun, Chief AI Scientist chez Meta, dans une récente interview pour GDIY[39], le développement des LLM n'est plus uniquement l'apanage des équipes de R&D. Les LLM sont désormais considérés comme des produits à part entière, intégrés rapidement dans les suites de services des géants technologiques (GAMAM[40]). Ces modèles sont en train de trouver leur place en un temps record dans divers domaines : la recherche sur internet avec Gemini de Google, les interactions sociales chez Meta, et la bureautique avec Copilot de Microsoft. Les LLM deviennent ainsi des éléments clés des offres de services de ces entreprises, transformant rapidement de nombreux aspects de la technologie et de notre quotidien.

Des LLM déjà utiles mais encore limités

Yann LeCun souligne également que le LLM ne font finalement que prédire le "token[41] suivant" (prédiction sémantique), sans véritable compréhension du problème posé et des concepts sous-jacents. Les LLM ne comprennent pas le contenu des images, n'ont pas de sens commun pour interpréter le monde qui les entoure, et sont actuellement incapables de construire des raisonnements complexes explicables. Selon lui, les LLM peuvent continuer à jouer un rôle majeur, mais leur potentiel seul est déjà limité. LeCun propose qu'ils soient utilisés comme composants de systèmes d'IA plus vastes et polyvalents, où d'autres architectures et algorithmes étendraient leurs capacités. Yann LeCun appelle les communautés de recherche et de R&D à explorer de nouvelles architectures pour aller au-delà des limitations actuelles des LLM.

L'assistant ChatGPT-o1, lancé par OpenAI en septembre 2024, représente peut-être un véritable progrès dans la direction souhaitée par LeCun, en proposant des capacités de raisonnement nettement améliorées. Ce modèle se distingue de ses prédécesseurs par un temps de réflexion plus long lors des interactions, permettant une analyse plus poussée avant de fournir une réponse. ChatGPT-o1 est en mesure de décomposer son raisonnement en plusieurs étapes ("chains of thoughts"), de vérifier ses réponses avant de les soumettre, et de présenter à l'utilisateur un rapport détaillé des étapes suivies pour arriver à une conclusion (explicabilité), renforçant ainsi la précision et la transparence du processus décisionnel.

[39] *Génération Do It Yourself, 5 juin 2024*

[40] GAMAM : Google, Apple, Meta, Amazon, Microsoft

[41] Un token est une unité de base pour analyser et traiter le texte. Généralement, un token représente un mot, une ponctuation, ou un autre élément du langage, selon le contexte et l'objectif de l'analyse.

Créer, Adopter ou Adapter

Dans la bataille de l'IA générative, 3 approches sont possibles pour les entreprises européennes :

1. Création : l'entreprise développe et entraîne son propre modèle, parfaitement adapté à son contexte et à celui de ses clients.

2. Adoption : l'entreprise intègre directement un modèle généraliste existant (ex : GPT-4, LLaMA 2, Claude 2, Mistral 7B, etc.).

3. Adaptation : l'entreprise adapte un modèle générique existant à son contexte spécifique ou à celui de ses clients (spécialisation des modèles) via des techniques comme le fine-tuning ou le RAG (voir §2.2).

Il n'est plus pertinent, à ce stade, de développer un modèle entièrement nouveau pour chaque entreprise, tant d'un point de vue économique qu'écologique.

Opter pour des LLM généralistes proposés par les grands acteurs technologiques, "powered by...", offre des solutions rapides et éprouvées pour chaque entreprise, mais peut entraîner une dépendance technologique et une perte de contrôle. Ces modèles, conçus par des tiers, ne reflètent pas toujours nos valeurs et ne sont pas forcément adaptés à nos langues ou à nos spécificités culturelles et sectorielles.

Adapter ces modèles aux contextes locaux et sectoriels permet de mieux répondre aux besoins spécifiques des industries européennes tout en respectant les particularités de chaque filière, et en préservant une certaine souveraineté technologique. Bien que cette spécialisation engendre des coûts supplémentaires, parfois significatifs, elle optimise les modèles existants pour des applications ciblées, assurant ainsi un équilibre entre performance et adéquation aux réalités locales.

1.3.2. Adapter les LLM pour une filière

Pour chaque filière, les bénéfices et usages des LLM peuvent être très spécifiques :

- **Dans le secteur de l'éducation,** ces modèles peuvent être utilisés pour créer des environnements d'apprentissage personnalisés et interactifs.
- **Dans le domaine de la santé,** ils peuvent améliorer le diagnostic et le suivi des patients en fournissant des analyses rapides et précises des données médicales. Ils peuvent aussi faciliter les démarches, tant pour le personnel de santé que pour les patients.
- **Dans les domaines de la mobilité et du tourisme,** ils permettent de fournir au voyageur des assistants, ou compagnons, capables de comprendre ses envies et ses besoins et de lui fournir des services et itinéraires personnalisés.
- Chaque filière a la possibilité d'emprunter une voie alternative en s'organisant pour coopérer et mutualiser ses ressources, permettant ainsi de réaliser l'adaptation/spécialisation des LLM, souvent onéreuse, à l'échelle de toute la filière plutôt qu'au niveau de chaque entreprise, de manière isolée.

Cette approche collaborative permet de partager les coûts et les efforts d'optimisation, tout en assurant que les modèles soient mieux adaptés aux besoins spécifiques de la filière. En unissant leurs forces, les entreprises d'une même filière peuvent non seulement bénéficier d'économies d'échelle, mais aussi développer des solutions plus cohérentes et performantes, renforçant ainsi leur indépendance technologique.

Si la filière se mobilise, l'idée même de créer son (ou ses) propre modèle devient réaliste, car les coûts et l'impact écologique sont largement répartis.

Cette démarche favorise également l'innovation collective, offrant à toutes les parties prenantes un accès à des technologies avancées, tout en réduisant la dispersion des efforts et des investissements. En mutualisant les ressources, la filière génère une dynamique de coopération qui accélère l'adoption de l'IA, tout en renforçant sa compétitivité face aux acteurs internationaux, notamment américains.

DES LLM POUR ET PAR LA FILIÈRE DU VOYAGE

Dans le cadre du projet Gen4Travel, les acteurs de la filière, sous la gouvernance du Data Space EONA-X, se sont engagés à développer ensemble une plateforme multi-LLM répondant aux spécificités de leur secteur.

Ce projet favorise **la mutualisation des coûts et le partage d'expériences et de données d'entraînement, réduisant ainsi les dépenses individuelles tout en augmentant l'efficacité globale** grâce à la participation de nombreux contributeurs.
L'option de s'appuyer sur un LLM unique n'est pas retenue pour le moment, afin de permettre à chaque acteur de choisir et exploiter le modèle qui lui convient, y compris des modèles plus génériques ou non européens.

Toutefois, pour ceux qui le souhaitent, l'entraînement ou le réentraînement des modèles pourra être mené collectivement au niveau de la filière, sous la gouvernance du Data Space.

L'APPROPRIATION DES LLM PAR LES DIFFÉRENTES FILIÈRES PEUT SUIVRE TROIS SCÉNARIOS :

1. "Powered by others" (utilisation d'un modèle généraliste existant propriétaire ou open source) :

Les acteurs économiques opèrent de manière individuelle, et adoptent directement des modèles existants.

→ *Assure une exécution rapide, mais limite l'adaptation aux spécificités métiers et risque de recréer des dépendances similaires à celles du modèle "powered by Google" qui prévalait il y a 20 ans.*

2. "Do It Yourself" (nouveau modèle personnalisé/adapté pour et par chaque acteur) :

Les acteurs économiques opèrent de manière individuelle en créant un modèle spécifique ou en adaptant et personnalisant un modèle existant à leurs besoins par leurs propres moyens.

→ *Permet une adaptation fine au contexte métier, mais impose à chaque acteur de supporter les coûts très élevés liés à la création ou, un peu moins élevés, à la spécialisation d'un modèle.*

3. "Join Forces" (nouveau modèle personnalisé/adapté pour et par une filière) :

Les acteurs économiques se regroupent en consortium (comme un Data Space), par exemple au sein d'une filière, pour développer ou adapter ensemble des modèles.

→ *Permet d'adapter les modèles au contexte métier tout en mutualisant et réduisant les coûts pour chaque participant. Renforce également la souveraineté des filières concernées.*

EN PLAÇANT LA « CONFIANCE » AU CŒUR DES IA GÉNÉRATIVES, L'EUROPE PEUT SE DÉMARQUER ET DEVENIR UN LEADER MONDIAL

II. CO-PRODUIRE UNE IA GÉNÉRATIVE DE CONFIANCE

2.1 FACILITER LA COOPÉRATION GRÂCE À L'OPEN SOURCE

2.1.1. Une stratégie Open Source de différenciation pour l'Europe

La stratégie de l'Open Source offrirait à l'Europe une véritable opportunité de se démarquer, plutôt que de suivre les tendances technologiques dominantes. **En misant réellement sur l'Open Source dans l'IA, l'Europe pourrait se positionner comme le champion de la transparence, de la coopération et de l'innovation ouverte**, permettant notamment aux filières de co-investir dans des infrastructures technologiques répondant à des besoins communs.

Ce socle Open Source permettrait également à l'Europe d'affirmer ses valeurs, de faire respecter ses régulations et d'imposer ses standards, afin de contrôler et d'orienter les développements futurs de l'IA générative et plus largement de l'IA. Cette approche renforcerait l'autonomie technologique de l'Europe face aux géants numériques étrangers, tout en permettant à ses entreprises de rester compétitives à l'échelle mondiale. Une vision soutenue par le Président de la République française, qui, dans sa déclaration d'août 2024, appelait de ses vœux une "solution commune d'IA agréée et ouverte, basée sur l'Open Source pour ne pas être dépendants de solutions, et donc de standards, non européens".

Ce choix s'avère également pragmatique face aux moyens colossaux des modèles propriétaires. Le modèle Llama de Meta illustre parfaitement la pertinence de cette **stratégie de challenger, qui peut s'avérer gagnante**. En effet, **Llama 3.1 rivalise désormais avec GPT-4** en termes de performances techniques et se démarque par une adoption rapide parmi les LLM, surpassant **OpenAI** et **Microsoft** en termes de dépenses liées à l'IA, grâce à son approche Open Source. Cela souligne l'efficacité des stratégies Open Source pour concurrencer les modèles propriétaires tout en renforçant l'autonomie technologique, l'Europe pourrait s'en inspirer.

2.1.2. Modèles "Open Weights" vs modèles Open Source complets

Il est essentiel de **préciser ce que signifie "réellement Open Source"** dans le contexte de l'IA, car le terme est souvent mal utilisé. De nombreux acteurs, notamment les GAMAM, détournent ce concept en proposant des modèles dits "Open Weights", où seuls certains paramètres ou informations, comme la répartition des données par langue, sont publiés, mais pas les données d'entraînement elles-mêmes. L'Open Source Initiative (OSI[42]) a lancé un processus collaboratif et ouvert visant à élaborer une définition précise de l'Open Source AI (OSAID[43]). Cette définition, dont la publication est attendue d'ici la fin de l'année 2024, établira des critères stricts pour l'utilisation du terme "Open Source" dans le domaine de l'IA. Selon les règles déjà proposées par l'OSAID, des modèles comme Llama de Meta ne pourront plus se revendiquer comme étant Open Source. Bien que l'OSAID n'impose pas la publication des jeux de données d'entraînement, cette nouvelle définition apporte un premier niveau de clarté nécessaire dans un domaine où la transparence est cruciale.

[42] https://opensource.org/
[43] https://opensource.org/deepdive

Si l'Europe, et ses filières, souhaitent être réellement Open Source, cela inclut :

- **Une licence véritablement ouverte et sans restriction :** permettant une utilisation, modification et redistribution sans restriction dans le domaine commercial et sans limitation sur des secteurs et/ou des volumes de puissance mobilisés (exemple des modèles Llama de la société Meta).
- **Le code source du modèle :** les scripts et outils de pré-entraînement accompagnés des méthodologies d'évaluation et de test.
- **Les jeux de données d'entraînement, d'évaluation et d'alignement** disponibles sous licence libre permettant le réentraînement du modèle.
- **La gouvernance ouverte et collaborative :** un modèle de développement qui encourage la participation active de la communauté ouverte et distribuée.

OPENAI VS IA OPEN SOURCE

Malgré ce que son nom pourrait laisser penser, OpenAI s'est progressivement éloignée de son modèle initial axé sur la science ouverte, optant pour une approche fermée, propriétaire et commerciale. Sous la direction de Sam Altman, l'entreprise limite désormais l'accès à ses modèles les plus avancés, qui sont devenus propriétaires et opaques.

En revanche, des initiatives véritablement Open Source, telles qu'OpenLLM[44] France, font la promotion de la transparence, de la collaboration et de l'accessibilité tout en restant fidèles aux valeurs qu'elles défendent.

[44] https://www.openllm-france.fr/

2.1.3. Modèles de développement et de commercialisation : vers une approche européenne distinctive

Dans la quête de souveraineté numérique européenne en matière d'IA générative, il est essentiel d'examiner attentivement les modèles de développement et de commercialisation adoptés par les différents acteurs. Bien que des entreprises comme Mistral AI ou Aleph Alpha apparaissent comme des candidats sérieux pour représenter l'Europe, leur approche n'est pas toujours en phase avec les valeurs européennes et les logiques d'intérêt général pourtant souhaitées politiquement. Mistral AI constitue une alternative industrielle européenne crédible face à la domination américaine. Toutefois, son modèle, basé sur des levées de fonds massives et une ambition mondiale, reste proche de celui des grandes entreprises américaines, en proposant des solutions génériques destinées à un marché global plutôt qu'adaptées aux besoins spécifiques de l'Europe et de ses filières.

En parallèle, une approche Open Source, potentiellement construite également autour des modèles ouverts de Mistral AI, pourrait mieux répondre aux particularités des filières et des États membres européens, offrant une alternative plus cohérente avec les ambitions de l'Europe :

- **Indépendance technologique** : en ayant un accès complet au code, aux outils et aux données, les entreprises et institutions européennes peuvent maintenir un contrôle total sur leurs solutions d'IA, sans dépendance envers des acteurs extra européens.

- **Collaboration et innovation :** l'Open Source favorise la collaboration entre entreprises, universités et institutions publiques, créant un écosystème d'innovation dynamique au niveau européen.
- **Adaptabilité culturelle :** les modèles peuvent être plus facilement adaptés aux différentes langues et contextes culturels européens, préservant ainsi la diversité qui fait la richesse de l'Europe.
- **Transparence et confiance :** La transparence inhérente à l'approche Open Source renforce la confiance des utilisateurs et des régulateurs, un aspect crucial dans le contexte européen de protection des données et des droits individuels.

Pour les entreprises européennes, investir dans l'Open Source ne se limite pas à un choix éthique, c'est également une **stratégie pragmatique qui permet une mutualisation des coûts et donc une réduction des barrières à l'entrée "level the playing field"** :

- L'utilisation de logiciels Open Source élimine les coûts des droits d'usage imposés par les solutions d'IA propriétaires.
- Les entreprises peuvent partager les coûts de développement et les risques avec d'autres membres de la communauté Open Source, rendant ainsi les projets de grande envergure plus réalisables et moins risqués tout en favorisant la création de communs numériques.
- **Les LLM Open Source sont plus légers nécessitant des infrastructures d'inférence plus efficaces avec des durées de vie plus longues, voire permettant de réutiliser des parcs de GPU à recycler.**
- Les LLM pré-entraînés sur les langues européennes permettent aussi de diminuer les coûts d'inférence et plus particulièrement les coûts de l'opération de tokenisation qui correspond à la transformation prompts de l'utilisation en fraction de mots, elles-mêmes représentées par des nombres utilisés par le LLM. Les premiers retours d'expérience tendent à montrer que les modèles d'OpenAI (ChatGPT 4) ou Llama (entraînés majoritairement sur la langue anglaise) engendrent un surcoût de de l'ordre 30% lorsqu'ils sont utilisés en langue française par rapport au coût en langue anglaise pour une même demande[45].

2.1.4. Des travaux pionniers en Europe

Des écosystèmes Open existent déjà

Sur le chemin ouvert par le projet BigScience qui a donné entre autres naissance au LLM Bloom[46], de nouvelles initiatives comme celles menées dans en France par Kyutai[47] ou OpenLLM montrent qu'une autre voie, véritablement ouverte, est possible.

Ces projets ne se contentent pas de publier du code, mais créent des écosystèmes complets autour de leurs modèles et des données d'entraînement publiées sous licence Open Source, favorisant la réutilisation ainsi qu'une innovation collaborative et transparente.

A ce titre, OpenLLM France avec ses plus de 850 membres (octobre 2024) constitue l'espace francophone d'échange et de partage le plus large autour de l'IA générative.

Réplicabilité et scalabilité : les clés du succès européen

Pour accélérer l'adoption et le développement de solutions Open Source d'IA générative en Europe, plusieurs actions sont toutefois nécessaires :

[45] L'anglais, langue principale des données d'entraînement, optimise la conversion texte-token avec un ratio d'environ 0,75 mot par token. En français, ce ratio est de 0,57, nécessitant jusqu'à 30 % de tokens supplémentaires, augmentant ainsi les coûts (source : OpenAI Developer Forum)

[46] https://bigscience.huggingface.co/blog/bloom

[47] https://kyutai.org/

- **Soutien institutionnel renforcé** : des programmes de financement et des politiques favorisant explicitement les projets d'IA réellement Open Source.
- **Collaboration inter-sectorielle :** encourager les partenariats entre recherche académique, industrie et secteur public autour d'initiatives partagées et ouvertes.
- **Formation et sensibilisation :** développer les compétences sur l'IA Open Source dans les cursus d'IA et sensibiliser les décideurs à ses avantages.
- **Standards et certifications :** établir des normes européennes pour l'IA Open Source (Open Source Act), garantissant qualité et interopérabilité.
- **Plateforme d'échange :** créer une plateforme européenne pour le partage et la collaboration autour de projets d'IA Open Source afin de fédérer les initiatives déjà lancées par les États membres.

A RETENIR

Comme tout choix, celui de l'Open Source présente à la fois des avantages et des inconvénients. Il est donc crucial pour les entreprises d'évaluer soigneusement leurs besoins, leurs compétences internes et leur capacité à gérer ces technologies à long terme. Cela permet d'éviter les inconvénients potentiels, tels que la complexité accrue lors de l'intégration des LLM (nécessitant plus de compétences spécialisées) ou la dépendance vis-à-vis de la communauté pour la maintenance, la gestion des failles de sécurité et la réactivité face aux incidents.

En revanche un choix clair de l'Open Source permettra de mutualiser les coûts, par exemple au niveau d'une filière, d'augmenter la transparence, de limiter les dépendances et effets de lock-in, et de faciliter grandement la coopération entre les acteurs.

2.2 SPÉCIALISER SANS RÉENTRAINER ? (RAG VS FINE-TUNING)

2.2.1. La spécialisation des modèles : quelles sont les options ?

La coopération rendue possible par l'Open Source est une condition nécessaire, mais non suffisante, pour une voie européenne pérenne en matière d'IA générative. Pour rester compétitives, les filières doivent également être en mesure de spécialiser (de personnaliser/adapter), les technologies d'IA génératives, et donc les modèles/LLM, à des besoins très spécifiques, par exemple pour des filières telles que la santé, l'éducation, la mobilité ou l'industrie.

Mais **doit-on personnaliser les modèles eux-mêmes en les réentraînant, ou simplement personnaliser l'expérience utilisateur, sans toucher au modèle lui-même ?** La spécialisation peut se faire à travers plusieurs techniques. Les deux approches les plus répandues sont les suivantes :

- **Le fine-tuning :** Cette méthode consiste à ajuster un modèle générique préexistant en le réentraînant sur des données spécifiques à un domaine.

Par exemple, une banque française peut vouloir adapter un assistant virtuel aux particularités du droit bancaire français. En réentraînant le modèle sur un corpus de textes juridiques et de cas pratiques, l'assistant peut fournir des réponses précises sur des produits financiers complexes tout en respectant la réglementation locale européenne.

- **Le Retrieval-Augmented Generation (RAG) :** Cette méthode consiste à combiner la génération de texte via LLM avec la récupération d'informations externes provenant de bases de données ou de documents spécifiques, issus du Système d'Information d'une entreprise. Les données tierces vont "nourrir et enrichir" les prompts ("prompts engineering"), ou augmenter les réponses, afin d'affiner l'utilisation du modèle, sans pour autant nécessiter de réentraînement spécifique de ce même modèle.

De plus en plus utilisée dans les expérimentations récentes, notamment pour les chatbots d'assistance client, cette méthode permet à l'assistant IA d'accéder en temps réel aux données client, aux contrats et conditions spécifiques de l'entreprises, aux dernières mises à jour réglementaires, ainsi qu'à l'historique des interactions, pour des réponses personnalisées et pertinentes.

RAG & FINE TUNING

Vous souhaitez utiliser un modèle de langage générique comme GPT-4 (pré-entraîné sur une vaste quantité de données provenant d'internet) pour répondre à des questions spécifiques liées au service client d'une entreprise.

APPROCHE FINE-TUNING :

Collecte de données spécifiques : Vous rassemblez un ensemble d'exemples de conversations réelles avec des clients, couvrant des questions fréquentes telles que "Où en est ma commande ?", "Comment retourner un produit ?", ou "Quelle est votre politique de remboursement ?".

Réentraînement du modèle : Vous utilisez ces données pour ajuster le modèle ("fine-tuning"). Cela consiste à réentraîner le modèle pour qu'il comprenne et réponde avec précision aux questions types du service client, tout en adoptant le ton et le style de la marque.

Résultat : Après ce processus, le modèle est capable de fournir des réponses nettement plus pertinentes lorsqu'il est déployé dans un chatbot ou un système de réponse automatique pour le service client. Il peut, par exemple, détecter les nuances fines dans les demandes de remboursement ou de retour, et répondre en fonction des politiques précises de l'entreprise.

APPROCHE RAG :

En optant pour une approche RAG plutôt que pour le fine-tuning, le modèle conserve sa nature générique, tandis que les prompts utilisés pour répondre aux clients sont enrichis en temps réel par des informations et documents récupérés dans les bases de données internes de l'entreprise.

Cette approche permet d'assurer que les réponses s'appuient sur les données les plus récentes et les plus fiables, sans nécessiter un réentraînement spécifique (et coûteux) pour ce domaine particulier.

À RETENIR :

Le **fine-tuning** permet d'adapter un modèle pour qu'il devienne expert dans un domaine spécifique, comme la fonction service client d'une entreprise, en le réentraînant sur des données de l'entreprise.

Le **RAG** permet de fournir des réponses adaptées au contexte spécifique d'un domaine ou d'une entreprise, sans réentraînement, mais en réutilisant simplement les données de l'entreprise en temps réel pour augmenter la génération des prompts.

2.2.2. Préférer le RAG au fine-tuning

Nos différentes auditions avec des grandes entreprises de différents secteurs nous conduisent à préférer en général l'approche RAG par rapport au fine-tuning, non seulement pour des questions de coût, mais aussi de performance et de confiance.

Pour les entreprises, l'approche RAG offre des avantages significatifs par rapport au fine-tuning. Avec RAG, il n'est pas nécessaire de transmettre au modèle d'IA l'ensemble des données sensibles liées à la propriété intellectuelle, aux connaissances métier, à la confidentialité ou à la vie privée des clients. En effet, cette approche permet au LLM de se concentrer sur sa capacité à interagir en langage naturel avec les utilisateurs, ce qui est déjà extrêmement précieux, sans avoir à assimiler les informations internes concernant le fonctionnement de l'entreprise ou les données personnelles de ses clients. Concrètement, les avantages sont les suivants :

- **Coût :** Le RAG est souvent moins coûteux que le fine-tuning, car il ne nécessite pas de réentraînement complet du modèle. Il repose sur un modèle pré-entraîné générique mais capable d'accéder à des bases de données (structurées et non structurées) au sein des Systèmes d'Information existants des entreprises, pour extraire les informations nécessaires et aider à générer des prompts.

- **Flexibilité et temps réel :** Le RAG permet aux entreprises d'exploiter des données en temps réel pour améliorer les réponses générées par le modèle. Par exemple, un modèle RAG peut accéder à une base de données constamment mise à jour, garantissant que les réponses sont toujours basées sur les informations les plus récentes, un atout crucial pour les secteurs où les données évoluent très rapidement. À ce stade, cette actualisation en temps réel n'est pas possible avec le fine-tuning.

- **Réduction des biais :** En s'appuyant sur les données des Systèmes d'Information des entreprises, l'approche RAG aide à réduire les biais des modèles pré-entraînés. Avec cette approche, les entreprises peuvent choisir et contrôler les sources de données pour générer des réponses. Ces sources, provenant de bases de données fiables et vérifiées, diminuent ainsi le risque de biais provenant de sources externes non contrôlées ou de mauvaise qualité.

- **Fiabilité :** Le RAG permet de limiter les hallucinations des LLM en contraignant les réponses à un ensemble de données spécifiques provenant du Système d'Information de l'entreprise, plutôt que de s'appuyer sur des sources externes non contrôlées ou de mauvaise qualité.

- **Conformité réglementaire :** Le RAG aide les entreprises à se conformer aux réglementations (comme l'IA Act le RGPD, etc.) en gardant les données sensibles (données personnelles, données protégées par la propriété intellectuelle, le secret professionnel, etc.) dans les Systèmes d'Information internes, plutôt que de les exposer directement dans un modèle dont l'usage est difficile à contrôler.

- **Maintenance :** Le RAG simplifie la gestion et la mise à jour des données, car celles-ci peuvent être actualisées indépendamment du modèle. Cela facilite l'évolutivité, éliminant le besoin de réentraînement du modèle à chaque mise à jour. De plus, il permet l'ajout de nouvelles sources de données, y compris en temps réel, ou l'extension à d'autres domaines d'application sans modifier le modèle existant.

Par ailleurs, nos entretiens ont révélé que les grands modèles tels que Mistral AI, Llama ou GPT, sont déjà très performants lorsqu'ils sont utilisés dans des contextes métiers très spécifiques, ils n'ont pas besoin de réentraînement pour toutes les tâches demandées.

Le réentraînement et le *fine-tuning* font déjà partie intégrante des méthodes de spécialisation des modèles et seront inévitablement adoptés par certains acteurs. Dans ce rapport, nous encourageons simplement ceux qui envisagent cette approche à bien évaluer le rapport coût/bénéfice du réentraînement, tout en prenant en compte les risques associés aux données sensibles (données personnelles et données liées à la propriété intellectuelle). Par ailleurs, de nouvelles approches et architectures combinent déjà les avantages du RAG et du fine-tuning, comme le RAFT[48] (Retrieval Augmented Fine-Tuning), qui surpasserait le RAG à modèle équivalent. Nous anticipons que les mois et années à venir verront émerger de nombreuses nouvelles approches et méthodes pour spécialiser les modèles, mais la question clé restera de savoir si les données sensibles et les processus métier doivent être intégrés directement dans le modèle, ou traités à part.

A RETENIR :

MISER SUR LE RAG

Pour les entreprises, le RAG permet d'exploiter pleinement les capacités des LLM tout en offrant des avantages significatifs par rapport au fine-tuning, notamment en termes de confiance, de conformité réglementaire, de réduction des coûts, de performance, de flexibilité, et de maintenance.

Le RAG permet également de compartimenter des données confidentielles et sensibles d'entreprises, comme les données personnelles des clients, qui sont traitées dans une "couche data", et non pas dans une "couche IA", plus opaque et plus difficile à maîtriser.

LES LLM GÉNÉRALISTES "PARLENT DÉJÀ TOURISME"

Une entreprise internationale du secteur du voyage que nous avons interviewée, a constaté après un an de développement de son « Assistant spécialisé » basé sur des LLM, dont plusieurs mois en production, que la majorité des modèles mis en place fonctionnent parfaitement sans nécessiter de *fine-tuning* spécifique, même pour des tâches métier très spécialisées comme la réservation. Selon cette entreprise, la seule différence notable concerne parfois la gestion de la toxicité, mais cela est bien maîtrisé grâce au prompt engineering. Les LLM généralistes actuels «parlent déjà correctement le langage spécifique du tourisme», rendant inutile leur réentraînement.

Plus surprenant encore, les entreprises interrogées qui ont utilisé le fine-tuning pour des tâches spécialisées, telles que les constructions tarifaires ou les règles liées aux programmes de fidélité, ont constaté que cette méthode peut parfois être contre-productive. En effet, **le *fine-tuning* semble générer davantage d'hallucinations par rapport à l'approche RAG, qui, en s'appuyant sur des données internes actualisées, réduit ces erreurs et donc augmente la fiabilité de l'expérience proposée au client final.**

[48] https://techcommunity.microsoft.com/t5/ai-ai-platform-blog/raft-a-new-way-to-teach-llms-to-be-better-at-rag/ba-p/4084674

DEVENONS LEADERS DES LAM PLUTÔT QUE SIMPLES CONSOMMATEURS DES LLM

2.3. LE PARI DES LAM ET DE L'AGENTIFICATION

2.3.1. Quels sont les paris possibles ?

Pour devenir un leader dans le domaine de l'Intelligence artificielle, l'Europe ne peut pas se contenter de simplement rattraper son retard en matière d'IA générative et de LLM. Cela implique de prendre des positions audacieuses et d'adopter les recommandations du rapport de Mario Draghi[49], qui souligne l'importance d'assumer des parti-pris stratégiques dans les paris technologiques. **Sans cette prise de risques, sans une forme de singularité dans nos choix, l'Europe ne pourra pas rivaliser avec les puissances comme les États-Unis et la Chine,** qui n'hésitent pas à investir massivement dans l'IA.

Au-delà des approches actuelles de spécialisation des modèles comme le fine-tuning ou le RAG, l'IA générative européenne doit anticiper les tendances futures et proposer des innovations de rupture :

- **Plus petit (des LLM au SML) :** face à l'impasse du gigantisme et du manque d'adaptation à des problèmes spécifiques des LLM, le marché produit déjà des SLM (Small Language Models). Les LLM sont des modèles de grande taille, souvent composés de milliards de paramètres, qui nécessitent des ressources importantes en termes de calcul, de mémoire et de stockage. Les SLM en revanche sont des modèles de plus petite taille, avec un nombre réduit de paramètres, généralement dans l'ordre de centaines de millions à quelques milliards de paramètres, par exemple phi-2 de Microsoft, ou encore Llama3.2 1B et 3B de Meta. Ces modèles sont plus légers et consomment moins de ressources, ce qui les rend plus faciles à déployer sur des appareils ou infrastructures aux ressources limitées.

- **Plus d'action (des LLM au LAM) :** là où les LLM sont déjà passés maîtres dans l'art de la génération de contenus multiples à partir de langage naturel (text2text, text2image, voice2text, etc.), les LAM (Large Action Models) en revanche, sont axés sur la prise de décision et l'exécution d'actions dans des environnements physiques ou simulés à partir de prompts en langage naturel (text2action, voice2action), impliquant souvent des tâches de contrôle ou d'action automatisée.

« Small is beautiful », notamment en termes d'efficacité et d'adaptation à des besoins spécifiques, et les SML sont surtout moins gourmand en ressources. Cependant, nos nombreux échanges avec les acteurs métiers de différentes filières, qui cherchent à exploiter tout le potentiel des LLM, nous poussent à penser que les Large Action Models (LAM), avec leur nouvelle capacité d'interagir avec le monde réel (via les Systèmes d'Information de nos entreprises), représentent un axe de développement encore plus prometteur, et aux implications plus générales.

2.3.2. Passer des LLM au LAM

Les LAM (Large Action Models) représentent une évolution naturelle pour l'IA générative[50], qui passe de la simple génération de contenu à partir de langage naturel, à la capacité d'interagir directement avec les Systèmes d'Information des entreprises. Le LAM permet l'exécution automatisée d'actions complexes, initiées par des commandes en langage naturel des utilisateurs finaux.

[49] https://commission.europa.eu/document/97e481fd-2dc3-412d-be4c-f152a8232961_en

[50] https://medium.com/version-1/the-rise-of-large-action-models-lams-how-ai-can-understand-and-execute-human-intentions-f59c8e78bc09

L'adoption et le développement des LAM représentent un atout stratégique pour l'Europe. En investissant dans cette technologie émergente, l'Europe peut se positionner à la pointe de l'innovation en IA, en offrant des solutions plus sophistiquées, adaptées aux besoins complexes des entreprises, des administrations et des individus.

Avec les LAM, la prochaine génération d'assistants IA pourra convertir les demandes des utilisateurs ou clients en actions concrètes, multiples et complexes, telles que l'organisation, la réservation, la planification, la diffusion d'information ou encore la réalisation de tâches administratives. Avec les LAM, les assistants IA peuvent devenir de puissants outils de productivité.

Comment fonctionnent les LAM ?

- **Le LAM décompose la demande en langage naturel de l'utilisateur en une séquence plus ou moins complexe de sous-tâches**, qu'il exécute ensuite via des programmes dédiés capables d'accéder aux systèmes d'information des entreprises (par exemple, un système de réservation dans le secteur des transports).
- **Les programmes permettant d'accéder aux Systèmes d'Information des entreprises et d'interagir avec eux sont appelés des agents.** Un LAM peut activer une série d'agents proposés par une seule entreprise, ou par un ensemble d'entreprises.
- La séquence d'actions est générée automatiquement par le LAM, qui fait preuve d'agilité en proposant une nouvelle action ou sous-tâche en cas d'erreur ou d'imprévu.
- Tout comme un LLM, le LAM peut s'améliorer en apprenant sur un grand nombre d'exemples et d'interactions avec des utilisateurs, devenant ainsi plus performant au fil du temps.
- L'utilisateur final initie l'action, en garde le contrôle, mais la supervision reste minimale : **c'est la machine qui exécute l'action dans son ensemble, et les sous-tâches associées, au nom de l'utilisateur et à sa place (principe de "délégation d'action").**

EXEMPLE D'USAGE DE LARGE ACTION MODELS (LAM)

Les Large Action Models (LAM) sont une avancée en Intelligence Artificielle conçue pour transformer des commandes textuelles ou vocales en actions réelles dans des environnements numériques. Ils vont au-delà de la compréhension du langage pour exécuter des tâches complexes dans diverses applications.

Avec un LAM, il est possible d'énoncer un désir de manière complexe et intégrée, comme "Organise une soirée cinéma avec un bon film, commande de la pizza pour 20h et prépare l'éclairage du salon." Le LAM analyserait cette demande complexe, comprendrait les différentes actions requises et les exécuterait de manière séquentielle ou simultanée à travers différentes applications et dispositifs. Le modèle pourrait choisir un film basé sur vos préférences passées, passer commande sur le site web de votre pizzeria préférée en utilisant vos informations de paiement sécurisées, et ajuster l'éclairage via une interface intelligente domestique, le tout en une seule opération fluide.

EXEMPLES DE LAM

Le cas le plus connu est Microsoft Copilot, directement intégré dans les applications Microsoft 365, qui aide les utilisateurs à accomplir des tâches complexes comme générer des rapports, planifier des réunions, ou automatiser des processus professionnels.

A noter aussi la démonstration prometteuse de Rabbit R1 qui exécute des commandes complexes comme naviguer sur internet pour réserver un vol ou acheter des produits en ligne en utilisant des instructions verbales simples de l'utilisateur.

Les communautés Open Source, comme crewai[51] , proposent des boîtes à outils déjà très sophistiquées permettant de mettre en œuvre des processus complexes de raisonnement à base d'agents.

[51] https://www.crewai.com/

CRÉER UN LAM DU VOYAGE POUR TOUS LES TRAVELS ASSISTANTS

Le projet Gen4Travel a déjà identifié plusieurs fonctionnalités de Travel Assistant qui devront s'appuyer sur un LAM, comme par exemple la réservation (train, avion, hôtel, activités touristiques, etc.).

La **délégation d'action** offerte par les LAM permettra, par exemple, à un voyageur de **demander verbalement à son assistant de réserver un voyage vers une destination précise pour le samedi suivant.** En prenant en compte les informations personnelles du voyageur, l'assistant pourra analyser les différentes options disponibles (ex : trajets multimodaux) et les confronter aux contraintes spécifiques du voyageur (agenda, préférences budgétaires, statut Personne à Mobilité Réduite PMR, cartes de fidélité, etc.). L'assistant sera alors capable d'effectuer les diverses réservations correspondantes, avec l'accord du voyageur, mais sans nécessiter de supervision détaillée de sa part.

2.3.3. L'Agentic AI comme allié naturel des LAM

Les actions (tâches et sous-tâches) déclenchées par les LAM sont exécutées par des programmes appelés "agents". **L'approche Agentic AI** est une évolution récente dans l'industrie, reposant sur une **architecture où plusieurs agents autonomes interagissent en réseau, augmentant ainsi la flexibilité du système et sa capacité à répondre à des besoins encore plus complexes.**

Chaque agent autonome est spécialisé dans une tâche ou fonction spécifique. Cette architecture peut être organisée de deux manières :

- **Centralisée :** Un orchestrateur central reçoit la demande initiale de l'utilisateur final (en langage naturel), puis active et supervise la constellation d'agents. Si un agent fournit une information inattendue ou imprévue, l'orchestrateur adapte les actions en conséquence, ajustant dynamiquement les processus pour mieux répondre aux besoins évolutifs.
- **Décentralisée :** Sans orchestrateur, les agents collaborent directement entre eux, s'invoquant mutuellement en fonction des exigences de l'utilisateur. L'Agentic AI facilite des transferts de contrôle (handoffs) entre agents, permettant l'exécution fluide de processus complexes tout en conservant une transparence élevée et un contrôle granulaire sur l'utilisation des outils et des contextes.

Le LAM intervient pour soutenir ce réseau d'agents de manière directe. Dans un système centralisé, il aide l'orchestrateur à déterminer la séquence optimale d'agents à activer. Dans un environnement décentralisé, le LAM peut aussi assister les agents eux-mêmes, améliorant ainsi leur coordination et leur réactivité.

En résumé, l'approche Agentic AI, associée aux LAM, permet de créer des systèmes plus autonomes, adaptatifs et robustes, capables de gérer des interactions complexes tout en maintenant un haut niveau de contrôle et de flexibilité.

Bien que peu mature pour le moment, l'approche Agentic AI est déjà au cœur des stratégies des acteurs majeurs de l'IA générative. Open AI a par exemple lancé en 2024 un outil Open Source appelé Swarm[52] (essaim en français). Swarm est un framework multi-agents permettant de gérer une constellation d'agents autonomes associés au LLM, et de les coordonner. A noter que des projets Open Source tels que CrewAI[53] proposent des approches équivalentes et tout aussi prometteuses.

[52] *OpenAI Introduces Swarm : a Framework for Building Multi-Agent Systems*

[53] *https://www.crewai.com*

OFFRIR DE L'AUTONOMIE DE DÉCISION AUX ASSISTANTS

En intégrant l'approche **Agentique AI** (avec des agents multiples réalisant diverses fonctions, et un orchestrateur sélectionnant les agents appropriés selon la situation), **les assistants Gen4Travel seront plus agiles et capables de s'adapter à des contextes dynamiques et complexes.**

Un exemple pertinent est la gestion d'une perturbation lors d'un voyage déjà réservé. **Si le vol du voyageur est annulé, l'assistant sera suffisamment autonome pour reprogrammer un autre vol,** réserver un hôtel pour la nuit, **annuler la location de voiture** à destination, et ajuster d'autres éléments de l'itinéraire. **Il pourra coordonner les actions et négocier avec les différents prestataires,** n'exigeant de l'utilisateur que les validations intermédiaires ou finales, sans nécessiter d'intervention directe tout au long du processus. L'assistant pourra réadapter la réponse en fonction des retours des différentes parties-prenantes.

La constellation des agents mis à contribution contiendra des agents multiples : listing des offres avion/train/hôtel, outils de réservation de trajets/chambres d'hôtel, analyse des conditions tarifaires (et de gestion des disruption) de chaque acteur, récupération du contexte personnel de l'utilisateur (ses données d'identité, ses préférences, ses historiques d'achat, ses programmes fidélité, etc.), matching entre le contexte utilisateur et les offres disponibles, etc.

Conformément à l'approche Agentic AI, chaque agent pourra être mis à disposition de Gen4Travel par chacune des entreprises concernées (compagnie aérienne/ferroviaire, hôtelier, loueur de voiture, entreprise de taxi/VTC, etc.), et disposera d'un niveau d'autonomie avancé. Avec l'Agentic AI, ce n'est pas simplement l'assistant IA qui répond au besoin du voyageur, mais c'est toute la filière qui se coordonne pour lui proposer une solution sur mesure à un problème complexe.

Cet exemple illustre aussi clairement l'avantage stratégique de l'exploitation des données des entreprises : **la filière est capable de proposer des services d'une valeur exceptionnelle, que des acteurs comme OpenAI ne peuvent offrir à ce stade,** sauf si bien sûr que les entreprises de la filière du voyage choisissent délibérément d'ouvrir à OpenAI les accès à leurs Systèmes d'Information, risquant ainsi de créer une dépendance nouvelle dangereuse vis à vis des géants du numérique.

2.3.4. Un LAM de filière ?

Nous suggérons donc que les LAM, et l'architecture Agentic AI, soient développés au niveau d'une filière ou d'un consortium étendu.

Un développement coopératif de LAM par filière permettrait de réduire les coûts, mais également **de développer de nouveaux services d'assistants capables d'aider les utilisateurs finaux de manière très transverse**, pour l'ensemble des tâches de leur quotidien.

La coopération dans les LAM permettrait également aux acteurs européens de **répondre à des besoins collectifs et sociétaux majeurs, comme ceux liés à l'environnement**, à la santé, à l'efficacité énergétique, à la mobilité, au tourisme et à l'emploi.

UN LAM UNIQUE POUR LA FILIÈRE DU VOYAGE ?

L'ambition de **Gen4Travel** est de permettre aux acteurs du tourisme de créer plusieurs assistants, évitant ainsi un monopole.

Cependant, pour offrir une expérience utilisateur optimale, **chaque assistant doit être intégré à un écosystème ouvert et riche** (si chaque assistant ne peut réserver des services que pour une seule entreprise, l'expérience utilisateur sera décevante).

En puisant dans un vaste réseau de partenaires et de fournisseurs cela permet d'accéder à une large gamme de services et de personnaliser les solutions pour chaque voyageur.

L'objectif est que tout assistant puisse permettre de réserver/organiser/réorganiser un voyage complet, incluant divers services (transports, hébergement, activités). Pour cela, **il est crucial de développer un LAM unique pour toute la filière, intégré à une infrastructure digitale mutualisée, permettant aux assistants d'effectuer des actions multiples et interconnectées.**

DEVENONS LEADERS DES LAM PLUTÔT QUE SIMPLES CONSOMMATEURS DES LLM.

En misant sur des technologies fiables et éthiques, et en passant du logos au praxis avec les Large Action Models, **l'Europe peut se positionner comme un leader mondial de l'IA générative au service de ses entreprises,** tout en garantissant que cette révolution technologique bénéficie à l'ensemble de ses entreprises et citoyens.

PRIVATE DATA
IS THE NEW OIL

2.4. MUTUALISER ET PARTAGER LES DONNÉES DE NOS ENTREPRISES

2.4.1 Passer de l'Open Data au Shared Data

Sortir de la naïveté du tout "Open Data"

L'ère de l'Open Data a favorisé la transparence et l'innovation grâce à l'ouverture des données publiques. Toutefois, il est temps de passer à la phase du "Shared Data", qui encourage le partage stratégique, et contrôlé, des données, tant des institutions publiques que des entreprises privées. Cette approche reconnaît la valeur collective des données, tout en respectant la confidentialité et la propriété intellectuelle. En créant des écosystèmes de partage sécurisé et contrôlés, l'Europe peut stimuler l'innovation tout en protégeant les intérêts de tous, se posant ainsi comme une alternative face au Big Data américain et chinois.

L'enthousiasme pour l'Open Data a parfois été naïf, pensant que libérer les données publiques suffirait à innover. En réalité, toutes les données n'ont pas la même valeur, et un partage sans discernement, et sans business model, peut poser des problèmes de confidentialité et de sécurité, ou manquer d'adéquation avec les besoins réels du marché.

Le "Share Data" en revanche, où les détenteurs de données maîtrisent et peuvent même monétiser les données, permet un partage structuré, laissant le contrôle à ceux qui partagent, maximisant l'impact tout en réduisant les risques.

"Shared data" la réponse européenne au "Big data"

L'émergence du Big Data dans les années 2000 et 2010, combinée à l'expansion d'internet et de la connectivité numérique (3G+, fibre), a généré une immense quantité de données (textes, images, sons), indispensables pour entraîner des modèles d'IA avancés. Les informations issues d'Internet, des réseaux sociaux et des smartphones ont offert un aperçu inédit du comportement humain, rendant l'IA plus pertinente et applicable à diverses situations réelles.

Cependant, l'Europe a déjà perdu la course aux données : seuls les géants technologiques américains et chinois ont su tirer parti des effets de réseau (Loi de Metcalfe[54]). Si la bataille pour connecter les 8 milliards d'humains est derrière nous, la connexion des entreprises est un défi que nous devons relever, car nous avons les atouts nécessaires. Pour cela, **nous devons créer un marché commun des données, atteignant une taille critique pour rivaliser avec le Big Data des Big Techs.** Il ne s'agit pas d'un marché ouvert au pillage, mais d'un espace ou le partage des données est encadré et contrôlé, où le partage se fait entre pairs qui se font confiance, selon des règles et des valeurs qui nous sont propres.

2.4.2 Données privées des entreprises : notre trésor commun

Le partage des données privées, clé du succès pour l'IA générative

Pour s'imposer comme leader dans l'ère de l'IA générative, et plus encore de l'agentification, l'Europe doit adopter une stratégie audacieuse et innovante qui combine :

- **Une coopération pan-européenne basée sur l'Open Source,** favorisant l'innovation collective et la transparence.
- **Une spécialisation des LLM avec pas (ou peu) de réentraînement.**

[54] La loi de Metcalfe est un principe qui affirme que la valeur d'un réseau de communication (comme un réseau social ou un réseau informatique) est proportionnelle au carré du nombre d'utilisateurs connectés à ce réseau. Autrement dit, plus il y a de participants dans un réseau, plus la valeur ou l'utilité du réseau croît de manière exponentielle.

- **Les technologies Language Action Models (LAM) associés à des réseaux multi-agents (Agentic AI),** facilitant l'interaction fluide entre les agents IA et les Systèmes d'Information des entreprises, ouvrant la voie à une automatisation intelligente et une prise de décision autonome optimisée.

Ces trois axes suffisent-ils pour se démarquer des géants actuels ? Malheureusement non :

- L'Open Source offre transparence et coopération, mais les modèles propriétaires et semi-propriétaires (open weights) dominent souvent en termes de performance pure et d'adoption (Meta).
- La spécialisation est déjà possible et efficace avec les technologies existantes (le RAG est déjà l'approche privilégiée par de nombreuses entreprises en Europe comme aux États-Unis).
- Le LAM et l'Agentic AI sont déjà au cœur des stratégies de plusieurs acteurs majeurs du marché, y compris parmi les géants américains (Open AI, Microsoft, etc.).

Il est donc crucial pour les filières d'aller au-delà de ces trois approches pour réellement se distinguer et prendre l'avantage.

Mais alors, que reste-t-il ?

Ce qu'il reste, ce sont nos données : les données privées, celles de nos entreprises, de nos administrations, de nos citoyens et de nos filières !

Face aux ressources financières colossales des géants américains et à leurs investissements massifs, il est raisonnable de penser qu'ils ont déjà accès à toutes les données publiquement disponibles (web, Open Data, etc.) ou qu'ils y accéderont tôt ou tard. C'est donc dans les données privées que réside le véritable levier de souveraineté, celui qui nous permettra de nous démarquer et garder le contrôle.

Si l'Europe souhaite se différencier, ce sera en exploitant mieux, tout en les protégeant, les données privées (personnelles et non personnelles) issues de ses entreprises, de ses administrations et de ses filières.

En mutualisant, en partageant de manière contrôlée, les données privées au niveau sectoriel et intersectoriel, les acteurs européens peuvent créer un écosystème de données souverain, riche et varié, amplifiant ainsi la puissance et la pertinence des systèmes d'IA générative.

En somme, **notre trésor collectif de données privées est bien plus qu'une simple ressource à exploiter individuellement ; c'est le fondement sur lequel nous pouvons bâtir, ensemble, une nouvelle forme d'IA générative. En cultivant ce continuum - des espaces de données partagés (Data Spaces) aux RAG puis aux LAM - nous définissons une nouvelle voie vers une IA générative de confiance, performante et éthiquement responsable.** Cette approche collaborative, basée sur le partage à grande échelle, permet non seulement d'améliorer significativement les performances des modèles, mais aussi de conserver des leviers de souveraineté collectifs, renforçant ainsi la position de l'Europe dans le paysage mondial de l'IA.

Des assistants augmentés grâce aux données privées partagées

Le passage au Shared Data ouvre également la voie à la création d'assistants augmentés, capables de tirer parti des données privées (issues des Systèmes d'Information de nos entreprises et administrations) partagées pour offrir des services plus sophistiqués et personnalisés. Ces assistants, qu'ils soient utilisés dans le domaine de la santé, de l'éducation, ou de l'industrie, peuvent analyser des volumes massifs de données pour fournir des conseils précis, anticiper les besoins et optimiser les processus.

Grâce au Shared Data, les assistants peuvent surtout agir et commander des actions sur les Systèmes d'Information de nos entreprises (réserver, planifier, organiser, etc.). Nous pouvons développer des assistants qui ne se contentent pas de répondre aux questions des utilisateurs, mais qui peuvent également proposer des solutions proactives, automatisées et sur mesure.

A RETENIR

"If we don't share our data, others will centralize it"

Soit nous partageons les données entre nous, soit elles sont données aux Big AI.

"Private data is the new oil"

Les Big Techs, malgré leurs moyens technologiques et financiers sans pareil, n'ont pas accès aux données ayant le plus de valeur : celles de nos entreprises[55]. Faisons de ces données privées un avantage compétitif pour nos filières.

"Shared data is the new Open data"

En abandonnant la naïveté du tout "Open Data" et en reconnaissant la valeur des données privées des entreprises comme un trésor commun, nous pouvons créer un écosystème où les données sont partagées de manière stratégique et sécurisée. Cette approche permet non seulement de protéger les intérêts des parties prenantes, mais aussi de stimuler l'innovation et la compétitivité à travers des assistants augmentés et d'autres applications avancées.

[55] Cette affirmation met de côté la question du Cloud Act, et part du principe que les hypersaclers américains hébergeant les données d'entreprises européennes respectent bien notre régulation. CQFD.

DONNÉES PRIVÉES PARTAGÉES ET TRAVEL ASSISTANTS

Le Shared Data permettra aux assistants Gen4Travel d'exploiter, de manière contrôlée, de nombreuses données privées :

Données personnelles (requérant consentement RGPD explicite de l'utilisateur) :

- Informations d'identification (régaliennes et non régaliennes)
- Préférences de voyage
- Informations de paiement
- Historique de voyage
- Données de santé (ex : statut PMR, allergies, etc.)
- Cartes de fidélité et programmes de récompenses
- Préférences personnelles
- Agenda et calendriers
- Données biométriques
- Données de géolocalisation

Données non personnelles (issues des Systèmes d'Information des entreprises) :

- Horaires et disponibilités des services
- Données des offres de services et événements avec conditions tarifaires
- Offres promotionnelles et réductions
- Informations sur les destination
- Politiques d'entrée et de sortie
- Avis et notes des services
- Notifications en temps réel (perturbations et disruptions)

SOIT NOUS PARTAGEONS LES DONNÉES ENTRE NOUS, SOIT ELLES SONT DONNÉES AUX BIG AI

2.5. MISER SUR LES DATA SPACES

2.5.1. Les Data Spaces au coeur de la stratégie des données de l'UE

Le partage des données privées, ou "Shared Data", est une question complexe, nécessitant la coopération entre entreprises et administrations européennes, et la création d'écosystèmes basés sur des standards d'interopérabilité et de confiance communs.

Le Shared Data constitue précisément l'objectif de la stratégie européenne de données[56] lancée en 2020. Cette stratégie vise à créer un marché unique des données avec des règles d'accès équitables et transparentes, tout en respectant notamment la confidentialité et la protection des données, ainsi que les lois sur la concurrence.

Le cadre réglementaire, comprenant principalement la loi sur la gouvernance des données (Data Governance Act - DGA) et la loi sur les données (Data Act - DA), introduit le concept d'« espaces communs de données » ou Data Spaces.

La Commission européenne investit près de 10 milliards d'euros (2022-2026) à travers le programme Digital Europe pour développer ces Data Spaces et assurer leur coordination à l'échelle du continent. Ce projet constitue un effort de standardisation inédit dans une quinzaine de secteurs, tels que la santé, la finance, l'agriculture, l'industrie, l'énergie, la culture, le Green Deal, la mobilité, l'administration publique, les compétences, et bien d'autres. Bien que le partage de données et la coordination des acteurs ne soient pas des concepts nouveaux, l'ampleur et l'ambition de ce projet sont, quant à elles, historiques.

DATA SPACE INVESTMENTS & SUBSIDIES

Source	Programme	€	Comment
Germany	National Funding	421 M	Data Ecosystems: Gaia-X Funding Competition (11 Projects), Manufacturing-X, Catena-X, Gaia-X 4 Future Mobility, EuProGigant, Energy data-X.
Spain	National Funding	150 M	Industrial Data Spaces Open Call
	National Funding	50 M	Tourism Data Spaces Open Call
France	National Funding	110 M	40 M Data4industry-X, 70 M€ for new call for tender
Luxembourg	National Funding	20 M	
Denmark	National Funding	4,8 M	
Finland	Sitra	2,6 M	Sitra invested 2,6 mil. of which EUR 625,000 was used to co-finance 5 pilot projects related to data spaces. The co-financing rate covered by Sitra per project was 70%, the rest 30% was covered by project consortia members.
EU	Digital Europe Work Programme 2021-2022	206 M	For topics deploying the sectorial data spaces and the related support activities, including the High Value Data Sets. This set of calls includes the DSSC (14M).
EU	Digital Europe Work Programme 2023-2024	151 M	For topics deploying the sectorial data spaces and the related support activities including actions on Digital Product Passport.
EU	EU4Health	280 M	Implementation of European Health Data Space
EU	Horizon Europe	100 M	Energy data spaces and R&I projects
EU	Digital Europe Work Programme 2021-2022	150 M	Destination Earth initiative
EU	Digital Europe Work Programme 2023-2024	90 M	Destination Earth initiative
TOTAL		**1,735.4**	

INVESTISSEMENT DÉJÀ RÉALISÉS DANS LES DATA SPACES (GAIA-X)

[56] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

QUE SONT LES DATA SPACES ?

Un Data Space est un écosystème d'organisations qui partagent volontairement des données (personnelles et non personnelles) de manière décentralisée, contrôlée, standardisée et régulée. Il repose sur une gouvernance commune qui établit, de façon collaborative, les règles de partage entre les participants. Une fois ces règles définies, une infrastructure technique mutualisée de partage de données est mise en place, respectant les principes de cette gouvernance.

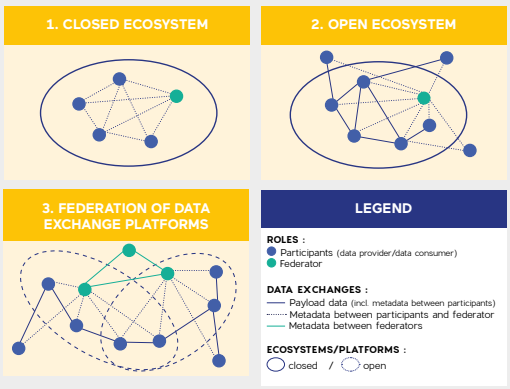
L'adoption de standards communs assure l'interopérabilité et renforce la confiance entre les participants d'un Data Space, et même entre plusieurs Data Spaces, générant ainsi un effet de réseau large et multi-sectoriel.

Chaque Data Space applique des standards et labels transsectoriels pour le partage de données (comme l'identité des organisations et des individus, ou les contrats de partage de données, les règles de partage et de gouvernance, etc.). Ces standards communs sont établis collectivement via des organisations telles que Gaia-X[57].

Dans un même secteur, les Data Spaces s'accordent sur des standards sémantiques spécifiques à leur domaine d'activité, suivant le principe de subsidiarité (par exemple, les Data Spaces liés à la mobilité définissent leurs propres standards sectoriels).

Les Data Spaces peuvent être créés de deux manières : par une approche descendante (top-down), initiée à l'échelle sectorielle et européenne (comme l'EMDS, l'Espace européen des données de mobilité), ou par une approche ascendante (bottom-up), où des consortiums d'acteurs locaux prennent l'initiative (par exemple, EONA-X, développé par des acteurs français et espagnols de la filière, ou le German Mobility Data Space, conçu par des acteurs allemands). Une approche d'urbanisation globale permet d'intégrer ces différents niveaux (par exemple, EONA-X et le German Mobility Data Space font tous deux partie de l'EMDS).

L'ensemble des Data Spaces européens forme ainsi un écosystème departage de données fédéré et ouvert, interopérable et de confiance.



DES ÉCOSYSTÈMES FERMÉS AUX DATA SPACES

Blueprint Common Energy Data Space[58]

[57] https://gaia-x.eu/

[58] https://enershare.eu/wp-content/uploads/Blueprint_CEEDS_v2.pdf

La stratégie européenne reposant sur les Data Spaces introduit une standardisation inédite du partage de données, visant à briser les silos entre les entreprises des 27 pays de l'UE et à favoriser une collaboration transnationale et intersectorielle. Les Data Spaces incarnent une nouvelle forme de souveraineté des données, offrant une alternative aux solutions cloud, données et IA dominées par des acteurs non européens. Conçus pour le marché européen, ils sont essentiels à la création d'un véritable espace numérique unifié, qui soutient l'écosystème des clouds européens tout en fournissant la matière première indispensable pour l'entraînement et l'exploitation des IA.

En mutualisant les données, les Data Spaces renforcent l'interopérabilité et la confiance, stimulant ainsi l'innovation à travers des collaborations intersectorielles. Ils permettent aux IA génératives d'accéder de manière sécurisée et contrôlée aux données des entreprises et administrations, sous la supervision des détenteurs de ces données, garantissant ainsi la souveraineté numérique de l'Europe et protégeant les acteurs locaux contre la désintermédiation.

Profiter enfin de l'effet de réseau grâce à un marché unique de la donnée

L'un des principaux atouts des Data Spaces est leur capacité à générer un effet de réseau puissant à travers un marché commun de la donnée. En connectant des sources de données multiples, les organisations peuvent accéder à une richesse d'informations auparavant inatteignable. Cela permet non seulement de diversifier les points de vue, mais aussi d'améliorer significativement la prise de décision, l'efficacité opérationnelle, l'expérience utilisateur et l'innovation produit. Les entreprises européennes peuvent ainsi collaborer plus étroitement, développer des solutions communes et s'ouvrir à des marchés plus vastes et diversifiés.

2.5.2 Les Data Spaces au service d'une IA capable et responsable

Les Data Spaces jouent un rôle essentiel dans la création d'écosystèmes interopérables et de confiance, fondamentaux pour le développement de l'Intelligence Artificielle. En assurant que les données partagées sont de haute qualité, contrôlées par leurs détenteurs, sécurisées, standardisées, et conformes aux réglementations de protection des données, ces écosystèmes permettent aux technologies d'IA de fonctionner de manière optimale et éthique. Il est en effet "important qu'un utilisateur de ces services ait la vision complète et transparente de l'ensemble de la chaîne".[59]

Cela **permet aux entreprises de développer des modèles et des assistants IA plus précis et fiables, basés sur des données vérifiées et provenant de sources multiples, renforçant ainsi la confiance des utilisateurs et partenaires.** Ce point est d'autant plus crucial que 60 % à 80 %[60] des projets d'IA échouent à atteindre leurs objectifs en raison d'un manque d'accès aux données.

Selon la thèse développée dans ce rapport, les Data Spaces fourniront aux assistants AI, via le RAG notamment, des données issues des Systèmes d'Information des entreprises, permettant une spécialisation avancée et des interactions sophistiquées. Ils offriront également une prise en compte globale et transverse du contexte de l'utilisateur des Assistants IA, avec des données provenant de diverses entreprises, tout en préservant un contrôle fort pour les détenteurs et producteurs de ces données.

[59] *"Valeur de Gaia-X dans le développement de l'économie européenne utilisant des solutions d'IAG"*, Hub France Gaia-X, mai 2024

[60] The One Practice That Is Separating The AI Successes From The Failures" Ron Schmelzer, Forbes, 2022

Des consortiums de filière pour partager et mutualiser

Pour maximiser les bénéfices des Data Spaces, la formation de consortiums de filière est essentielle. Ces groupements d'entreprises et d'organisations au sein d'une même industrie peuvent partager et mutualiser leurs données de manière structurée et sécurisée. En travaillant ensemble, ces consortiums peuvent identifier des tendances sectorielles, optimiser leurs chaînes de valeur et développer des innovations conjointes qui profitent à l'ensemble de la filière. Cette collaboration inter-entreprises permet de surmonter la problématique des silos de données et d'accélérer la transformation numérique de secteurs entiers.

Un avantage compétitif sur les enjeux d'harmonisation des données

Contrairement aux États-Unis, où l'on observe l'émergence d'un besoin pour des couches d'harmonisation des données nécessaires au développement de l'IA générative ("harmonized data layers"[61]), l'Europe a déjà pris une longueur d'avance grâce à ses initiatives de Data Spaces. Les Data Spaces représentent un élément clé dans le développement d'applications basées sur des agents collaboratifs multiples qui nécessitent une couche d'harmonisation pour fonctionner correctement. En Europe, cette harmonisation est déjà largement facilitée par les Data Spaces, permettant ainsi aux entreprises européennes de se concentrer davantage sur l'innovation (IA générative, LAM, Agentic AI, etc.) et la spécialisation, plutôt que sur la mise en place d'infrastructures de partage de base. Les agents IA orchestrés via ces Data Spaces, peuvent collaborer de manière intelligente en s'appuyant sur des LLM/LAM qui leur permettent de traiter les données dans un écosystème adaptatif et semi-autonome. Ces agents utilisent les données disponibles dans les Data Spaces pour synchroniser leurs actions avec les objectifs des entreprises et des utilisateurs d'assistants AI, renforçant ainsi la cohérence et l'efficacité des processus métier.

Des standards d'interopérabilité et de confiance

Le succès des Data Spaces repose sur l'établissement de standards d'interopérabilité et de confiance. Ces normes garantissent que les données peuvent être partagées facilement entre différents systèmes et organisations, tout en maintenant leur intégrité et leur sécurité. Des standards clairs et largement adoptés permettent de réduire les frictions et les coûts associés au partage de données, tout en augmentant la fiabilité et la transparence des interactions entre les parties prenantes. En adoptant ces standards, l'Europe peut créer un environnement de données plus cohérent et performant, qui pourra devenir le socle d'une Intelligence Artificielle Capable et Responsable.

A RETENIR

1. **La "solution" Data Space semble avoir trouvé son "problème" avec l'IA générative. Sans ces écosystèmes de confiance, la bataille de l'IA générative serait très probablement ingagnable.**
2. **Un pour tous, tous pour un : coopérer en faisant des Data Spaces les écosystèmes de confiance pour créer des Modèles de filière et des assistants IA souverains .**

[61] *"From LLMs to SLMs to SAMs, how agents are redefining AI"*, Dave Vellante, Scott Hebner and George Gilbert, septembre 2024

LE DATA SPACE DE LA FILIÈRE DU VOYAGE

Comme mentionné précédemment, le Data Space à l'origine du projet Gen4Travel s'appelle EONA-X.

Ce Data Space dédié à la filière du voyage (mobilité-transport-tourisme) a pour mission de réunir les acteurs du secteur, en leur offrant un cadre de gouvernance partagé ainsi qu'une infrastructure mutualisée et standardisée pour le partage de données (personnelles et non personnelles).

Le projet Gen4Travel, ainsi que les assistants de voyage qu'il permettra de créer, étant développé au sein du Data Space EONA-X, la gouvernance de l'infrastructure IA (GenAI, LAM, Agentic AI) y est strictement encadrée et contrôlée par l'ensemble des acteurs de la filière.

2.5.3 Participer à un espace de données de confiance

L'adhésion à un Data Space représente une opportunité inédite pour les entreprises et les institutions de collaborer autour de la valorisation des données, y compris sur les sujets d'IA générative. Un Data Space est un environnement sécurisé et régulé (par le cadre européen DGA, DA, etc.) où des organisations peuvent partager et échanger des données, stimulant ainsi l'innovation et la compétitivité.

En rejoignant un Data Space, vous pouvez accéder à des ressources informationnelles diversifiées, bénéficier des insights de partenaires de votre secteur et développer des solutions communes qui répondent aux défis de votre filière.

Prenez part à la gouvernance de votre filière

Participer à la gouvernance de votre filière au sein d'un Data Space est essentiel pour influencer les orientations stratégiques et garantir que les besoins spécifiques de votre secteur sont pris en compte. En prenant part aux décisions, vous pouvez aider à établir des standards, des protocoles de sécurité et des mécanismes de partage qui favorisent une collaboration efficace et équitable. Votre implication dans la gouvernance assure que les intérêts de votre organisation et de votre filière sont représentés et protégés.

Chassez en meute les financements

L'obtention de financements pour les initiatives liées aux Data Spaces peut être facilitée par une approche collective. En formant des consortiums avec d'autres acteurs de votre filière, vous pouvez "chasser en meute" pour accéder à des fonds plus importants et obtenir un soutien plus solide des institutions financières et des programmes de subventions. Cette approche collaborative permet de présenter des projets plus robustes et de démontrer un impact potentiel plus large, augmentant ainsi les chances de succès dans les appels à projets et les

demandes de financement. Dans ce domaine nous devons nous rapprocher des allemands qui ont cette culture de la coopération, et qui jusqu'à présent profitent davantage que nous des financements européens grâce à leur capacité former des alliances technologiques européennes. D'où la nécessité pour la France de s'organiser autour d'initiatives de places technologiques européennes comme celles proposées par Digital New Deal Do Tank[62].

Rejoignez un Data Space existant

Rejoindre un Data Space existant présente de nombreux avantages, notamment l'accès immédiat à une infrastructure établie et à un réseau de partenaires prêts à collaborer. Cela permet de gagner du temps et de bénéficier de l'expérience et des ressources déjà disponibles dans le Data Space. De nombreux secteurs ont déjà initié des Data Spaces, offrant ainsi des opportunités prêtes à l'emploi pour les entreprises souhaitant se lancer dans le partage de données de manière sécurisée et régulée (comme par exemple Agdatahub pour l'agriculture, Omega-X pour l'énergie, Prometheus-X pour l'éducation, Catena-X pour l'automobile, EONA-X pour la mobilité et le tourisme, etc).

Initiez un Data Space si votre filière ne l'a pas déjà fait

Si votre filière n'a pas encore de Data Space, il peut être judicieux d'en initier un. Cependant, il est crucial de ne pas multiplier les Data Spaces de manière excessive. Il vaut mieux créer de grands Data Spaces qui couvrent largement les problématiques de la filière, permettant ainsi de développer de nombreux cas d'usage impliquant différents acteurs. Cela maximise l'effet de réseau et évite la fragmentation qui pourrait diminuer l'efficacité et l'impact des initiatives de partage de données. Dit plus prosaïquement "C'est la même logique que celle des infrastructures physiques : on ne construit pas plusieurs routes ou plusieurs réseaux d'égout ayant les mêmes objectifs en parallèle".[63]

[62] Digital New Deal a créé un "Do tank" visant à structurer cette approche : création d'écosystème de confiance Cloud-Data-IA comme les Data Spaces, lancement de consortium pour répondre aux appels d'offres européens (exemple avec InfrateX, co-gagnant du call Simpl à 150 M€), etc

[63] Extrait du rapport "Les infrastructures publiques de partage de données", Laura Létourneau, Digital New Deal - Terra Nova, septembre 2024

DÉVELOPPER UNE CULTURE DES DATA SPACES

La création des Data Spaces est un processus complexe, exigeant non seulement des compétences techniques avancées en architecture, ingénierie et cybersécurité, mais également des capacités en gestion de projet, accompagnement du changement et conformité juridique. La rareté des talents dans ces domaines, amplifiée par la concurrence mondiale pour les compétences digitales représentent un défi majeur.

Il est d'abord impératif pour les entreprises d'adopter une culture de la donnée, en dépassant la simple utilisation d'outils pour s aisir la valeur des données partagées et leur exploitation stratégique, y compris au niveau d'une filière. C'est pourquoi ces transformations requièrent également une révision des programmes de formation au sein des entreprises, une planification rigoureuse, une formation adéquate des collaborateurs, ainsi qu'une stratégie RH adaptée

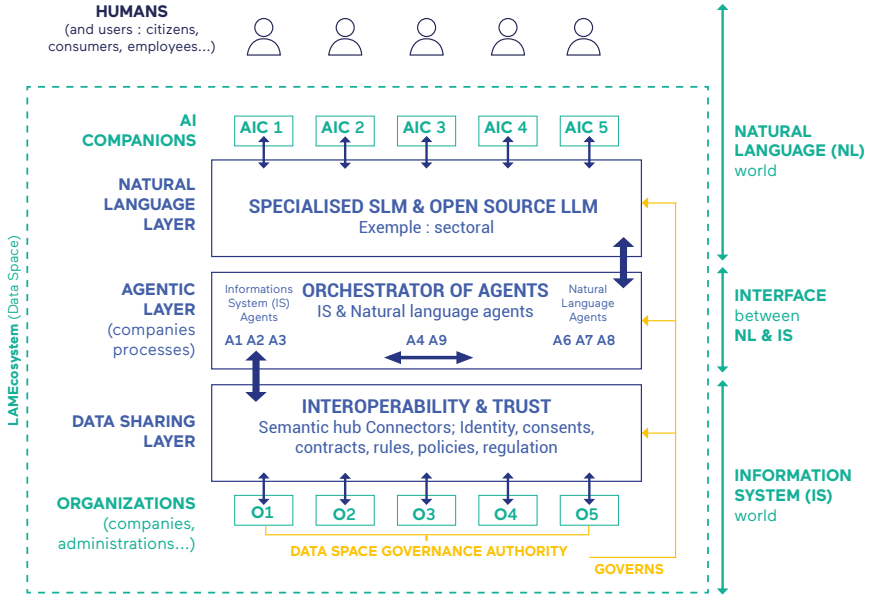
Cependant, en rejoignant un Data Space, les entreprises peuvent s'appuyer sur les compétences déjà présentes dans ces espaces, limitant ainsi le besoin de recrutements spécifiques. Le principal enjeu restant de développer une culture data suffisamment solide pour tirer parti efficacement de ces ressources et des potentialités permises par la mise en commun de moyens et des compétences au sein du Data Space.

A RETENIR

1. **Le Data Space, cadre d'une coopération numérique :** bâtir un écosystème de confiance par filière autour du partage des données (ex : EONA-X pour transport-mobilité- tourisme)
2. **La data de confiance, socle d'une IA de confiance :** s'appuyer sur cette infrastructure de données gouvernées pour exploiter grâce au RAG une IA générative de filière sur des données d'entreprises (ex : Gen4Travel depuis EONA-X)
3. **Des modèles d'IA réellement Open Source :** s'affranchir des GAMAM tout en innovant, en tirant partie de la collaboration ouverte et de la combinaison entre recherche publique et privée (ex : Gen4Travel avec Open LLM)
4. **L'agentification, "climax" Data-IA :** entraîner des "AI companion" ou "AI Assistants" souverains grâce au LAM et aux agents de la filière.

LA DATA DE CONFIANCE, SOCLE D'UNE IA DE CONFIANCE

III. CRÉER UNE NOUVELLE ARCHITECTURE POUR UNE IA CAPABLE ET RESPONSABLE

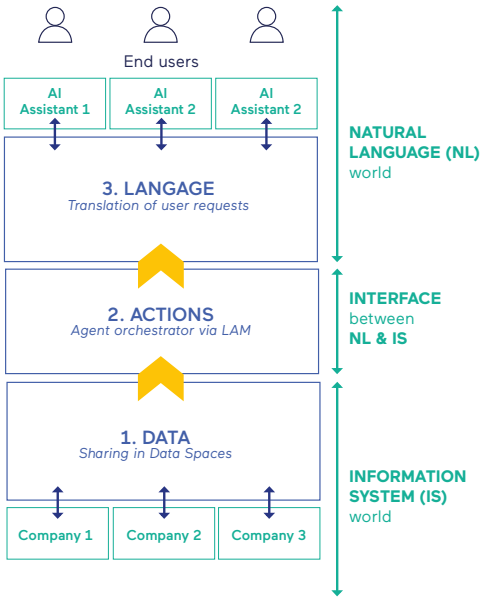


En résumé, l'architecture d'ensemble que nous proposons pour le développement d'une IA capable et responsable — vecteur d'une stratégie véritablement différenciante pour l'Europe — repose sur un nouveau modèle de coopération : les Data Spaces. Ces espaces de données permettent une gouvernance sectorielle optimisée pour le partage de données et l'intégration de services à forte valeur ajoutée, tels que l'IA générative.

Nous proposons d'appeler ce nouveau modèle l'architecture DUCAI (Data Union for Capable AI).

Les Data Spaces, dans toutes les filières, concentrent actuellement leurs développements sur le partage de données uniquement. Nous proposons une architecture de Data Space plus adaptée aux nouveaux besoins de l'IA en y ajoutant 2 niveaux. pour obtenir 3 couches :

1. Données
2. Actions
3. Langage



La couche 1 - Données est une composante classique des Data Spaces. Elle regroupe l'ensemble des outils logiciels qui permettent aux participants d'un **Data Space** (souvent au sein d'une filière) de partager des données entre eux, et avec d'autres Data Spaces à l'avenir. Cette couche **assure l'interopérabilité et la confiance entre les participants**, elle s'appuie sur des régulations et des standards européens. Elle intègre notamment des éléments tels que la gestion des identités, des contrats, des consentements, des connecteurs sémantiques, etc. Solidement ancrée dans l'univers des Systèmes d'Information des entreprises, cette couche garantit que chaque donnée (structurée ou non) est associée à un détenteur de droits (organisation ou individu), avec des règles strictes concernant son usage et ses finalités.

La couche 2 - Actions permet d'exploiter les données partagées au sein du Data Space pour réaliser diverses actions à la demande de l'utilisateur final, avec un degré d'autonomie variable pour chaque agent (LAM et approche Agentic AI). Elle regroupe une constellation d'agents (services ou programmes) capables d'accomplir des tâches spécifiques et variées, comme lister des offres, faire correspondre des offres avec des profils d'individus, réserver ou reprogrammer des services métier, etc. Cette couche inclut également un **orchestrateur, qui coordonne les agents pour exécuter des tâches complexes impliquant plusieurs agents.** Les agents peuvent être proposés par le Data Space lui-même si leur usage est général (ex : agents de traduction linguistique), mais ils peuvent également être proposés par des participants spécifiques (ex : tel acteur de l'hôtellerie externalise son processus de booking de chambres d'hôtel sous forme d'agent).

En outre, la couche Actions intègre un Large Action Model **(LAM**) unique (au niveau de la filière par exemple), qui **apporte de l'intelligence à l'orchestrateur et optimise les interactions entre les agents, en fonction des besoins des utilisateurs finaux** (les individus qui utilisent les Assistants IA). Enfin la couche 2 (Actions) contient des **agents en charge des requêtes RAG qui vont faire le pont avec la couche 3 (Langage)**. Ici, des approches multi-RAG mixant par exemple des recherches dans des bases documentaires classiques (après les avoir intégrées dans des bases vectorielles) avec des requêtes complexes s'appuyant sur des bases de type Knowledge Graph (KB) permettent de répondre de manière fiabilisée à des demandes complexes de l'utilisateur entraînant le déclenchement d'action dans plusieurs systèmes d'informations du Data Space.La couche 2 (Actions) fait ainsi le lien entre le monde des Systèmes d'Information (couche 1 - Données) et celui du langage naturel (couche 3 - Langage).

La couche 3 - Langage est dédiée à l'interaction avec les humains, et à **l'interprétation des demandes des utilisateurs finaux en langage naturel.** Elle est directement en lien avec la couche 2 (Actions), via les agents RAG notamment, qui permet de traduire ces demandes humaines en actions concrètes. Elle intègre plusieurs Large Language Models (LLM), les rendant interopérables. Cette couche, axée sur le traitement du langage naturel, est **directement utilisée par les assistants IA qui s'appuient sur le Data Space.**

L'importance de la séparation des couches

Prétraitement des données pour la couche 3 (Langage) :

Les modèles de langage (LLM) de la couche 3 (Langage) peuvent être créés de toute pièce, ou affinés et réentraînés à partir de modèles génériques si nécessaire. Cependant, conformément au principe de séparation des couches, les données brutes de la couche 1 (Données) ne sont jamais utilisées directement par ces modèles, elles doivent passer par la couche 2 (Action).

Les données brutes ne sont pas ce que recherche l'utilisateur final, ce sont bien les agents qui apportent de l'intelligence à la réponse en s'appuyant sur ces données. **Cette approche garantit également la protection des informations sensibles, qu'il s'agisse de données personnelles ou de propriété intellectuelle, tout en préservant l'expertise métier.** Un prétraitement, tel que l'anonymisation (pour le cas des données personnelles), est obligatoire avant que ces données ne puissent être exploitées par la couche 3.

Rôle central de la couche 2 (Agents) :

La véritable intelligence du système, définie comme la capacité à développer des raisonnements complexes et à les appliquer aux processus métier, est confinée à la couche 2 (Agents). **Cette couche est le cœur du système, où s'opèrent les analyses et les prises de décision sophistiquées.**

Fonction spécifique de la couche 3 (Langage) :

Dans ce contexte, la couche 3 (Langage) joue principalement le **rôle d'une interface de traitement du langage naturel.** Sa fonction se concentre sur la compréhension et la génération du langage, servant de pont entre les utilisateurs et le système, sans participer directement aux processus de raisonnement avancés et à l'intelligence métier.

Gouvernance du système DUCAI (Data Union for Capable AI)

Même si les données du système proviennent des participants, et que les participants peuvent proposer et opérer des agents eux-mêmes, c'est bien le Data Space lui-même qui assure la fonction de gouvernance pour les 3 couches. **La gouvernance du Data Space s'assure du respect de la séparation des couches et de la bonne application des règles communes sur chaque couche**.

L'ARCHITECTURE DUCAI DE GEN4TRAVEL

Par rapport à notre architecture en couches Data Union for Capable AI (DUCAI), c'est naturellement le Data Space EONA-X qui gouverne l'ensemble du projet Gen4Travel.

EONA-X dispose déjà de la couche 1 (Données). Le projet Gen4Travel vise par conséquent à augmenter Gen4Travel des couches 2 (Agents) et 3 (Langage), afin de permettre aux acteurs de la filière et technologiques de ce consortium de développer leurs Travel Assistant.

DERRIÈRE LA QUESTION DE L'IA GÉNÉRATIVE, SE PROFILE CELLE DE L'IA GÉNÉRALE

CONCLUSION

Si l'histoire nous a appris quelque chose, c'est que la grandeur d'une civilisation repose sur sa capacité à unir ses forces. C'est précisément ce que l'Europe doit accomplir pour devenir leader dans le développement d'une IA générative capable et responsable.

Face aux monopoles, l'Europe doit impérativement opter pour la voie de la coopération, en faisant des Data Spaces le cœur battant de ce marché numérique unique que nous appelons de nos vœux.

Cette vision porte un parfum de renaissance : une Europe qui se réinvente, non en se refermant sur elle-même, mais en construisant une collaboration intelligente entre l'humain et la machine, pour façonner ensemble un avenir qui soit choisi, et non subi.

En intégrant les Data Spaces, les entreprises européennes ne se contentent pas de sécuriser leurs données, elles établissent les fondations d'une Europe capable de fixer ses propres règles, en accord avec ses valeurs d'éthique et de transparence. Cette nouvelle architecture numérique est une opportunité pour positionner l'Europe en tant que leader mondial de l'IA. Il est crucial que les entreprises s'unissent autour de ces espaces communs, à l'image de la filière voyage, afin de créer les Gen4Legal, Gen4Finance ou Gen4Health de demain.

Cela revêt une importance stratégique d'autant plus grande que, derrière la question de l'IA générative, se profile celle de l'IA générale. La course vers l'AGI (Artificial General Intelligence), capable de surpasser les capacités humaines dans de nombreux domaines, structure déjà le marché américain, alimentée par des investissements massifs visant à concrétiser une vision libertarienne de la technologie.

Adopter une stratégie propre à l'Europe dans le domaine de l'IA générative permettrait non seulement de préserver notre autonomie, mais aussi d'éviter que quelques milliardaires américains ne dictent une nouvelle fois leur vision de l'avenir. L'Europe n'a ni les ressources ni l'intérêt de s'engager dans une nouvelle course au gigantisme technologique et financier. Elle doit plutôt aborder l'IA sous l'angle de l'urbanisation numérique et de la création d'écosystèmes de confiance, afin de rester au centre de l'architecture mondiale du numérique.

Si l'Europe parvient à exploiter les données privées pour tirer parti de l'IA générative, elle sera potentiellement mieux préparée à affronter les défis de l'IA générale.

Des intelligences artificielles éthiques, humanistes, reposant sur un réseau d'agents IA urbanisés, nourris par des données privées de confiance – voilà l'idéal européen que nous vous proposons.

OLIVIER DION

Olivier Dion, ingénieur télécom et pionnier de l'Open Data, est un expert reconnu en Europe dans le domaine du partage des données. Consultant et enseignant en data et en intelligence artificielle, il a fondé en 2011 la startup Onecub, spécialisée dans la portabilité et le partage des données personnelles. Il collabore avec les principaux acteurs des Data Spaces, interagissant avec diverses parties prenantes de la communauté européenne des données. Il a contribué à l'élaboration du RGPD (notamment le droit à la portabilité) avec la CNIL, du DGA sur les enjeux d'intermédiaires de données, et sur la stratégie européenne des données avec la Commission européenne. Olivier Dion joue un rôle clé dans la création de Data Spaces dans des secteurs variés (tourisme, mobilité, espace, énergie, éducation, etc.). Il est également co-auteur du rapport "Data de confiance : le partage des données, clé de notre autonomie stratégique" pour le Think-tank Digital New Deal, et coordonne des consortiums Data-IA, tels que InfrateX, Gen4Travel et Themis-X, initiés par le Do Tank - Digital New Deal.

MICHEL-MARIE MAUDET

Passionné par le potentiel de Linux dès les années 90, Michel-Marie a participé à la création de LINAGORA en 2000 et s'est engagé à promouvoir les standards et les technologies ouvertes anticipant l'importance cruciale de l'Open Source dans le paysage technologique d'aujourd'hui. Face à la domination des GAFAM dans le domaine de l'IA, Michel-Marie a lancé en 2016 LinTO, un assistant personnel entièrement Open Source, incarnant sa vision d'une technologie accessible, inclusive et transparente. En juin 2023, il a fondé la communauté OpenLLM France, visant à développer des communs numériques souverains, réellement open source, basés sur des données publiques et des algorithmes explicables.

Convaincu que l'avenir numérique de la France et de l'Europe se joue dans le code et les algorithmes, Michel-Marie œuvre pour une IA souveraine comme levier de croissance économique et de compétitivité industrielle. Il s'engage dans la promotion d'une troisième voie numérique, alliant ouverture, éthique et responsabilité, pour contrebalancer l'hégémonie américaine et chinoise dans le domaine de l'IA.

ARNO PONS

Délégué général du Think-tank-Digital New Deal, co-auteur de quatre rapports sur le numérique de confiance (Cloud de confiance ; Infrastructures du numérique de confiance ; Data de confiance ; et IA de confiance pour France 2030). Il a fondé l'activité Do Tank - Digital New Deal, dédiée aux enjeux de coopération pour accompagner les entreprises à structurer les filières en écosystèmes de confiance via des alliances technologiques. Le Do Tank a lancé depuis 2022 plusieurs initiatives de place numérique : SCOVERY - The European Cyber Scoring Agency ; INFRATEX - consortium européen sélectionné par la Commission européenne pour SIMPL ; THEMIS-X ayant fusionné ensuite avec EONA-X pour créer un data space unique ; GEN4TRAVEL consortium dédié à une IA générative pour la filière Voyage.

Il a également enseigné à SciencesPo sur les enjeux de souveraineté numérique liés à la centralisation des pouvoirs par les Big Tech, et a créé auparavant plusieurs start-ups (Checkfood - gaspillage alimentaire, Medicimo - Moteur de recherche Canadien, CityLuxe - Tourisme en ligne premium,…).

SYNTHÈSE DES PROPOSITIONS

RÉSUMÉ DE LA STRATÉGIE PROPOSÉE

LES AXES CLÉS DE LA STRATÉGIE GENAI :

- Prioriser la **confiance** (éthique et fiabilité) plutôt que la puissance.
- Adopter une approche d'**écosystème intégré Cloud-Data-IA**, plutôt que de se limiter à des solutions d'Intelligence Artificielle isolées.
- Anticiper l'essor des **assistants IA** en tant que nouvelles interfaces pour les utilisateurs humains avec nos entreprises et administrations.
- Se positionner comme pionnier dans la prochaine révolution de l'**agentification et des Large Action Models (LAM).**
- Exploiter le **partage et l'accès aux données privées** comme levier concurrentiel et stratégique pour la souveraineté.

COMMENT PASSER À L'ACTION :

Unir nos forces au niveau de chaque filière

1. Construire des **socles communs d'IA générative**, en s'appuyant sur l'Open Source, pour faciliter la **coopération sectorielle**.
2. Adapter et **personnaliser les modèles d'IA générative** pour répondre aux besoins spécifiques de chaque secteur.
3. Favoriser l'**approche RAG** (Retrieval-Augmented Generation), qui n'exige pas de réentraînement, pour personnaliser les modèles, plutôt que de recourir systématiquement au fine tuning.
4. Créer des **LAM sectoriels** (Large Action Models), associés à une architecture multi-agents (Agentic AI), capables de transformer les assistants conversationnels en véritables **agents intelligents**.
5. Permettre aux LAM d'accéder aux **données privées provenant des systèmes d'information des entreprises**, grâce au **Shared Data**.
6. Développer ces initiatives sectorielles au sein des **Data Spaces** dédiés.
7. Développer une **architecture DUCAI (Data Union for Capable AI)** en 3 couches distinctes (Données, Actions, Langage) au niveau du Data Space.

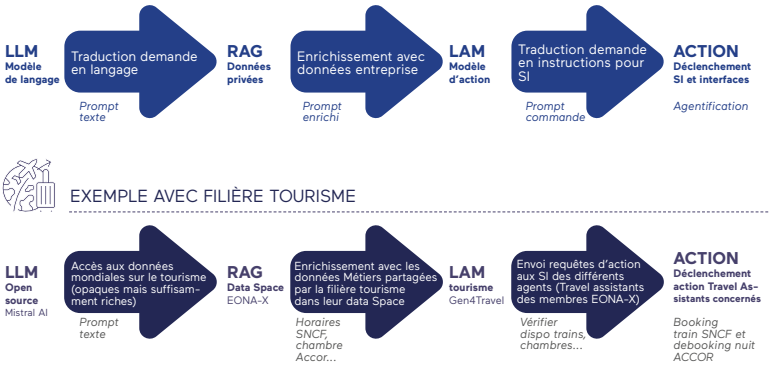
EN RÉSUMÉ :

Créons ensemble une IA "capable" et "de confiance"

- **Saisissons l'opportunité des Data Spaces** pour propulser nos secteurs clés (santé, mobilité, éducation, etc.) à l'avant-garde de l'IA générative.
- Faisons du **partage des données privées** de nos entreprises, le cœur de notre valeur, et notre principal levier de souveraineté.
- Dépassons les LLM, et **concentrons nous dès maintenant sur les LAM**, en créant de nouvelles capacités d'interaction avec le monde réel, qui vont transformer les assistants conversationnels en véritables agents à tout faire du quotidien.

NOTRE RECOMMANDATION EN 1 FORMULE POUR UNE
IA GÉNÉRATIVE AUTONOME DE CONFIANCE :

LLM OPEN SOURCE + RAG DES DONNÉES PARTAGÉES D'ENTREPRISES = LAM DE CONFIANCE



Digital New Deal

RÉSUMÉ EN 1 PAGE

Ce rapport propose une stratégie pour les entreprises européennes face à l'émergence de l'Intelligence Artificielle générative (GenAI).

Ne pas répéter les erreurs du passé

L'objectif est d'éviter la même naïveté que celle qui a marqué l'arrivée du Web2, il y a 25 ans, et qui nous a coûté cher en termes de compétitivité et de souveraineté. Nous devons prévenir ce que nous appelons un "enfermement au carré" (**GenAI = Web2²**). Si nous n'agissons pas, nos entreprises risquent de se retrouver encore plus coincées par les BigTechs, prises d'un côté par leurs **BigClouds** (hyperscalers comme Azure, AWS ou GoogleCloud) et de l'autre par leurs **BigAI** (comme OpenAI ou Gemini).

Se montrer pragmatique face au gigantisme

Nous devons être réalistes et concentrer nos efforts sur des objectifs atteignables. La course à l'accumulation de données pour les LLM est révolue (les Large Language Models sont déjà des produits standards selon Yann LeCun). Il ne sert à rien de chercher à scraper l'ensemble des données du web comme le font les géants. Non seulement cela soulève des questions éthiques et légales, mais cela nous fourvoie. L'Europe doit se concentrer sur une catégorie de données, celles qui ont **le plus de valeur et les seules que les BigTechs n'ont pas encore : les données de nos entreprises.**

Capitaliser sur les données privées et collaboratives

Nous proposons donc de capitaliser sur ces données privées en nous appuyant sur la **Data Strategy européenne**. Stratégie consistant à encourager la mutualisation et le partage des données entre entreprises pour créer un marché commun de la donnée comparable aux marchés américain et chinois, et pouvoir, enfin, bénéficier de l'effet de réseau. Concrètement, cela passe par la participation aux **Data Spaces**, des écosystèmes de confiance offrant la bonne échelle et gouvernance pour créer des IA génératives adaptées à chaque filière. C'est la meilleure solution pour éviter la dépendance au "Powered by OpenAI", ou la tentation contre-productive d'un modèle développé par une entreprise seule.

Faire le pari industriel des LAM (Large Action Models)

En réponse au rapport Draghi qui encourage à prendre des paris technologiques audacieux, nous proposons de miser sur les Large Action Models (LAM). Plutôt que de se focaliser sur les modèles linguistiques (LLM), nous pensons que les Data Spaces devraient servir à créer des IA génératives dédiées à **l'exécution autonome** d'actions (**agentification**). Ces modèles innovants offrent à l'Europe un avantage stratégique grâce à l'accès aux données d'entreprises, un domaine où les Data Spaces sont particulièrement adaptés.

Suivre l'exemple concret du data space transport-tourisme EONA-X avec Gen4Travel

EONA-X a lancé le consortium Gen4Travel pour développer un **LAM dédié au secteur du voyage.** Ce modèle permettra aux "Travel assistants conversationnels" des entreprises membres de devenir des **agents intelligents autonomes** pouvant automatiser des tâches complexes, telles que la réservation ou l'annulation de vols et d'hôtels en cas de problème (incluant la gestion des paiements). Prémunissant ainsi la filière du tourisme de la désintermédiation de trop par un "Booking ou Tripadvisor agentifié".

NOS REMERCIEMENTS

POUR LEURS CONTRIBUTIONS :

Arno Amabile — rapporteur général de la Commission IA

Guillaume Avrin — coordonnateur national pour l'intelligence artificielle (DGE)

Julien Chiaroni — ex-directeur Grand Défi Intelligence Artificielle (SGPI-France 2030)

Maxence Demerlé — directrice du numérique, MEDEF

Frédéric Josué — fondateur 18M.i.o, enseignant à Sciences Po

Francis Jutand — ex directeur général de l'Institut Mines Télécom (IMT)

Alexis Kasbarian — responsable du pôle transition et innovation, MEDEF

David Krieff — président EONA-X, directeur des systèmes d'information des Aéroports de Paris (ADP), administrateur du Cigref

Yann Lechelle — co-founder CEO :Probabl, co-fondateur France Digitale, entrepreneur in residence INSEAD et Inria

Yves Nicolas — deputy group CTO, AI group program director, Sopra Steria

POUR LEUR ASSISTANCE :

ChatGPT (Open AI), **Le Chat** (Mistral AI) et **Perplexity** qui ont aidé les auteurs à la rédaction de ce rapport.[64]

[64] En cas d'erreurs factuelles, grammaticales ou syntaxiques dans ce rapport, cela restera l'entière responsabilité des auteurs...

DIGITAL NEW DEAL
THINK-DO-TANK DE LA NOUVELLE DONNE

Digital New Deal accompagne les décideurs privés et publics dans la création d'un Numérique des Lumières, Européen et Humaniste. Notre conviction est que nous pouvons offrir une troisième voie numérique en visant un double objectif : défendre nos valeurs en proposant un cadre de confiance par la régulation (think-tank) ; et défendre nos intérêts en créant des écosystèmes de confiance par la coopération (do-tank).

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

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

SÉBASTIEN BAZIN
PDG AccorHotels

NATHALIE COLLIN
General Manager, Consumer and Digital Division La Poste Group

NICOLAS DUFOURCQ
DG of Bpifrance

AXELLE LEMAIRE
Former Secretary of State for Digital Technology and Innovation

ALAIN MINC
President AM Conseil

DENIS OLIVENNES
DG Libération

ODILE GAUTHIER
DG de de l'Institut Mines-Telecom

ARNO PONS
General Delegate of the Digital New Deal think tank

JUDITH ROCHFELD
Associate Professor of Law, Panthéon Sorbonne

OLIVIER SICHEL
President Digital New Deal
DGA Caisse des Dépôts

BRUNO SPORTISSE
PDG Inria

ROBERT ZARADER
PDG Bona fidé

Digital New Deal

OUR PUBLICATIONS

Infrastructures publiques de partage des données : les grandes oubliées I Laura Létourneau - *septembre 2024*

Strengthening pan-european democracy in the era of AI I Axel Dauchez, Hendrik Nahar - *avril 2024*

Le numérique au service d'un futur durable I Véronique Blum, Maxime Mathon - *juin 2023*

IA de confiance, opportunité stratégique pour une souveraineté industrielle et numérique | Julien Chiaroni, Arno Pons - *juin 2022*

Data de confiance, le partage des données, clé de notre autonomie stratégiques | Olivier Dion, Arno Pons - *septembre 2022*

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*

Novembre 2024

www.thedigitalnewdeal.org