



HUB
FRANCE
IA

Intelligence Artificielle Générative

**40 mesures pour faire de la France
et de l'Europe des leaders de l'IA d'ici 5 ans**

Février 2024



L'IA Générative, un saut technologique et d'usage

Les Grands Modèles de Langue (ou LLM) et l'IA Générative représentent un saut technologique majeur dans la compréhension et le traitement du langage, des images, des vidéos, des sons. Ces modèles et technologies constituent des leviers de productivité stratégiques pour les acteurs économiques quelles que soit leur taille. Ces technologies auront également des impacts importants au niveau individuel et sociétal du fait de leur diffusion fulgurante et leurs usages multiples.

Il est par conséquent nécessaire et urgent de promouvoir leur développement dans un cadre et des conditions qui assurent l'excellence tout autant que la souveraineté, la transparence et la maîtrise des risques.

Pour le Hub France IA, quatre objectifs principaux doivent être tenus pour permettre à la France et à l'Europe de se démarquer pour devenir d'ici 5 ans les leaders de l'IA Générative dans le respect des valeurs européennes.

Développer une approche holistique, respectueuse de l'environnement et des règles éthiques européennes, rassemblant fournisseurs ET consommateurs publics ET privés.

Anticiper les évolutions de l'IA en structurant le marché de l'accès aux données d'entraînement et l'architecture des super-calculateurs européens autour des prochaines ruptures.

Garantir un environnement qui permette l'émergence d'un marché réellement concurrentiel tout en préservant le marché des médias et de la culture.

Apporter le soutien nécessaire à la montée en compétence mais également à la gestion des risques et la mise en œuvre des réglementations.

Rim Tehraoui

Présidente du Hub France IA



IA Générative : les constats...

- ***Une adoption fulgurante mais non sans risques***
- ***Une offre européenne qui peine à décoller***
- ***Des impacts environnementaux probablement significatifs bien qu'encore mal quantifiés***
- ***Un big bang pour l'univers des médias et de la culture***

40 mesures pour...

- **Favoriser la souveraineté et l'excellence**
- **Stimuler et soutenir la croissance de l'écosystème**
- **Faciliter l'accès aux données dans un cadre de confiance**
- **Accompagner les entreprises dans la gestion des risques et la régulation**
- **Piloter et optimiser l'empreinte environnementale**
- **Accélérer la montée en compétence pour favoriser l'adoption**



Un marché rapide, des impacts majeurs, l'Europe à la traine en termes de financement et de concurrence

1. Une adoption fulgurante mais non sans risques

Il est généralement admis aujourd'hui que les LLM, et l'IAG (IA générative) plus généralement, font partie des ruptures technologiques majeures de ces dernières années avec pour fait caractéristique une courbe d'adoption fulgurante aussi bien dans le monde professionnel que dans la sphère personnelle.

L'appétence des entreprises, toutes tailles confondues, pour développer des technologies, des produits ou des services qui s'appuient sur des modèles d'IA générative augmente de façon exponentielle.

*A titre d'illustration, les projets développés dans le cadre du dispositif Pack IA¹ centrés autour de l'usage de l'IAG représentent **environ 50% des appels entrants sur les 6 derniers mois**. L'objectif vise généralement à améliorer l'efficacité opérationnelle dans le cadre de la recherche documentaire, du service après-vente ou encore l'extraction de données non structurées.*

*On observe également sur la même période que les dossiers soumis **utilisent de plus en plus souvent des LLMs** (sans que ce soit le sujet principal nécessairement, ni la seule technologie mise en œuvre). Ces technologies prennent de plus en plus souvent le relais de briques de NLP (Natural Language Processing).*

Il est intéressant de souligner une prise en main de plus en plus directe par les collaborateurs, ce qui représente un changement de paradigme par rapport aux projets IA plus classiques qui nécessitent généralement des temps de déploiement et d'appropriation plus longs.

Avec l'IAG, l'adoption apparaît ainsi plus rapide mais présente également plus de risques :

¹ Le Pack IA est un dispositif de la Région Ile de France adressant principalement les PME et ETI pour une aide au financement de leur premier projet d'IA. Le Hub France IA est copilote du dispositif qui a permis d'accélérer 100 PME sur le sujet de l'IA. Source : <https://www.packia.fr/>

- Mauvaise utilisation, incompréhension des hallucinations, lassitude et désengagement/ désintérêt
- Risques liés à la propriété intellectuelle ou à la protection des données,
- Désinformation incontrôlable, risques cyber
- Fracture numérique accélérée entre ceux qui sont déjà en pleine maîtrise et ceux qui en ont peur et/ou vont avoir besoin de se former
- Précarisation d'emplois (travailleurs du click)

A cela s'ajoute le fait que la majorité des entreprises souhaitant s'appuyer sur des modèles d'IA générative se dirigent aujourd'hui vers des solutions américaines pour développer leurs technologies et ce pour deux raisons simples : d'une part, une expérience utilisateur très fluide (offres ultra packagées, API prêtes à l'usage, accompagnement clé en main sur le « prompting ») et d'autre part, une tarification ultra compétitive.

Une entreprise souhaitant minimiser les risques liés à la souveraineté et/ou la confidentialité des données et personnaliser ses modèles grâce au « transfer learning » (apprentissage par transfert), devra investir à un coût relativement rédhibitoire dans de la puissance de calcul et des capacités de stockage massifs (machines, cartes graphiques, mémoire, espace disque SSD) à quoi s'ajoutent le coût important de l'électricité, sans oublier celui des experts.

2. Une offre européenne qui peine à décoller

Selon une étude² du Hub France IA publiée en janvier 2024 conjointement avec AppliedAI et d'autres partenaires européens, 10,6% des start-ups formant l'écosystème européen, sont spécialisées dans l'IAG, les start-ups françaises représentant 17,5% de cette sous-catégorie. Si le potentiel de développement de nouveaux produits et services basés sur l'IAG paraît immense, ces start-ups font état de nombreuses difficultés qui peuvent freiner leur croissance.

2

https://www.hub-franceia.fr/wp-content/uploads/2024/01/Study_Generative_AI_in_the_European_Startup_Landscape_2024.pdf

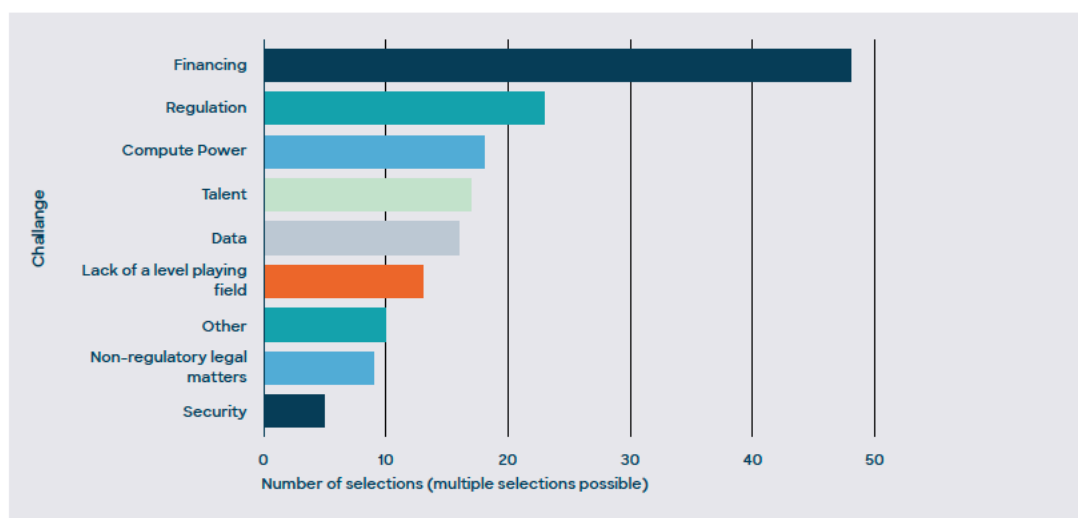


Figure 14: Challenges of Gen. AI Startups

Extrait de : https://www.hub-franceia.fr/wp-content/uploads/2024/01/Study_Generative_AI_in_the_European_Startup_Landscape_2024.pdf

Les difficultés principales rencontrées par les startups de l'IAG en Europe :

- L'accès au financement au niveau européen
- La complexité et la crainte réglementaire
- L'accès à un coût maîtrisé aux puissances de calcul

S'agissant du **financement**, l'un des principaux défis mentionnés par les startups est la **faiblesse structurelle des financements européens**, en particulier par rapport aux États-Unis et à la Chine. Cette difficulté découle de causes multiples, certaines étant conjoncturelles et liées à la situation des taux d'intérêt ou les difficultés des start-ups à lever des fonds européens de façon plus générale, d'autres étant plus structurelles et liées à un manque de compréhension et/ou de confiance des investisseurs dans les potentiels de l'IA générative, leurs sous-jacents technologiques et les coûts associés de développement et enfin les modèles économiques.

Sur les aspects **réglementaires**, si de nombreuses start-ups restent attachées à l'importance de protéger les valeurs européennes et la société dans son ensemble, de nombreuses préoccupations émergent en raison d'un cadre législatif perçu tour à tour comme incertain, trop complexe et restrictif. **Qu'il s'agisse de l'AI Act, du RGPD ou des législations protégeant les droits d'auteur ou les citoyens contre les risques de manipulation, de nombreuses préoccupations émergent face au**

potentiel de violation par inadvertance, de charges de preuve déraisonnables, du poids et du coût de mise en conformité pour de petites entreprises, ou encore l'effet paralysant de la protection des données sur les clients potentiels.

Ces startups font également état de grandes difficultés à accéder à de la **puissance de calcul** suffisante ou à absorber le coût élevé des GPU, ainsi que les coûts liés à l'utilisation du cloud. Elles indiquent également avoir du mal à capitaliser pleinement sur les capacités que l'EuroHPC est supposé mettre à disposition de l'écosystème.

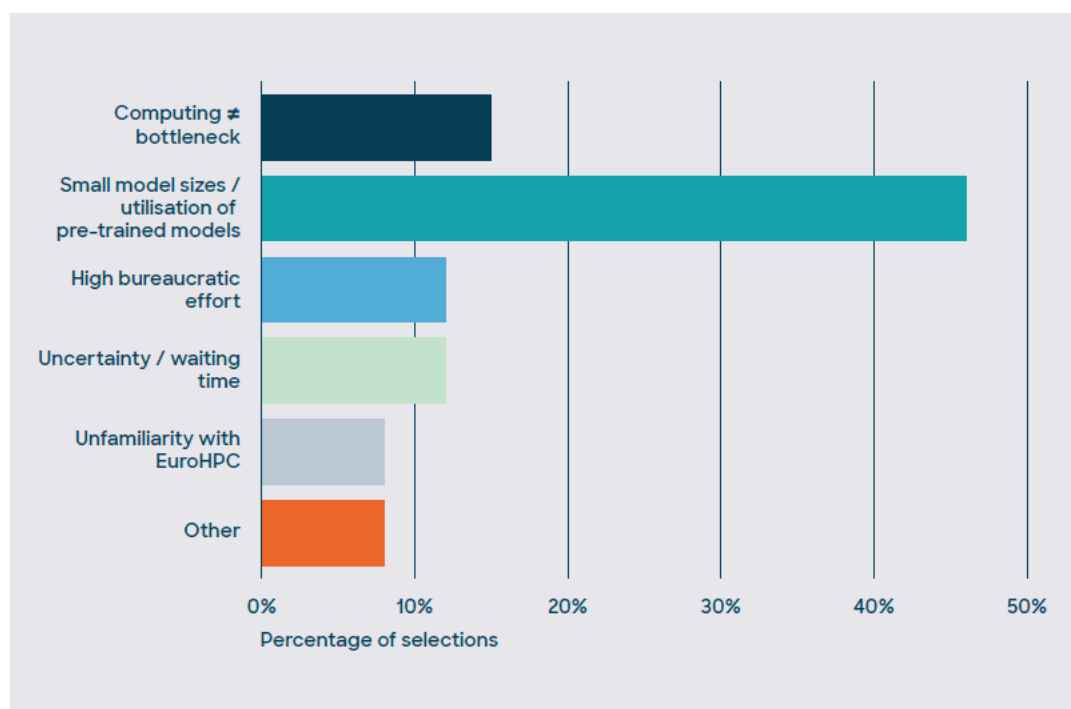


Figure 17: Reasons for not requiring EuroHPC access

Extrait de : https://www.hub-franceia.fr/wp-content/uploads/2024/01/Study_Generative_AI_in_the_European_Startup_Landscape_2024.pdf

Enfin s'agissant **des données**, les startups de l'IA générative sont également confrontées à plusieurs obstacles lorsqu'il s'agit de l'accès et la collecte de données de haute qualité. Là encore, les causes sont duales avec à la fois des difficultés liées aux limitations et incertitudes juridiques diverses et variées (RGPD, droits d'auteur, ...), en particulier dans des domaines sensibles tels que les ressources humaines, la santé ou la défense, et d'autres concernant les difficultés techniques à accéder aux données du fait de la complexité, et dans certains cas, l'obsolescence des infrastructures informatiques où elles sont stockées. Or la disponibilité de données de haute qualité est une condition sine qua non pour développer des modèles spécifiques, pertinents et frugaux et établir un avantage concurrentiel à long terme.

3. Des impacts environnementaux probablement significatifs bien qu'encore mal quantifiés

Il existe aujourd'hui très peu d'études reposant sur une analyse de cycle de vie complète d'une solution basée sur l'IA générative faute de données d'impact transparentes de la part des constructeurs et fournisseurs.

De telles études nécessiteraient d'intégrer le cycle de vie des équipements numériques, qui comprend en amont l'extraction des matières premières, la fabrication et la distribution des serveurs et de ses composants électroniques (GPU, en plus des composants d'un serveur classique : CPU, RAM, disques, carte mère, alimentation), auxquels il faut ajouter la fin de vie et la consommation électrique des serveurs accueillant des modèles d'IA.

Des comparaisons des impacts liés aux phases d'entraînement et d'inférence sont sporadiquement disponibles à l'instar de celle conduite sur BLOOM³, un modèle de langage très proche de GPT-3⁴. Ces études montrent que dans le cas de l'IA, **il semblerait que la part des émissions de gaz à effet de serre liées à la phase d'inférence dépasse largement celle liée à l'entraînement⁵, du fait essentiellement du très grand nombre d'utilisateurs.**

4. Un big bang pour l'univers des médias et de la culture

Le monde de la culture et des médias est en pleine ébullition depuis l'apparition de ChatGPT. Entre l'emballlement pour des gains de productivité permettant aux acteurs de l'information de se recentrer sur l'essentiel et la **crainte de risques de désinformation et de manipulation** encore plus grands que ceux déjà générés par les réseaux sociaux, le secteur fait face à un problème aussi important qu'inédit, celui du **pillage** de ses contenus pour entraîner des modèles de plus en plus gigantesques.

L'enjeu est de trouver un juste équilibre entre un extrême particulièrement injuste pour les auteurs et les créateurs où les développeurs de modèles LLM utilisent les données disponibles dans le web sans aucun droit d'auteur et un autre où ils devraient acheter toutes les données qui leur sont nécessaires avec le risque que seuls les mastodontes de l'IA puissent soutenir ce modèle dans la durée.

³ <https://arxiv.org/pdf/2211.02001.pdf>

⁴ <https://towardsdatascience.com/chatgpts-electricity-consumption-7873483feac4>

⁵ <https://www.semianalysis.com/p/the-inference-cost-of-search-disruption>



Se posent donc deux questions :

Quel modèle pour **soutenir l'industrie européenne des LLM** pour créer de la concurrence sans léser l'industrie de la culture

Dans quelle mesure un tel modèle serait **reproductible pour tous les droits d'auteur** : livres, actualités, musique, films, jeux, documentaires scientifiques...

Quatre modèles existent aujourd'hui sur le marché : Sacem like (musique), Droits voisins (internet), Droits directs (tous), Taxe de création (films).

Côté LLM, nous pouvons distinguer trois types de besoins :

- Données pour l'entraînement de LLM génériques : de très grands ensembles de données (ordre de grandeur en trillion de tokens) avec beaucoup de variété
- Données pour une production de LLM spécifiques : grands jeux de données (ordre de grandeur en milliards de tokens) sur des domaines spécifiques (scientifiques, actualités, ...)
- Données pour l'entraînement conversationnel ou le réentraînement et/ou la mise à jour : de grands jeux de données (ordre de grandeur en milliards de tokens) de haute qualité pour l'apprentissage par renforcement, l'optimisation des préférences directe... ainsi que pour la mise à jour des réponses (en temps réel comme pour les actualités, ou régulièrement)

Du point de vue des LLM, tous ces usages ont besoin de jeux de données volumineux de plus ou moins bonne qualité et relativement actualisés, notamment pour les deux derniers points.

Les recommandations de l'écosystème IA

1. Favoriser la souveraineté et l'excellence

Les acteurs européens fournissant ou adoptant des solutions d'IA générative sont de facto strictement contraints dans leurs choix du fait de la situation quasi monopolistique d'un nombre limité de grands acteurs tout au long de la chaîne de valeur de l'IA générative. Or il ne peut y avoir de souveraineté sans la possibilité de choix que ce soit en termes de solutions logicielles, d'architectures ou de GPU.

Il importe par conséquent de déployer une stratégie industrielle globale de mise à disposition des leviers nécessaires à chaque étape de la chaîne de déploiement, à un coût maîtrisé et avec une empreinte environnementale minimisée.

Cette stratégie doit être européenne « by design » pour s'acheminer vers un plan d'investissement massif de plusieurs dizaines de Mds d'euros supportés par les secteurs publics et privés, complétés par des investissements d'achats publics et privés au moins équivalents.

1.1 Pour l'entraînement des modèles

1. Adresser les difficultés rencontrées par les start-ups pour accéder aux **capacités de calcul de l'EuroHPC**
2. Tirer les leçons de cette expérience pour optimiser les choix d'architecture ainsi que les **accès à Jean Zay et d'ici 2025 Jules Verne**
3. A court-terme, encourager un **environnement plus compétitif** pour optimiser le coût et les délais d'accès aux GPUs
4. A moyen-terme, investir de façon ambitieuse pour établir **un écosystème de champions industriels européen** pour traiter l'amont de la chaîne de valeur (équipements pour la gravure, gravure, fabrication et commercialisation des cartes graphiques)
5. Préparer Jules Verne pour répondre aux évolutions de l'IA anticipées pour 2025-2030, notamment l'hybridation entre simulation physique, apprentissage et génération

1.2 Pour l'inférence

6. Inciter les travaux de recherche sur la performance des modèles réduisant le besoin de CPU pour optimiser l'équilibre performance – robustesse – coût d'inférence
7. Optimiser l'inférence en **capitalisant sur des offres européennes de CPUs** présentant le double avantage de la souveraineté et de la frugalité énergétique
8. Mettre en place des **dispositifs incitatifs** pour permettre aux **fournisseurs de cloud souverain** de mettre à niveau les offres PaaS et GPU et de mettre à disposition des capacités pour l'entraînement et l'inférence à **des coûts contrôlés et avec une empreinte écologique maîtrisée**
9. Simplifier les processus d'accès à l'EuroHPC pour les besoins de fine-tuning/transfer learning

1.3 Pour l'intégration des IAG

10. Proposer **des bacs à sable pour expérimenter et stimuler l'innovation**, comme par exemple les « Regulatory Sandboxes » dans le cadre de l'AI Act ou les TEF (Testing and Experimentation Facility), de la Commission Européenne⁶, et mettre en place les ressources et relais pour encourager les entreprises françaises à utiliser ces dispositifs, qui doivent intégrer ou donner accès à des capacités de calcul importantes
11. Favoriser l'émergence de **briques d'intégration européennes** (LLM Ops, Systèmes d'IA) via une double démarche d'achats publics et d'accessibilité dans les places de marché des hyperscalers et des intégrateurs

2. Stimuler et soutenir la croissance de l'écosystème

12. Lancer un « EU foundation model challenge » et **accélérer l'initiative « IEC Scale-up Europe »**
13. **Favoriser l'« open source »** pour garantir l'indépendance, la souveraineté et baisser les barrières à l'entrée pour les acteurs européens

⁶ [https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI\(2022\)733544_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf) <https://digital-strategy.ec.europa.eu/en/activities/testing-and-experimentation-facilities>

14. Favoriser la concurrence en rendant obligatoire l'accès aux modèles des hyperscalers **en dehors de leurs environnements cloud**
15. Sécuriser un **quota fixe de marchés publics** pour les startups françaises et adapter les processus de référencement en conséquence (« small business act » dédié à l'IA)
16. **Favoriser l'achat par les acteurs privés de solutions européennes**, en incitant par exemple le fléchage de 10% du budget IT des entreprises sur CAC 40 ou du SBF 120 vers ces solutions, soit environ 10Md par an via une incitation fiscale type CIR
17. Développer des **places de marché performantes chez les fournisseurs de cloud et les ESN français/européens**
18. Adresser les problématiques **de concurrence et de règles équitables** en sécurisant pour les fournisseurs français une présence **dans les places de marché des « hyperscalers »**
19. Adapter le programme IA Booster France 2030 pour accompagner les PME et les ETI dans l'adoption de l'IAG en intégrant **les problématiques de coût de développement (ou de licence)**
20. Flécher **le support public vers le développement d'offres B2B** pour développer avec les bons talents et les bons outils des produits d'IAG respectueux des règles éthiques et des réglementations européennes, basés sur des modèles plus petits, moins chers à produire, plus frugaux et plus pertinents pour les usages des entreprises

3. Faciliter l'accès aux données dans un cadre de confiance

21. Créer des fédérations de **bases de données européennes** (texte, audio, vidéo) **normées et qualifiées** pour diversifier et améliorer les modèles de langage et les sécuriser dans un cloud européen souverain et mettre en place **des bacs à sable pour l'entraînement hébergés sur ces bases données au GENCI** (et autres équivalents européens)
22. Mettre en œuvre la proposition du ministère de l'Économie et des Finances **d'un marché unique de la donnée avec une place de marché** fondée sur un modèle économique permettant d'éviter une spéculation effrénée à l'avantage des acteurs en situation de monopole
23. Lancer un chantier transverse sur la **gestion des droits d'auteur permettant l'émergence des modèles européens tout en préservant le marché des médias et de la culture, avec un cadre clair et transparent de monétisation de la donnée et des moyens concrets pour le mettre en œuvre**

4. Accompagner les entreprises dans la gestion des risques et la régulation

Avec l'adoption de l'AI Act, le principal enjeu consiste à présent à accompagner les acteurs fournissant ou adoptant des solutions d'IA avec des lignes directrices claires sur la manière de mettre en œuvre les exigences de cette réglementation.

24. Garantir une **mise en place équilibrée et progressive** de l'AI Act permettant d'associer innovation et protection pour les fournisseurs de modèles ou de systèmes d'IA générative
25. **Déployer un dispositif d'accompagnement opérationnel des fournisseurs et des consommateurs de solutions pour implémenter l'AI Act.** Organiser une boucle de retours terrains vers les autorités compétentes afin de faire évoluer le texte sur les mesures les plus problématiques à travers le dispositif d'actes délégués
26. Faciliter l'accès à des outils de mise en conformité pour la donnée, les modèles et l'inférence
27. Intégrer la taxonomie et les recommandations NIST⁷ pour se prémunir des risques cyber
28. Promouvoir la recherche sur la « trustworthiness » en s'appuyant sur les dispositifs européens existants

5. Piloter et optimiser l'empreinte environnementale

29. Déployer un **cadre normalisé pour l'analyse du cycle de vie de l'IA générative** intégrant le hardware (serveurs, cartes graphiques, ...), l'entraînement et l'inférence
30. **Valoriser la frugalité des solutions souveraines européennes** et l'énergie décarbonée française à un prix accessible comme un atout compétitif
31. Acculturer et éduquer les utilisateurs (entreprises et individus) sur les enjeux environnementaux afin **d'appuyer l'utilisation de l'IA uniquement vers les usages qui le nécessitent vraiment** (ex. il n'est pas nécessaire d'utiliser des LLM pour faire une simple recherche. Souvent des briques de NLP classiques ou un modèle de classification fine-tuné suffisent largement)
32. Accompagner les développeurs de solutions dans **l'adoption d'approches plus frugales** en fonction des cas d'usage (techniques de « few shot learning », choix de modèles de petite taille...)

⁷ https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf?mc_cid=8b7c480e03&mc_eid=c6e04f00af

33. Inciter les fournisseurs de solutions à mettre en place des systèmes automatiques pour **orienter vers des modèles pertinents et frugaux** par opposition aux modèles les plus puissants
34. Limiter la multiplication des solutions et favoriser les solutions open source et partagées afin d'éviter le gaspillage de ressources et la redondance
35. **Mesurer et piloter l'effet rebond** et les incidences sur l'environnement.

6. Accélérer la montée en compétence pour favoriser l'adoption

36. Favoriser l'enseignement de l'informatique et de l'IA dès l'école primaire
37. Établir une **cartographie des besoins en compétences IAG** au sein des entreprises en fonction de leur typologie et de la typologie des métiers impactés, notamment la transformation des métiers au-delà des métiers IA et IT
38. **Accélérer le processus de labellisation de nouvelles formations**, certifiantes de préférence : parcours d'acculturation, certification de compétence sur les outils d'IA générative, re-skilling à l'échelle (stratégie de massification)
39. Promouvoir **la transparence des solutions intégrant de l'IAG** pour faciliter l'acceptabilité par les utilisateurs
40. Mettre en place **un observatoire de l'IAG afin de piloter dans le temps l'évolution de l'adoption, des usages, de la prise en main des métiers, des risques et des impacts** (productivité, emploi, effets rebonds, environnement) en s'appuyant par exemple sur le projet européen « DeployAI »⁸.

⁸ Le projet DeployAI financé par la CE a débuté en janvier 2024 pour une durée de 3 ans, dans le cadre du programme : [DIGITAL-2022-CLOUD-AI-B-03-AI-ON-DEMAND](#) est lié à « AI-On-Demand »



A propos du Hub France IA :

Le Hub France IA est une association fédérant l'écosystème IA français et européen. Le Hub France IA agit pour accélérer le développement et l'adoption d'une IA responsable, éthique et souveraine par l'ensemble du tissu économique. Il est né en 2017 d'une volonté commune des acteurs privés et publics de création d'un écosystème français de l'intelligence artificielle. Le Hub France IA compte environ 150 membres (startups, grands groupes, institutions porteuses de projets d'IA) regroupés au sein de l'association pour développer des projets opérationnels. Notre action suit 3 axes principaux :

- la création d'un acteur de référence en interaction étroite avec les autorités compétentes en France et en Europe ;
- la fédération et l'animation de l'écosystème IA français et européen ;
- l'accompagnement à l'adoption et à la montée en compétence des acteurs économiques.

Plus d'informations sur : <https://www.hub-franceia.fr>

Envie de discuter de ces propositions avec nous ?

Votre point de contact : mehdi.triki@hub-franceia.fr



HUB FRANCE IA