



**OPECST**

**ChatGPT, et après ?  
Bilan et perspectives de l'intelligence  
artificielle**

*Alexandre Sabatou, député,  
Patrick Chaize, sénateur,  
Corinne Narassiguin, sénatrice,  
rapporteurs*

**28 novembre 2024**

*ChatGPT, et après ?*  
*Bilan et perspectives de l'intelligence artificielle*

*« Le progrès n'est plus dans l'homme, il est dans la technique,  
dans le perfectionnement des méthodes capables de permettre  
chaque jour une utilisation plus efficace du matériel humain »*

Georges Bernanos, *La France et les Robots*,  
Éditions de la France libre, 1946, page 11.



## SOMMAIRE

	<u>Pages</u>
L'ESSENTIEL.....	13
AVANT-PROPOS .....	15
<b>PREMIÈRE PARTIE : COMPRENDRE LES TECHNOLOGIES D'INTELLIGENCE ARTIFICIELLE .....</b>	<b>21</b>
<b>I. HISTOIRE DE LA NOTION D'INTELLIGENCE ARTIFICIELLE .....</b>	<b>21</b>
A. LA PRÉHISTOIRE DE L'INTELLIGENCE ARTIFICIELLE ET SES REPRÉSENTATIONS CULTURELLES .....	21
1. <i>De la mythologie antique aux machines à calculer .....</i>	<i>22</i>
2. <i>Un thème traditionnel de la science-fiction.....</i>	<i>25</i>
B. LES AVANCÉES DEPUIS UN SIÈCLE : DU BAPTÊME DE 1956 À LA CONFIRMATION DE 2017 .....	27
1. <i>L'école d'été de Dartmouth de 1956, le moment fondateur de la définition de l'IA .....</i>	<i>27</i>
2. <i>Printemps et Hivers de l'IA .....</i>	<i>32</i>
3. <i>Les années 2010 : une décennie d'innovations et de progrès spectaculaires .....</i>	<i>38</i>
4. <i>« L'intelligence artificielle n'existe pas » : la question de la « frontière » de l'IA.....</i>	<i>41</i>
C. LA SUMMA DIVISIO DE L'IA : INTELLIGENCE ARTIFICIELLE SYMBOLIQUE ET INTELLIGENCE ARTIFICIELLE CONNEXIONNISTE .....	42
1. <i>L'intelligence artificielle symbolique .....</i>	<i>42</i>
a) Principes généraux des IA symboliques .....	42
b) Quelques illustrations de ces technologies .....	43
c) Des limites sémiotiques qui les éloignent de l'intelligence .....	44
2. <i>L'intelligence artificielle connexionniste .....</i>	<i>46</i>
a) Cadre et définition de ces « superstatistiques ».....	46
(1) À l'origine de tous ces systèmes : les classifieurs linéaires .....	46
(2) La pierre angulaire théorique : le théorème d'approximation universelle.....	48
b) Les réseaux de neurones artificiels : aux origines de l'apprentissage profond ou <i>Deep Learning</i> .....	48
(1) Les premières théories dans les années 1940 .....	48
(2) Les premiers perceptrons.....	51
(a) <i>Les perceptrons monocouches .....</i>	<i>51</i>
(b) <i>Les perceptrons multicouches (MLP) et les réseaux de neurones à propagation avant (FNN) .....</i>	<i>54</i>
(3) La rétropropagation du gradient ( <i>Back-propagation</i> ).....	60
(4) Les réseaux de neurones convolutifs (CNN) .....	62
(5) Les réseaux de neurones récurrents (RNN) .....	65
(6) Les réseaux de neurones à mémoire court et long terme (LSTM).....	67
c) Les autres systèmes d'apprentissage .....	68
(1) Les machines à vecteurs de support (SVM) .....	68
(2) Les modèles markoviens ou « chaînes de Markov ».....	73
(3) La contribution des réseaux bayésiens à l'IA.....	75
(4) L'apport de la « régularisation statistique » de Vapnik .....	75

---

<b>II. LES IA GÉNÉRATIVES : PRINCIPALE INNOVATION TECHNOLOGIQUE EN INTELLIGENCE ARTIFICIELLE DEPUIS 2017 .....</b>	<b>76</b>
<b>A. LES PRINCIPALES AVANCÉES EN MATIÈRE DE GÉNÉRATION DE CONTENU CES DERNIÈRES ANNÉES .....</b>	<b>76</b>
1. « Attention is all you need » : <i>la technologie Transformer inventée en 2017</i> .....	77
a) La « tokenisation » : découper préalablement les mots.....	78
b) Le plongement lexical : vectoriser les <i>tokens</i> .....	79
c) Le mécanisme d'attention : intégrer le contexte .....	84
d) Un apprentissage auto-supervisé et des données synthétiques .....	85
2. <i>Distinguer les « modèles de fondation » des systèmes d'intelligence artificielle</i> .....	88
a) La notion de « modèle de fondation » et ses ambiguïtés .....	88
b) Le « <i>fine-tuning</i> » : passer du modèle de fondation aux applications.....	89
c) Pourquoi les IA deviennent-elles « <i>woke</i> » ? .....	93
3. <i>D'autres innovations récentes que la technologie Transformer en matière d'IA générative</i> .....	96
a) La Génération par diffusion : une innovation de pointe pour la création de contenus visuels.....	97
b) Générer du contenu avec des réseaux de neurones convolutifs : les auto-encodeurs variationnels (VAE).....	98
c) Les réseaux génératifs antagonistes (GAN) .....	100
d) La dernière innovation de 2024 : l'architecture Mamba et son association avec un <i>Transformer</i> dans le modèle Jamba.....	102
<b>B. LES GRANDES QUESTIONS TECHNOLOGIQUES ET LES POSSIBLES ÉVOLUTIONS À VENIR .....</b>	<b>104</b>
1. <i>Les problématiques technologiques de l'intelligence artificielle</i> .....	104
a) L'IA « boîte noire » : le double défi de l'explicabilité.....	104
b) Des biais à plusieurs niveaux : données réelles ou synthétiques ainsi que choix de programmation .....	105
c) Les erreurs graves ou « hallucinations » .....	106
2. <i>Les tendances de la recherche et les principales perspectives technologiques</i> .....	107
a) Moins halluciner : la « <i>Retrieval Augmented Generation</i> » (RAG) par les « <i>Retrieval Augmented Transformers</i> » (RAT) .....	107
b) Manipuler en entrée et en sortie des données de nature variée : les IA génératives multimodales .....	109
c) Faire des systèmes d'IA des interfaces devenant la principale plateforme d'accès aux services numériques .....	110
d) Aller vers plus d'autonomie : le défi de l'agentivité.....	111
e) Faire plus avec moins : vers une IA frugale et efficace .....	113
f) L'exemple de la méthode « <i>Mixture of Experts</i> » (MoE) .....	114
3. <i>Synthèse et articulations entre les modèles d'IA</i> .....	116
a) Des technologies enchâssées et souvent conjuguées .....	116
b) Les Arbres de pensées ou <i>Trees of Thought</i> (ToT) : l'IA « symboliconnexionniste ».....	117
c) La fécondité des hybridations IA symboliques/IA connexionnistes, notamment pour doter ces systèmes d'une représentation du monde réel .....	119
d) La variété des domaines de l'intelligence artificielle .....	122
<b>III. LE GRAND MARCHÉ DE L'IA : UNE CHAÎNE DE VALEUR ÉTENDUE DES MATIÈRES PREMIÈRES AUX CONSOMMATEURS .....</b>	<b>123</b>
<b>A. DU SILICIUM AUX APPLICATIONS : LA CHAÎNE DE VALEUR COMPLEXE DE L'INTELLIGENCE ARTIFICIELLE.....</b>	<b>123</b>
1. <i>Un schéma souvent simplifié en quatre étapes</i> .....	123

---

2. La complexité de la dizaine d'étapes de la chaîne.....	124
a) L'amont : du silicium aux microprocesseurs .....	124
b) La couche multiforme des infrastructures .....	128
c) L'étape de définition des modèles : de leur architecture au réglage fin .....	129
d) L'aval : les applications.....	129
<b>B. LA MULTIPLICATION DES MODÈLES PROPOSÉS AU GRAND PUBLIC .....</b>	<b>133</b>
1. Une domination nette de l'IA générative par la Big Tech américaine .....	133
2. Le marché très évolutif des grands modèles de langage (LLM) .....	136
3. Les autres modèles d'IA générative disponibles sur le marché .....	138
4. Des modèles plus ou moins ouverts : la question de l'open source .....	139
<b>DEUXIÈME PARTIE : LES ENJEUX DE L'INTELLIGENCE ARTIFICIELLE .....</b>	<b>141</b>
<b>I. DES PROBLÉMATIQUES POLITIQUES PRÉOCCUPANTES, SURTOUT À L'HEURE DE L'IA GÉNÉRATIVE.....</b>	<b>141</b>
<b>A. UNE SOUVERAINETÉ DE PLUS EN PLUS MENACÉE : LES ENJEUX GÉOPOLITIQUES DE LA CHAÎNE DE VALEUR DE L'IA .....</b>	<b>141</b>
1. Tirer les conséquences du fait que l'IA est d'abord américaine .....	142
2. La France et l'Europe, « colonies numériques », ont-elles déjà complètement perdu la course mondiale à l'IA ?.....	144
3. Une souveraineté française qui peut encore être défendue en tant que puissance intermédiaire .....	148
<b>B. DES RISQUES DE MANIPULATIONS POLITIQUES VOIRE DE DÉSTABILISATION .....</b>	<b>152</b>
1. Désinformation au carré .....	152
2. Hypertrucages réalistes (« deepfakes »).....	153
3. Cybersécurité et risque d'attaques à grande échelle .....	153
a) La typologie des usages malveillants de l'IA générative .....	154
b) Comment répondre à ces nouveaux risques en matière de sécurité face à l'essor de l'IA générative aujourd'hui ?.....	155
<b>C. LA SINGULARITÉ ET LE RISQUE EXISTENTIEL : DE L'IAG AUX SCÉNARIOS À LA TERMINATOR.....</b>	<b>157</b>
1. Que sont l'intelligence artificielle générale (IAG) et la Singularité ? .....	157
a) La perspective possible mais pas certaine de l'intelligence artificielle générale rend la singularité et le risque existentiel encore moins probables.....	157
b) L'hypothèse des lois d'échelle ou <i>scaling laws</i> .....	159
c) Quelques jalons sur les évolutions en cours : des capacités croissantes et la probabilité de plateaux.....	160
2. L'espace des positions face à l'intelligence artificielle générale et au risque existentiel .....	161
a) Les techno-pessimistes : l'IA est un risque existentiel et l'alternative un moratoire ou le transhumanisme .....	162
b) Les pessimistes modérés : l'IA est une menace, les mesures de prévention une solution .....	167
c) Les techno-optimistes : l'absence de menace, la poursuite du développement tranquille de l'IA avec un encadrement pragmatique.....	168

---

<b>II. DES EFFETS GLOBALEMENT POSITIFS POUR LA SOCIÉTÉ MALGRÉ DES IMPACTS ÉCONOMIQUES CONTRASTÉS .....</b>	<b>168</b>
<b>A. DES AMÉLIORATIONS POUR LE BIEN-ÊTRE ET LA SANTÉ GRÂCE À L'INTELLIGENCE ARTIFICIELLE .....</b>	<b>169</b>
1. <i>De nombreux outils pour la vie quotidienne .....</i>	<i>169</i>
2. <i>De meilleurs systèmes de soins.....</i>	<i>172</i>
3. <i>Des effets cognitifs à surveiller dans une économie de l'attention.....</i>	<i>174</i>
<b>B. UN IMPACT SUR LA CROISSANCE INCERTAIN.....</b>	<b>175</b>
1. <i>Des études divergentes, très optimistes ou très mesurées .....</i>	<i>175</i>
2. <i>Le retour du « paradoxe de Solow » ? .....</i>	<i>176</i>
3. <i>Des besoins d'énergie considérables et croissants laissant planer la menace de risques         environnementaux.....</i>	<i>177</i>
<b>C. UNE TRANSFORMATION CONSIDÉRABLE DU MARCHÉ DU TRAVAIL .....</b>	<b>179</b>
1. <i>Les études quantitatives sur la base des tâches et des compétences .....</i>	<i>180</i>
2. <i>Les études qualitatives sur la base d'observations et d'entretiens .....</i>	<i>182</i>
3. <i>Des effets encore incertains mais qui appellent un dialogue social.....</i>	<i>184</i>
<b>III. DES DÉFIS SANS PRÉCÉDENT POUR LA SPHÈRE CULTURELLE ET LE MONDE SCIENTIFIQUE .....</b>	<b>185</b>
<b>A. DE LA DOMINATION CULTURELLE ANGLO-SAXONNE À L'UNIFORMISATION COGNITIVE .....</b>	<b>185</b>
1. <i>L'IA est américaine et pas que sur un plan économique .....</i>	<i>185</i>
a) <i>Une domination par les données .....</i>	<i>185</i>
b) <i>Une domination par la culture et la langue .....</i>	<i>186</i>
c) <i>Ne pas donner notre langue à ChatGPT ! Pour la diversité linguistique et             culturelle .....</i>	<i>187</i>
2. <i>Le danger de l'uniformisation cognitive plus encore que culturelle .....</i>	<i>188</i>
a) <i>De nouvelles structures mentales .....</i>	<i>188</i>
b) <i>Vers une civilisation de poissons rouges ? .....</i>	<i>192</i>
c) <i>Éduquer à l'IA, éduquer par l'IA .....</i>	<i>193</i>
<b>B. LA CRÉATION AU DÉFI DE L'IMPACT DE L'INTELLIGENCE ARTIFICIELLE SUR LA PROPRIÉTÉ INTELLECTUELLE ET LES DROITS D'AUTEUR .....</b>	<b>195</b>
1. <i>Des régimes de propriété intellectuelle fragilisés .....</i>	<i>196</i>
2. <i>Des risques contentieux .....</i>	<i>199</i>
3. <i>Le modèle économique de la création artistique par l'IA .....</i>	<i>199</i>
4. <i>L'avenir de la création artistique .....</i>	<i>200</i>
<b>C. DES BÉNÉFICES CONSIDÉRABLES POUR LA RECHERCHE .....</b>	<b>201</b>
1. <i>La fertilisation des autres disciplines scientifiques par l'IA.....</i>	<i>201</i>
2. <i>Les cas emblématiques de l'analyse du repliement des protéines en 2018         et de la génomique en 2024 .....</i>	<i>203</i>
3. <i>Les jumeaux numériques et le perfectionnement des simulations .....</i>	<i>205</i>
4. <i>Adapter nos politiques de recherche aux perspectives ouvertes par l'IA.....</i>	<i>206</i>

---

<b>TROISIÈME PARTIE : LA GOUVERNANCE ET LA RÉGULATION DE L'INTELLIGENCE ARTIFICIELLE.....</b>	<b>207</b>
<b>I. LES DISPOSITIFS NATIONAUX OU RÉGIONAUX.....</b>	<b>209</b>
<b>A. UNE POLITIQUE FRANÇAISE DE L'INTELLIGENCE ARTIFICIELLE EN DEMI-TEINTE .....</b>	<b>209</b>
1. <i>La stratégie nationale en faveur de l'IA depuis 2017 : un retard à l'allumage.....</i>	<i>209</i>
2. <i>Le bilan critique de la stratégie et son évaluation par la Cour des comptes .....</i>	<i>213</i>
3. <i>Les perspectives de relance de la politique nationale de l'IA .....</i>	<i>215</i>
<b>B. D'AUTRES DISPOSITIFS NATIONAUX DANS L'UNION EUROPÉENNE .....</b>	<b>219</b>
1. <i>L'Allemagne : le pays le plus proche du nôtre .....</i>	<i>219</i>
2. <i>L'Italie : une stratégie de soutien et de vigilance.....</i>	<i>220</i>
3. <i>L'Espagne : un cadre complet avec un riche volet culturel .....</i>	<i>222</i>
4. <i>Les Pays-Bas : une coalition public-privé efficace et une régulation précoce .....</i>	<i>227</i>
5. <i>L'Estonie : un État numérique préoccupé par la sécurité de l'IA .....</i>	<i>230</i>
6. <i>La Finlande : une stratégie tournée vers l'appropriation de l'IA et l'éducation .....</i>	<i>231</i>
<b>C. LA GOUVERNANCE EUROPÉENNE DE L'INTELLIGENCE ARTIFICIELLE .....</b>	<b>236</b>
1. <i>Le travail préparatoire conduit par les institutions européennes entre 2018 et 2020.....</i>	<i>238</i>
a) <i>La communication de la Commission européenne d'avril 2018 .....</i>	<i>238</i>
b) <i>Le groupe d'experts de haut niveau sur l'IA .....</i>	<i>238</i>
c) <i>Le Livre blanc de la Commission européenne sur l'IA .....</i>	<i>240</i>
2. <i>De la proposition de règlement du 21 avril 2021 à la juxtaposition de deux dispositifs à la suite des amendements adoptés.....</i>	<i>241</i>
a) <i>Le volet issu du projet initial d'AI Act en 2021 : une régulation des usages selon leurs risques plutôt qu'une régulation de la technologie elle-même.....</i>	<i>242</i>
b) <i>Le volet ajouté par les co-législateurs : la régulation des modèles de fondation assortie d'un régime spécifique pour les modèles les plus puissants, dits « à risque systémique » .....</i>	<i>243</i>
c) <i>Les autres aspects de l'AI Act : une polysynodie institutionnelle, une portée extraterritoriale, un calendrier très complexe et une normalisation désinvestie .....</i>	<i>247</i>
3. <i>Une gouvernance européenne de l'IA à compléter .....</i>	<i>253</i>
a) <i>Mobiliser les entreprises et élaborer de la Soft Law : l'AI Pact et les bonnes pratiques.....</i>	<i>253</i>
b) <i>Deux projets de directive complémentaires à l'AI Act : l'intelligence artificielle comme produit sur le marché unique .....</i>	<i>254</i>
c) <i>Le soutien européen à la recherche et à l'innovation en IA au-delà du dispositif EuroHPC.....</i>	<i>256</i>
<b>D. PANORAMA D'AUTRES RÉGULATIONS NATIONALES DANS LE RESTE DU MONDE .....</b>	<b>258</b>
1. <i>Aux États-Unis, une régulation inachevée .....</i>	<i>258</i>
a) <i>Au niveau fédéral : de la stratégie de 2016 à l'Executive Order présidentiel d'octobre 2023 .....</i>	<i>258</i>
b) <i>Au niveau des États : plusieurs projets à commencer par celui de la Californie .....</i>	<i>261</i>
2. <i>En Chine, un développement rapide et centralisé de l'IA et de sa régulation .....</i>	<i>262</i>
a) <i>La politique chinoise en faveur d'une IA maîtrisée depuis 2017 .....</i>	<i>262</i>
b) <i>Un encadrement strict et assez exhaustif des IA génératives .....</i>	<i>264</i>
c) <i>Un rival sérieux des États-Unis intéressé par l'AI Act.....</i>	<i>266</i>



---

3. Quelques autres initiatives intéressantes.....	267
a) Le Canada.....	267
b) Le Brésil.....	268
c) Le Japon.....	269
d) La Corée du Sud.....	270
e) L'Inde.....	270
f) Le Royaume-Uni.....	273
g) Israël.....	278
h) L'Arabie Saoudite.....	279
i) Les Émirats arabes unis (EAU).....	280
<b>II. UNE DIZAINE DE PROJETS DE GOUVERNANCE INTERNATIONALE NON COORDONNÉS.....</b>	<b>280</b>
<b>A. LA RÉFLEXION TECHNIQUE LA PLUS ABOUTIE : LES PRINCIPES, RECOMMANDATIONS ET MÉTRIQUES DE L'OCDE.....</b>	<b>281</b>
1. <i>Les principes, les recommandations et la classification des systèmes</i> .....	281
2. <i>La méthodologie des métriques</i> .....	284
3. <i>L'Observatoire des politiques publiques de l'IA</i> .....	286
<b>B. LE CADRE MULTILATÉRAL EN CONSTRUCTION.....</b>	<b>287</b>
1. <i>La contribution multiforme mais encore inachevée de l'Organisation des Nations unies     (ONU).....</i>	287
2. <i>Le travail spécifique de l'Organisation des Nations unies pour l'éducation, la science et     la culture (Unesco).....</i>	290
3. <i>Les Principes pour l'IA du G20.....</i>	292
<b>C. LES FORUMS FERMÉS DU G7, DU PARTENARIAT MONDIAL SUR L'INTELLIGENCE ARTIFICIELLE ET DU CONSEIL DU COMMERCE ET DES TECHNOLOGIES UE-ÉTATS-UNIS.....</b>	<b>293</b>
1. <i>Le G7 et son « processus d'Hiroshima ».....</i>	293
2. <i>L'expertise du Partenariat mondial sur l'intelligence artificielle (PMIA) ou Global     partnership on artificial intelligence (GPAI).....</i>	295
3. <i>Le Conseil du commerce et des technologies (CCT) UE-États-Unis.....</i>	295
<b>D. LES INITIATIVES NON OCCIDENTALES.....</b>	<b>296</b>
1. <i>La Proposition chinoise de gouvernance mondiale de l'IA.....</i>	296
2. <i>Le travail des BRICS.....</i>	297
3. <i>La stratégie de l'Union africaine.....</i>	297
<b>E. LES AUTRES PROJETS DE GOUVERNANCE MONDIALE : CONSEIL DE L'EUROPE, FORUM ÉCONOMIQUE MONDIAL, INITIATIVES ÉMANANT DU SECTEUR PRIVÉ.....</b>	<b>299</b>
1. <i>La Convention-cadre sur l'IA du Conseil de l'Europe.....</i>	299
2. <i>L'Alliance pour la gouvernance de l'IA proposée par le Forum économique mondial.....</i>	300
3. <i>Des principes et bonnes pratiques proposés par les entreprises au Partnership on AI     lancé en 2016 par sept géants de l'IA.....</i>	300
4. <i>Le Forum sur les modèles de pointe ou Frontier Model Forum et les autres initiatives.....</i>	301
<b>F. LE CADRE EN CONSTRUCTION DES SOMMETS POUR LA SÉCURITÉ DE L'INTELLIGENCE ARTIFICIELLE ET DES INSTITUTS DE SÉCURITÉ DE L'IA.....</b>	<b>303</b>
1. <i>Un réseau international d'agences pour la sécurité de l'intelligence artificielle.....</i>	303
2. <i>Du sommet de Bletchley Park au rapport de Yoshua Bengio.....</i>	304
3. <i>Un sommet en France les 10 et 11 février 2025.....</i>	307

---

<b>III. LES PROPOSITIONS DE L'OFFICE .....</b>	<b>308</b>
<b>A. LES PROPOSITIONS À SOUTENIR DANS LE CADRE DU FUTUR SOMMET DE L'IA .....</b>	<b>308</b>
1. <i>Faire reconnaître le principe d'une approche transversale de l'IA et renoncer à l'approche exclusivement tournée vers les risques .....</i>	<i>308</i>
2. <i>Proposer de placer la gouvernance mondiale de l'IA sous l'égide d'une seule organisation internationale .....</i>	<i>309</i>
3. <i>Initier le cadre d'une régulation globale et multidimensionnelle de l'IA en s'inspirant des travaux de l'OCDE et de l'UE .....</i>	<i>310</i>
4. <i>Annoncer un programme européen de coopération en IA, associant plusieurs pays dont au moins la France, l'Allemagne, les Pays-Bas, l'Italie et l'Espagne .....</i>	<i>311</i>
5. <i>Associer le Parlement à l'organisation du sommet .....</i>	<i>311</i>
<b>B. LES PROPOSITIONS VISANT À FONDER UNE VÉRITABLE POLITIQUE NATIONALE DE L'IA .....</b>	<b>312</b>
1. <i>Développer une filière française ou européenne autonome sur l'ensemble de la chaîne de valeur de l'intelligence artificielle .....</i>	<i>312</i>
2. <i>Mettre en place une politique publique de l'IA avec des objectifs, des moyens et des outils de suivi et d'évaluation .....</i>	<i>313</i>
3. <i>Organiser le pilotage stratégique de la politique publique de l'intelligence artificielle au plus haut niveau .....</i>	<i>314</i>
4. <i>Former les élèves de l'école à l'Université, former les actifs et former le grand public à l'IA .....</i>	<i>314</i>
5. <i>Accompagner le déploiement de ces technologies dans le monde du travail et la société, notamment par la formation permanente .....</i>	<i>315</i>
6. <i>Lancer un grand dialogue social autour de l'intelligence artificielle et de ses enjeux .....</i>	<i>315</i>
7. <i>Mobiliser et animer l'écosystème français de l'IA .....</i>	<i>315</i>
8. <i>Reconduire le programme « Confiance.ai » ou mettre en place un projet équivalent .....</i>	<i>316</i>
9. <i>Soutenir la recherche publique en intelligence artificielle selon des critères de transversalité et de diversification des technologies .....</i>	<i>318</i>
10. <i>Relever le défi de la normalisation en matière d'intelligence artificielle .....</i>	<i>319</i>
11. <i>S'assurer du contrôle souverain des données issues de la culture française et des cultures francophones et créer des jeux de données autour des cultures francophones .....</i>	<i>319</i>
12. <i>Préparer une réforme des droits de propriété intellectuelle dont le droit d'auteur pour les adapter aux usages de l'IA générative .....</i>	<i>320</i>
13. <i>Confier à l'OPECST le suivi et l'évaluation régulière de la politique publique conduite par le Gouvernement en la matière .....</i>	<i>321</i>
<b>LISTE DES PERSONNES ENTENDUES .....</b>	<b>323</b>
<b>ANNEXES .....</b>	<b>337</b>
<b>LETTRE DE SAISINE DE L'OFFICE PAR LE BUREAU DE L'ASSEMBLÉE NATIONALE .....</b>	<b>337</b>
<b>LETTRE DE SAISINE DE L'OFFICE PAR LE BUREAU DU SÉNAT .....</b>	<b>339</b>



## **L'ESSENTIEL**

*ChatGPT, et après ?  
Bilan et perspectives de l'intelligence artificielle*



## AVANT-PROPOS

Le présent rapport répond à une **saisine de l'Office par les bureaux de l'Assemblée nationale et du Sénat en juillet 2023**, ce qui est exceptionnel<sup>1</sup>. Il s'agissait de manifester un soutien au plus haut niveau des deux assemblées aux travaux de l'OPECST à l'occasion de son 40<sup>e</sup> anniversaire. Quatre rapporteurs ont été désignés le 26 octobre 2023 dans une composition doublement paritaire : deux députés et deux sénateurs, deux femmes et deux hommes. Ils représentaient, de plus, la diversité de l'éventail politique national.

Le présent rapport n'est pas le premier travail de l'Office sur le thème de l'intelligence artificielle : en 2016 et 2017, nos anciens collègues **Claude de Ganay et Dominique Gillot** ont ainsi conduit des investigations très poussées sur ce sujet d'intérêt capital. Le **rapport pionnier** qu'ils ont rendu au nom de l'Office<sup>2</sup> a marqué l'histoire des analyses relatives à ces technologies et il y sera parfois renvoyé dans le cadre des développements du présent rapport.

Ce rapport exigeait, dans une modernité notable face à l'évolution rapide de ces techniques et du vocabulaire qui leur est associé, que l'ensemble de ces technologies soient en effet « *maîtrisées, utiles et fassent l'objet d'usages conformes à nos valeurs humanistes* »<sup>3</sup> : même si le fait de concevoir des machines réellement intelligentes restait surtout selon eux une préoccupation de long terme, la « *question de savoir comment aligner les valeurs de ces machines avec les valeurs morales humaines* mérite d'être posée dès aujourd'hui » disaient très sagement nos collègues.

Leur rapport a même apporté quelques réponses solides et argumentées à la question tout en précisant que ses préconisations devraient être « *remises en débat au fur et à mesure des nouvelles découvertes scientifiques, de leurs transferts et de leurs usages* ». Les deux rapporteurs tenaient surtout à ce que « *le point d'équilibre qu'ils ont cherché à atteindre dans leur rapport puisse évoluer, en fonction des évolutions du contexte résultant du jeu de ces variables* »<sup>4</sup>.

---

<sup>1</sup> Cf. les deux lettres de saisine signées par la Présidente de l'Assemblée nationale, Yaël Braun-Pivet, et le Président du Sénat, Gérard Larcher, annexées à la fin du présent rapport.

<sup>2</sup> Cf. le rapport de l'OPECST de Claude de Ganay et Dominique Gillot, mars 2017, « Pour une intelligence artificielle maîtrisée, utile et démystifiée », disponible sur le site du Sénat : <https://www.senat.fr/notice-rapport/2016/r16-464-1-notice.html> ainsi que sur celui de l'Assemblée : [https://www.assemblee-nationale.fr/dyn/14/dossiers/intelligence\\_artificielle\\_maitrisee\\_utile](https://www.assemblee-nationale.fr/dyn/14/dossiers/intelligence_artificielle_maitrisee_utile)

<sup>3</sup> Cf. le rapport de l'OPECST précité p.29.

<sup>4</sup> Cf. la conclusion générale du rapport de l'OPECST précité p.207.

Parmi les 15 propositions de ce rapport en mars 2017 figurait, par exemple, le fait de :

- favoriser des algorithmes et des robots sûrs, transparents et justes ;
- prévoir une charte de l'intelligence artificielle et de la robotique ;
- confier à un institut national de l'éthique de l'intelligence artificielle et de la robotique un rôle d'animation du débat public sur les principes éthiques qui doivent encadrer ces technologies ;
- encourager la constitution de champions européens en intelligence artificielle et en robotique ;
- redonner en IA une place essentielle à la recherche fondamentale et revaloriser la place de la recherche publique par rapport à la recherche privée ;
- mobiliser la communauté française de la recherche en intelligence artificielle ;
- accompagner les transformations du marché du travail en menant une politique de formation continue ambitieuse visant à s'adapter aux exigences de requalification et d'amélioration des compétences ;
- élargir l'offre de cursus et de modules de formation aux technologies d'intelligence artificielle dans l'enseignement supérieur et créer, en France, au moins un pôle d'excellence international et interdisciplinaire en IA ;
- former à l'éthique de l'intelligence artificielle et de la robotique dans les cursus spécialisés de l'enseignement supérieur :
- former à l'informatique dans l'enseignement primaire et secondaire ;
- former et sensibiliser le grand public à l'intelligence artificielle et à ses conséquences pratiques ;
- assurer une meilleure prise en compte de la diversité et de la place des femmes dans la recherche en intelligence artificielle.

Ces propositions, quand elles n'ont pas encore été traduites dans les faits, restent d'actualité.

Le dernier point a, par exemple, justifié l'organisation au Sénat d'une **audition sur la place des femmes dans l'intelligence artificielle**, le 7 mars 2024, au cours de laquelle on s'est demandé pourquoi on dénombrait si peu de femmes dans les métiers de l'IA<sup>1</sup>.

---

<sup>1</sup> À l'occasion de la Journée internationale des droits des femmes, la délégation sénatoriale aux droits des femmes, la délégation sénatoriale à la prospective et l'OPECST ont organisé au Sénat, jeudi 7 mars 2024, trois tables rondes réunissant des experts et expertes de l'IA, en particulier des femmes engagées et aux compétences reconnues, qui se sont penchées sur trois questions : pourquoi si peu de femmes dans les métiers de l'IA ? L'IA est-elle sexiste ? Comment faire de l'IA un atout pour l'égalité femmes-hommes ? Cf. le rapport d'information « Femmes et IA : briser les codes », n° 607 (2023-2024) : <https://www.senat.fr/notice-rapport/2023/r23-607-notice.html>

Le présent rapport répond à une saisine assez précise puisqu'il s'agit de **traiter des « nouveaux développements » de l'IA**, ceux intervenus en particulier depuis le premier rapport de 2017.

Et de nouveaux développements, il y a eu. Les innovations s'accroissent trimestre après trimestre. Depuis quelques années, on observe en effet que **les cycles de révolution en IA sont en moyenne de trois mois**.

En septembre 2017, quelques mois après la publication du rapport de l'OPECST, une nouvelle architecture d'intelligence artificielle était ainsi proposée par des chercheurs de Google : les *Transformers*. Ces algorithmes, comprenant des **centaines de milliards voire des milliers de milliards de paramètres** sont devenus des **systèmes d'IA générative accessibles au grand public** et ont notamment abouti à **des applications populaires telles que ChatGPT**, lancée en novembre 2022 et à laquelle le titre du rapport fait un clin d'œil appuyé en se demandant ce qui se joue au-delà de cette application en elle-même.

Ces technologies, capables de générer du texte, des images ou d'autres contenus en réponse à des commandes en langage naturel (ou *prompts* en anglais), représentent une **nouvelle étape significative dans la longue histoire de l'intelligence artificielle** et posent de **nouvelles questions** à nos sociétés.

Si c'est au milieu du XX<sup>e</sup> siècle, en même temps que l'informatique, que se développe et qu'apparaît formellement la notion d'intelligence artificielle, il ne fait aucun doute que **les sept dernières années ont marqué une étape fondamentale dans les progrès de ces technologies**. Ce constat valide près de 80 ans après le propos de Georges Bernanos placé en ouverture liminaire du présent rapport, qui nous rappelle avec un ton mi-positiviste, mi-nostalgique, que « *le progrès n'est plus dans l'homme, il est dans la technique, dans le perfectionnement des méthodes capables de permettre chaque jour une utilisation plus efficace du matériel humain* »<sup>1</sup>. À l'heure des IA génératives, cette phrase nous a semblé faire écho à la riche actualité de ces systèmes de plus en plus présents dans nos vies et dans nos sociétés.

Après le premier rapport de l'Office sur l'IA rendu public en mars 2017, il est nécessaire de remettre l'ouvrage sur le métier pour analyser **le fonctionnement de ces nouvelles IA et leurs défis**, sans chercher à récapituler tous leurs domaines d'application et leurs cas d'usage, qui sont innombrables.

---

<sup>1</sup> Cf. Georges Bernanos, *La France et les Robots*, Éditions de la France libre, 1946, page 11.



Cette **nécessité apparaît d'autant plus grande que très peu de travaux parlementaires sont consacrés à l'intelligence artificielle**. Outre le rapport de l'Office précité, on peut relever un ensemble de travaux de la **délégation à la prospective du Sénat**<sup>1</sup> et le rapport d'information de la **commission des lois de l'Assemblée nationale** déposé le 14 février 2024 en conclusion des travaux de sa mission d'information sur les défis de l'intelligence artificielle générative en matière de protection des données personnelles et d'utilisation du contenu généré dont les rapporteurs étaient nos collègues députés Philippe Pradal et Stéphane Rambaud<sup>2</sup>. Deux autres rapports établis au nom de la **commission des affaires européennes du Sénat** et proposant des résolutions en réponse à la stratégie européenne pour l'intelligence artificielle de la Commission européenne en 2019<sup>3</sup> et à la proposition de règlement européen sur l'intelligence artificielle en 2023<sup>4</sup> peuvent aussi être mentionnés. En outre, la **commission des lois du Sénat** a mis en place une mission d'information sur le point de rendre un rapport sur l'impact de l'IA sur les professions du droit<sup>5</sup>.

Vos rapporteurs ont voulu fournir des éléments de réponse à plusieurs interrogations. **Comment fonctionnent ces technologies, en particulier les IA génératives ? Quels avantages et quels inconvénients présentent-elles ? Quels biais persistent dans l'usage des données et dans les programmations ? Posent-elles des difficultés en termes de souveraineté, de sécurité ou de régulation ? Quelles gouvernances nationales, européennes**

---

<sup>1</sup> En effet, la *délégation à la prospective du Sénat* a travaillé sur cinq cas d'usage de l'intelligence artificielle dans les services publics. Ainsi, nos collègues sénateurs ont d'ores et déjà publié trois rapports. Sylvio Vermeillet et Didier Rambaud ont écrit le premier rapport : « Impôts, prestations sociales et lutte contre la fraude » ; Anne Ventalon et Christian Redon-Sarrazy, le deuxième rapport : « IA et santé » ; et Christian Bruyen et Bernard Fialaire, le troisième rapport : « IA et éducation ». Deux autres rapports sont prévus : « IA et environnement » ainsi que « IA, territoires et proximité ». Cf. les liens vers les trois rapports déjà publiés : « L'IA et l'avenir du service public, rapport thématique #1 : Impôts, prestations sociales et lutte contre la fraude » <https://www.senat.fr/rap/r23-491/r23-491-syn.pdf> ; « L'IA et l'avenir du service public, rapport thématique #2 : IA et santé » : <https://www.senat.fr/rap/r23-611/r23-611-syn.pdf> ; « L'IA et l'avenir du service public, rapport thématique #3 : IA et éducation » : <https://www.senat.fr/rap/r24-101/r24-101-syn.pdf>

<sup>2</sup> Cf. le rapport de la commission des lois de l'Assemblée nationale en conclusion des travaux de sa mission d'information sur les défis de l'intelligence artificielle générative en matière de protection des données personnelles et d'utilisation du contenu généré, rapport d'information n° 2207, 14 février 2024, 16<sup>e</sup> législature : [https://www.assemblee-nationale.fr/dyn/16/rapports/cion\\_lois/116b2207\\_rapport-information](https://www.assemblee-nationale.fr/dyn/16/rapports/cion_lois/116b2207_rapport-information)

<sup>3</sup> Cf. la proposition de résolution du 31 janvier 2019 d'André Gattolin, Claude Kern, Cyril Pellevat et Pierre Ouzoulias au nom de la commission des affaires européennes du Sénat sur les investissements dans l'intelligence artificielle en Europe : « Intelligence artificielle : l'urgence d'une ambition européenne » par : <https://www.senat.fr/rap/r18-279/r18-279.html>

<sup>4</sup> Cf. la proposition de résolution du 30 mars 2023 d'André Gattolin, Catherine Morin-Desailly, Cyril Pellevat et Elsa Schalck au nom de la commission des affaires européennes du Sénat relative à la proposition de règlement établissant des règles harmonisées concernant l'intelligence artificielle et modifiant certains actes législatifs de l'Union : <https://www.senat.fr/leg/ppr22-484.html>

<sup>5</sup> Cf. la présentation de cette mission de la commission des lois du Sénat : <https://www.senat.fr/travaux-parlementaires/commissions/commission-des-lois/intelligence-artificielle-et-professions-du-droit.html>

voire internationales sont mises en place ? Faut-il faire évoluer ces cadres de régulation ? Si oui, dans quelles directions ? Comment **la France** doit-elle se positionner par rapport à ces enjeux de gouvernance et par rapport aux évolutions technologiques et économiques en cours ? Et quelles **perspectives** la **recherche** permet-elle de dessiner pour le futur ? Va-t-on vers **l'intelligence artificielle générale** (IAG), voire vers une IA qui nous serait même supérieure ? Dans cette perspective, appelée singularité, devons-nous considérer que nous aurons à faire face à un **risque existentiel** ?



## PREMIÈRE PARTIE COMPRENDRE LES TECHNOLOGIES D'INTELLIGENCE ARTIFICIELLE

Vos rapporteurs ont jugé **indispensable**, pour traiter des nouveaux développements de l'IA depuis 2017, **de retracer tout d'abord l'histoire de ces technologies**, en mettant l'accent sur leur fonctionnement ainsi que sur le contenu et les contours du concept d'IA.

Avant même d'évoquer l'histoire de l'intelligence artificielle, il est proposé de **reconstruire brièvement sa préhistoire**.

Cette **archéologie du savoir** – pour reprendre la démarche de Michel Foucault – est utile en ce qu'elle permet d'identifier l'ensemble des représentations et des enjeux qui relèvent de ces technologies, en particulier d'un point de vue culturel. Ces perceptions traditionnelles de l'IA, qui peuvent être certes différentes de la réalité effective de ces technologies, influencent du reste la façon dont ces dernières sont conçues.

Ainsi que l'exprimait Marie Curie « *dans la vie, rien n'est à craindre, tout est à comprendre* », alors plutôt que de s'épouvanter en débattant des risques que ferait courir l'IA, il est en effet primordial en suivant l'invite de Marie Curie (un *prompt* dirait-on en anglais) de commencer par chercher à **comprendre ce que recouvre exactement la notion d'intelligence artificielle et d'appréhender avec rigueur le fonctionnement de ces technologies complexes**.

C'est là tout l'objet de cette première partie.

### I. HISTOIRE DE LA NOTION D'INTELLIGENCE ARTIFICIELLE

#### A. LA PRÉHISTOIRE DE L'INTELLIGENCE ARTIFICIELLE ET SES REPRÉSENTATIONS CULTURELLES

En 2021, a été mis en place un séminaire pluridisciplinaire autour de l'histoire culturelle de l'intelligence artificielle<sup>1</sup>, quelques années après le rapport précité de l'OPECST qui, en 2017, jugeait **nécessaire, pour mieux comprendre les IA modernes, de les replacer dans le contexte des incarnations de l'intelligence artificielle qui ont jalonné notre longue histoire**, qu'il s'agisse de mythes anciens ou de projets imaginés par des écrivains et des scientifiques<sup>2</sup>.

---

<sup>1</sup> Depuis trois ans, le projet CulturIA notamment à travers son séminaire « Pour une histoire de l'intelligence artificielle (IA) » aborde les aspects scientifiques, technologiques et culturels de l'histoire de l'intelligence artificielle et tend à diffuser les recherches en cours, dans l'attente d'une future publication. Grâce aux différentes spécialités convoquées, cette histoire vise à comprendre les multiples enjeux du développement de l'IA. Cf. la présentation du séminaire 2024/2025 : <https://doi.org/10.58079/12b1f>

<sup>2</sup> Cf. le rapport de l'OPECST précité « Pour une intelligence artificielle maîtrisée, utile et démystifiée », p. 31.

Le Conseil d'État explique dans une étude sur l'IA<sup>1</sup> que « *la très forte charge symbolique de l'expression intelligence artificielle, ainsi que l'absence de définition partagée et de consensus sur le contenu même de la notion, contribuent puissamment à la confusion et compliquent l'examen rationnel des avantages et des inconvénients de ce qui est, d'abord et avant tout, un ensemble d'outils numériques au service de l'humain* ». Démystifier l'IA impose de déconstruire les apports culturels ayant produit la cristallisation de cette très forte charge symbolique.

## 1. De la mythologie antique aux machines à calculer

Il est fait mention dès l'Égypte ancienne de statues articulées, animées par la vapeur et par le feu, qui hochaient la tête et bougeaient les bras, véritables ancêtres des automates. **Homère** a décrit dans *L'Iliade* des servantes en or douées de raison : « *Fabriquées par Héphaïstos, le dieu forgeron, elles ont, selon le poète, voix et force ; elles vaquent aux occupations quotidiennes à la perfection, car les immortels leur ont appris à travailler. Ce sont donc des robots, au sens étymologique de travailleurs artificiels* » ainsi que le rappelle Jean-Gabriel Ganascia, entendu par vos rapporteurs. **Ovide** dans ses « *Métamorphoses* » crée la figure de Galatée, statue d'ivoire sculptée par Pygmalion et à laquelle Vénus, déesse de l'amour, accepte de donner vie. **La Bible**, par le Psaume 139:16, a fondé le mythe du Golem, cette créature d'argile humanoïde que l'on retrouve souvent dans la tradition cabalistique juive.

Parallèlement à ces développements mythologiques, la science tâtonne pendant des siècles mais pose tout de même des jalons concernant la future intelligence artificielle. Ces technologies d'IA sont filles des mathématiques et se basent sur des algorithmes<sup>2</sup>.

---

<sup>1</sup> Cf. l'étude du Conseil d'État du 31 mars 2022, « *Intelligence artificielle et action publique : construire la confiance, servir la performance* » : <https://www.conseil-etat.fr/publications-colloques/etudes/intelligence-artificielle-et-action-publique-construire-la-confiance-servir-la-performance>

<sup>2</sup> Il s'agit d'un ensemble fini et précis de séquences d'opérations ou d'instructions permettant, à l'aide d'entrées, de résoudre un problème ou d'obtenir un résultat, ces sorties étant réalisées selon un certain rendement. Donald Knuth, pionnier de l'algorithmique moderne (*The Art of Computer Programming*), a identifié les cinq propriétés suivantes comme étant les prérequis d'un algorithme : la finitude (« Un algorithme doit toujours se terminer après un nombre fini d'étapes »), une définition précise (« Chaque étape d'un algorithme doit être définie précisément, les actions à transposer doivent être spécifiées rigoureusement et sans ambiguïté pour chaque cas »), l'existence d'entrées (« des quantités lui sont données avant qu'un algorithme ne commence. Ces entrées sont prises dans un ensemble d'objets spécifié ») et de sorties (« des quantités ayant une relation spécifiée avec les entrées ») et un rendement (« toutes les opérations que l'algorithme doit accomplir doivent être suffisamment basiques pour pouvoir être en principe réalisées dans une durée finie par un homme utilisant un papier et un crayon »).

Le mot **algorithme** est issu de la latinisation du nom du **mathématicien Al-Khawarizmi**, dont le titre d'un des ouvrages (« *Abrégé du calcul par la restauration et la comparaison* »), écrit en arabe entre 813 et 833, est également à l'origine du mot **algèbre**. Il est le premier à proposer des méthodes précises de résolution des équations du second degré, du type «  $ax^2 + bx + c = 0$  ».

La longue histoire des algorithmes est bien décrite par Serge Abiteboul et Gilles Dowek, dans leur ouvrage *Le temps des algorithmes*. Ils y rappellent que les algorithmes sont **utilisés depuis des milliers d'années** : **Euclide a inventé en l'an 300 avant notre ère** un algorithme de calcul du plus grand diviseur commun de deux nombres entiers. Pour se représenter au sens familier ce qu'est un algorithme, il faudrait penser à une sorte de recette de cuisine : en suivant une série d'instructions, un ensemble d'ingrédients (les *inputs*) permettent de parvenir à un plat précis (les *outputs*).

En 1495, en vue de festivités organisées à Milan, **Léonard de Vinci** imagine puis construit, bien que ce dernier point reste débattu, un « **chevalier mécanique** », sorte de robot automate revêtu d'une armure médiévale. Sa structure interne en bois, avec quelques parties en métal et en cuir, était actionnée par un système de poulies et de câbles.

Avec ses « *animaux-machines* », **René Descartes** proposa, quant à lui, dans la première moitié du XVII<sup>e</sup> siècle, de reproduire artificiellement les fonctions biologiques, y compris la communication et la locomotion. **Blaise Pascal** à 19 ans réfléchit à la création d'une machine à calculer mécanique et, trois ans plus tard, aboutit à la création en 1645 de sa « *machine d'arithmétique* », initialement désignée ainsi puis dénommée roue pascaline et, enfin, pascaline. Projet qu'il abandonnera tout comme ses autres entreprises scientifiques pour se consacrer à l'étude de la philosophie et à la religion<sup>1</sup>. À la fin du XVII<sup>e</sup> siècle, **Gottfried Leibniz** imagine ensuite une machine à calculer capable de raisonner. Il construit lui aussi un prototype de machine à calculer en 1694, basée sur un cylindre cannelé<sup>2</sup>.

---

<sup>1</sup> Dans ses *Pensées*, Pascal affirme : « la machine d'arithmétique fait des effets qui approchent plus de la pensée que tout ce que font les animaux mais elle ne fait rien qui puisse faire dire qu'elle a de la volonté, comme les animaux ».

<sup>2</sup> Cette technologie sera d'ailleurs au principe de la première machine à calculer commercialisée, l'arithmomètre, invention française de Charles Xavier Thomas de Colmar pour laquelle il dépose un brevet en France en 1820, puis en Angleterre et en Belgique en 1851. Sa machine « propre à suppléer à la mémoire dans toutes les opérations d'arithmétique » repose sur les cylindres de Leibniz et nécessite de tourner des boutons. Elle a été fabriquée à 900 exemplaires de son vivant mais souffre d'une forte concurrence internationale, qui se contente de copier son modèle (Burkhardt, Saxonia, Bunzel, Archimedes, Tate, etc.). Les machines à calculer mécaniques américaines, dotées d'un clavier à touches, domineront le marché mondial sous le nom de comptomètres à partir de la fin du XIX<sup>e</sup> siècle jusqu'à l'invention des calculatrices électroniques dans les années 1960.

Pendant le siècle des Lumières, le philosophe français **Julien de la Mettrie** anticipe le jour où les progrès de la technique permettront de créer un homme-machine tout entier, à l'âme et au corps artificiels. L'abbé Mical et Kratzenstein imaginent une **machine à parler** en 1780, bientôt construite par le **baron Von Kempelen** grâce à une cornemuse à tuyaux multiples, aujourd'hui propriété du « *Deutsches Museum* » de Munich.

Au milieu du XIX<sup>e</sup> siècle, le logicien britannique **George Boole** appelle à mathématiser la logique en faisant du raisonnement déductif une série de calculs<sup>1</sup>, l'économiste britannique **William Stanley Jevons** inspiré par Boole imagine puis construit un piano mécanique capable de raisonner selon des prémisses de logique pure<sup>2</sup> et, surtout, le mathématicien britannique **Charles Babbage** conçoit, avec sa « **machine analytique** », l'ancêtre mécanique des ordinateurs modernes<sup>3</sup> en **associant les inventions de Pascal** et de **Jacquard** : à savoir, d'une part, la **machine à calculer**, d'autre part, les **programmes des métiers à tisser inscrits sur des cartes perforées**. C'est sur cette machine, qu'à 27 ans, **Ada Lovelace** écrit en 1842 le premier véritable programme informatique<sup>4</sup>, selon le calcul des nombres de Bernoulli, allant au-delà du simple calcul numérique. Celle que son père Lord Byron appelait, enfant, la princesse des parallélogrammes avait choisi, par passion pour les mathématiques, et dès ses 17 ans, de s'associer aux recherches de Babbage, dont elle travaillera à la promotion jusqu'à sa mort prématurée en 1852.

En 1870, dans son ouvrage ambitieux en deux volumes *De l'Intelligence*, **Hippolyte Taine** propose de traiter les facultés cognitives à travers des lois mathématiques comme on le ferait pour n'importe laquelle des sciences expérimentales, dont la physique. Pour lui, « *la perception extérieure est une hallucination vraie* » et « *tous les problèmes concernant un être quelconque, moral ou physique, seraient au fond des problèmes de mécanique* ».

---

<sup>1</sup> George Boole, 1847, *The mathematical analysis of logic. Being an essay towards a calculus of deductive reasoning*, Barclay, & Macmillan.

<sup>2</sup> Dans les années 1860, Jevons travaille à la construction d'un piano logique, qu'il finit en 1869 et présente à la Royal Society en 1870. Sa machine possède un clavier qui ressemble à celui d'un piano, mais à la place des notes, les touches comportent des lettres, de A à D, redoublées. Il y a par exemple un A pour la proposition A est vraie, et un autre A pour la proposition A est fausse. Les touches suivantes (B, C, et D) fonctionnent de la même manière et selon ces quatre prémisses possibles que l'utilisateur va pouvoir combiner, la machine calculera le résultat.

<sup>3</sup> En 1834, pendant le développement d'une machine à calculer, Charles Babbage imagine le premier ordinateur sous la forme d'une « machine à différences » en utilisant la lecture séquentielle des cartes du métier à tisser Jacquard afin de donner des instructions et des données à sa machine. En cela, il fut le premier à énoncer le principe d'un ordinateur.

<sup>4</sup> Charles Babbage avait avant cela écrit, en 1838, un algorithme de calcul des coefficients du produit de deux polynômes, mais en tant que simple programme séquentiel, il n'est pas considéré comme le premier programme informatique.

## 2. Un thème traditionnel de la science-fiction

Dès 1818, **Mary Shelley** publie son roman « *Frankenstein ou le Prométhée moderne* », dans lequel elle imagine un savant capable de créer un être artificiel, le monstre Frankenstein. **Jules Verne**, dans son roman, *La Maison à vapeur*, paru en 1880, imagine un éléphant à vapeur géant capable de traverser l'Inde, sur terre, comme sur l'eau. Sa machine n'est cependant pas autonome.

Alors qu'**Isaac Asimov** affirmait que l'« *on peut définir la science-fiction comme la branche de la littérature qui se soucie des réponses de l'être humain aux progrès de la science et de la technologie* », force est de constater que l'intelligence artificielle est un thème de science-fiction particulièrement fécond pour la **littérature**, le **cinéma** et les **jeux vidéo**.

Dans le célèbre roman *Erewhon* de **Samuel Butler**, paru en 1872, en particulier dans les trois chapitres qui forment *The Book of the Machines*, les machines sont douées d'une intelligence comparable à celle des êtres humains et risquent de nous dépasser. Quelques années plus tôt, sous le pseudonyme de Cellarius, dans un article publié le 13 juin 1863 dans le journal néo-zélandais *The Press* et intitulé « *Darwin among the Machines* », il jugeait inéluctable qu'à long terme « *les machines détiendront la réelle suprématie sur le monde et ses habitants, c'est ce qu'aucune personne d'un esprit vraiment philosophique ne peut un instant remettre en question* »<sup>1</sup>. Il concluait en appelant à la destruction de toutes les machines dans cette guerre à mort.

En 1920, le terme « robot » apparaît avec la pièce de théâtre de science-fiction de **Karel Čapek** « *R. U. R. Rossum's Universal Robots* », ce néologisme ayant été créé par son frère Josef à partir du mot tchèque *robota* qui signifie travail. Arthur C. Clarke met l'ordinateur CARL au centre de l'intrigue de ses romans *La Sentinelle*, en 1951, *À l'aube de l'histoire*, en 1953, et *2001, L'Odyssée de l'espace*, en 1968.

---

<sup>1</sup> Extrait de l'article « *Darwin among the Machines* » dans *The Press* du 13 juin 1863 : « We refer to the question: What sort of creature man's next successor in the supremacy of the earth is likely to be. We have often heard this debated; but it appears to us that we are ourselves creating our own successors; we are daily adding to the beauty and delicacy of their physical organisation; we are daily giving them greater power and supplying by all sorts of ingenious contrivances that self-regulating, self-acting power which will be to them what intellect has been to the human race. In the course of ages we shall find ourselves the inferior race (...). Day by day, however, the machines are gaining ground upon us; day by day we are becoming more subservient to them; more men are daily bound down as slaves to tend them, more men are daily devoting the energies of their whole lives to the development of mechanical life. The upshot is simply a question of time, but that the time will come when the machines will hold the real supremacy over the world and its inhabitants is what no person of a truly philosophic mind can for a moment question ».



L'un des pères fondateurs de l'IA, Marvin Minsky en l'occurrence, servit d'ailleurs de conseiller à **Stanley Kubrick** et à **Arthur C. Clarke** pour l'adaptation de ces romans au cinéma, avec **l'une des premières apparitions de l'IA dans la culture populaire** : l'ordinateur CARL rebaptisé **HAL 9000**, dans le film *2001 : l'Odyssée de l'espace*, sorti en 1968. Dans ce film, la communication avec l'ordinateur passe par une **interface de synthèse vocale** qui permet d'interagir avec le système par le langage naturel. Son nom correspond à un rétrodécalage de chacune des lettres d'IBM et l'acronyme signifie *Heuristically programmed ALgorithmic Computer*, dont on serait à la version 9000.

Les **nombreux ouvrages devenus des classiques** d'Isaac Asimov et d'Arthur C. Clarke, mais aussi de Philip K. Dick, de William Gibson, de Frank Herbert, de Francis Rayer ou de Iain Banks, illustrent cet intérêt marqué de la science-fiction pour le thème de l'IA, intérêt qui se poursuit aujourd'hui comme en témoignent les romans de Becky Chambers ou d'Alain Damasio<sup>1</sup>.

L'intelligence artificielle est aussi omniprésente **au cinéma, et ce depuis 1927**, avec de nombreux films, comme par exemple, chronologiquement : « *Metropolis* », « *2001 : l'Odyssée de l'espace* », « *Le Cerveau d'acier* », « *THX 1138* », « *Mondwest* », « *Les Femmes de Stepford* », « *Les Rescapés du futur* », « *Génération Proteus* », « *Star Wars* », « *Blade runner* », « *Alien* », « *Tron* », « *Wargames* », « *Terminator* », « *Virtuosity* », « *Matrix* », « *L'Homme bicentenaire* », « *A.I.* », « *I, Robot* », « *Iron Man* », « *Wall-E* », « *Eva* », « *The Machine* », « *Transcendance* », « *Chappie* », « *Her* », « *Ex Machina* », « *Ghost in the Shell* », « *Interstellar* », ou, encore, « *Ready Player One* ».

Des **séries télévisées** comme « *Lost in Space* », « *Star Trek : La Nouvelle Génération* », « *Battlestar Galactica* », « *K 2000* », « *Person of interest* », « *Emma* », « *Westworld* », « *Silicon Valley* », « *Better than us* », « *Star Wars : The Clone Wars* » ou, surtout, « *Black Mirror* », « *Real Humans* » et « *Humans* » ont également exploité ce sujet.

Les thèmes de **l'hostilité de l'intelligence artificielle** ou des **risques que cette dernière ferait courir à l'espèce humaine** sont souvent au cœur de l'intrigue de ces œuvres. Des chercheurs de l'Université de Cambridge ont quant à eux identifié, en 2019, sur le fondement de l'étude de 300 œuvres, **quatre thèmes principaux** dans les fictions représentant l'IA : l'immortalité, l'espoir d'une vie libérée du travail, la satisfaction de nos désirs et, surtout, la soif de domination<sup>2</sup>. Leur analyse conclut sur l'idée que toutes ces perceptions des possibilités ouvertes par l'IA, qui peuvent être assez éloignées de la réalité des technologies, peuvent toutefois **influencer la façon dont elles sont développées, déployées et réglementées**.

---

<sup>1</sup> Les romans récents d'Antoine Bello (comme *Ada*), de Christian Léourier (*Helstrid*) ou de Steven Erikson (*Rejoice*) sont d'autres exemples de cette inspiration par le thème de l'IA.

<sup>2</sup> Cf. Stephen Cave et Kanta Dihal, 2019, « *Hopes and fears for intelligent machines in fiction and reality* », *Nature Machine Intelligence*, volume 1, n° 2 : <https://www.nature.com/articles/s42256-019-0020-9>

## B. LES AVANCÉES DEPUIS UN SIÈCLE : DU BAPTÊME DE 1956 À LA CONFIRMATION DE 2017

### 1. L'école d'été de Dartmouth de 1956, le moment fondateur de la définition de l'IA

L'intelligence artificielle a fêté cet été son **soixante-huitième anniversaire**, puisqu'elle est inventée en tant que concept et discipline en 1956 à l'occasion d'une école d'été qui a rassemblé les pères fondateurs de la discipline pendant huit semaines.

Le concept a fait l'objet de longs débats et il est dit *a posteriori* que le choix du mot doit beaucoup à la **quête de visibilité de ce nouveau champ de recherche**. Parler d'**intelligence artificielle** a pu apparaître comme **plus séduisant que de parler d'informatique avancée ou des sciences et technologies du traitement automatisé de l'information**. L'anthropomorphisme essentialiste<sup>1</sup> qui est exprimé par le choix du concept d'« intelligence artificielle » n'a sans doute pas contribué à apaiser les peurs suscitées par le projet prométhéen de construction d'une **machine rivalisant avec l'intelligence humaine**, même si ce n'était pas le projet en 1956 de cette discipline, dont l'ambition plus modeste était de **simuler tel ou tel aspect de nos fonctions cognitives dites « intelligentes »**.

Avant cet événement qui lui donna son nom, l'intelligence artificielle avait déjà été **imaginée par les pères fondateurs de l'informatique moderne**.

Dès 1936, **Alan Turing pose ainsi les fondements théoriques de l'informatique et introduit les concepts de programme et de programmation**. Il imagine en effet, à ce moment, un modèle de fonctionnement pour un appareil doté d'une capacité élargie de calcul et de mémoire, en recourant à l'image d'un ruban infini muni d'une tête de lecture/écriture. Un tel appareil sera appelé « **machine de Turing** », précurseur théorique de l'ordinateur moderne.

Puis, après avoir décrypté à Bletchley Park en 1942 le code nazi Enigma<sup>2</sup>, dans un article paru en 1950<sup>3</sup>, Alan Turing explore le problème de l'intelligence artificielle et propose une expérience dénommée « *the imitation game* », maintenant connue sous le nom de « **test de Turing** », qui est une tentative de définition, à travers une épreuve, d'un critère permettant de

---

<sup>1</sup> L'anthropomorphisme est l'attribution de caractéristiques du comportement humain ou de la morphologie humaine à d'autres entités comme des dieux, des animaux, des objets ou d'autres phénomènes. L'essentialisme est l'attribution à un être ou à un objet d'une existence propre « par essence », c'est-à-dire inhérente au sujet en question.

<sup>2</sup> Au sein d'une équipe composée de nombreux scientifiques – jusqu'à 7 000 ont participé en même temps aux projets de l'opération Ultra –, Turing s'est illustré par ses talents. Outre Enigma, ces cryptanalystes ont également cassé les codes Purple et Lorenz, notamment grâce à l'ordinateur Colossus fabriqué en 1943, premier ordinateur électronique numérique à voir le jour dans le monde, capable d'accomplir jusqu'à 5 000 opérations par seconde. Le 2<sup>e</sup> ordinateur Colossus appelé Mark II a notamment été utilisé à l'occasion du débarquement de Normandie. Dix machines Colossus ont au total été produites pendant la Seconde Guerre mondiale.

<sup>3</sup> Alan Turing, octobre 1950 « Computing Machinery and Intelligence », *Mind*.

qualifier une machine d'« intelligente »<sup>1</sup>. Il fait alors le pari que les machines vont réussir son test à moyen terme : « *d'ici à cinquante ans, il n'y aura plus moyen de distinguer les réponses données par un homme ou un ordinateur, et ce sur n'importe quel sujet* ». Cette prophétie d'Alan Turing quant aux progrès connus en l'an 2000 est en cours de réalisation aujourd'hui avec les IA génératives, certes avec 25 ans de retard mais c'est en réalité très peu au regard du rythme global du progrès technique.

C'est au milieu du XX<sup>e</sup> siècle, en 1955, en même temps que l'informatique se développe, qu'**apparaît formellement une première fois la notion d'« intelligence artificielle »** avec une définition peu connue proposée alors par John McCarthy, à ce moment jeune professeur assistant de mathématiques au Collège de Dartmouth. Il propose d'en faire :

*« la science et l'ingénierie de la fabrication de machines intelligentes ».*

On notera le caractère double de l'IA dès cette première définition : discipline scientifique et savoir-faire pratique pour assurer la production de produits. Mais, surtout, le problème de cette définition est qu'elle est **réursive**, elle tourne sur elle-même : l'intelligence artificielle est définie par la science et la fabrication de « machines intelligentes », dont la présence permettrait d'identifier l'intelligence artificielle. Elle n'est donc pas satisfaisante. Une autre définition de l'IA sera donnée un an plus tard, à l'occasion d'une école d'été qu'il organise et qui est intitulée « *The Dartmouth Research Project on Artificial Intelligence* ».

Lors d'une conférence au Collège de Dartmouth à l'été 1956, l'intelligence artificielle est en effet **définie et actée comme un champ de recherche au sein de l'informatique**.

Cette école d'été de huit semaines, organisée par John McCarthy et Marvin Minsky, mais surtout soutenue par la fondation Rockefeller, par Nathan Rochester, alors directeur scientifique d'IBM, et par Claude Shannon, ingénieur, mathématicien, chercheur aux laboratoires Bell et père des théories de l'information et de la communication, offre d'abord à John McCarthy l'occasion de **convaincre la vingtaine de chercheurs y participant d'accepter l'expression « intelligence artificielle »** en tant que nouveau nom pour ce domaine de recherche. La conférence pose donc les bases axiomatiques de l'IA dès 1956 avec l'idée que « *la discipline se fonde sur l'hypothèse que chaque aspect de l'apprentissage ou toute autre caractéristique de l'intelligence peut en principe être décrit avec une telle précision qu'une machine peut être conçue pour le simuler* ».

---

<sup>1</sup> Le test de Turing consiste à mettre en confrontation verbale un humain avec un ordinateur imitant la conversation humaine et un autre humain. Dans le cas où l'homme qui engage les conversations n'est pas capable de dire lequel de ses interlocuteurs est un ordinateur, on peut considérer que le logiciel de l'ordinateur a passé avec succès le test.

La rigueur pousse à observer que le projet n'est pas, en réalité, de construire une machine rivalisant avec l'homme mais de **simuler telle ou telle tâche que l'on réserve habituellement à l'intelligence humaine**. Il est alors affirmé que **tout aspect de l'intelligence humaine peut être décrit de façon assez précise pour qu'une machine le reproduise en le simulant**.

À l'occasion de cette conférence, une nouvelle définition du concept est donnée par Marvin Minsky, qui a l'avantage de ne pas être récursive et de mettre l'accent sur **le fait que l'IA reste de l'informatique** et sur l'aspect dynamique de ces technologies (« *pour l'instant* »), même si elle reste floue et peu rigoureuse :

*« Construction de programmes informatiques capables d'accomplir des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains ».*

Le projet d'école d'été de Dartmouth et quelques-uns de ses participants

# A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

August 31, 1955

*John McCarthy, Marvin L. Minsky,  
Nathaniel Rochester,  
and Claude E. Shannon*

■ The 1956 Dartmouth summer research project on artificial intelligence was initiated by this August 31, 1955 proposal, authored by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. The original typescript consisted of 17 pages plus a title page. Copies of the typescript are housed in the archives at Dartmouth College and Stanford University. The first 5 papers state the proposal, and the remaining pages give qualifications and interests of the four who proposed the study. In the interest of brevity, this article reproduces only the proposal itself, along with the short autobiographical statements of the proposers.

guage, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

## 1. Automatic Computers

If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have.

## 2. How Can a Computer be Programmed to Use a Language

It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning and rules of conjecture. From this point of view, forming a generalization consists of admitting a new

**W**e propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use lan-



John McCarthy



Marvin Minsky



Claude Shannon



Ray Solomonoff



Alan Newell



Herbert Simon



Arthur Samuel



Oliver Selfridge



Nathaniel Rochester



Trenchard More

Source : <https://akin-agunbiade.medium.com/the-first-wave-of-ai-1956-1973-f10860a807f9>

Cette idée directrice a contribué à définir les travaux en intelligence artificielle qui ont suivi. Toutefois, lors de son audition devant vos rapporteurs, Jean-Gabriel Ganascia, professeur émérite à Sorbonne-Université et président du comité d'éthique du CNRS a souligné qu'il ne s'agit là que de la **définition scientifique** de l'IA, en tant que domaine de recherche mais que le terme « intelligence artificielle » peut également avoir une **définition technologique** et une **conception populaire**, qui substantialise l'IA et prête un esprit aux machines. Outre la définition initiale de Marvin Minsky de 1955 - qui parlait de la construction de programmes informatiques capables d'accomplir des tâches du niveau de l'intelligence humaine - une des définitions technologiques les plus consacrées est élaborée par Peter Norvig et Stuart Russell dans le manuel de référence de la discipline depuis une trentaine d'années<sup>1</sup> : « *Agents qui perçoivent depuis leur environnement et exécutent des actions en conséquence* ». On retrouve la dualité évoquée en 1955 dans la première définition oubliée de l'IA proposée par John McCarthy : à la fois une science ET une ingénierie de la fabrication de machines.

Cette définition technologique, même si elle paraît vague, permet d'englober l'ensemble des systèmes d'IA, du robot conversationnel Eliza à l'IA d'IBM Watson ou encore les systèmes d'intelligence artificielle générative actuels comme ChatGPT.

En effet, les robots conversationnels reçoivent, du côté des perceptions, des intrants sous la forme des instructions de l'utilisateur, et génèrent des textes, du côté des actions, en réponse à ces invites.

L'OCDE donne de l'IA une définition assez proche puisqu'elle affirme qu'un système d'IA est un « *système qui fonctionne grâce à une machine et est capable d'influencer son environnement en produisant des résultats (tels que des prédictions, des recommandations ou des décisions) pour répondre à un ensemble donné d'objectifs. Il utilise les données et les intrants générés par la machine et/ou apportés par l'homme afin de (i) percevoir des environnements réels et/ou virtuels ; (ii) produire une représentation abstraite de ces perceptions sous forme de modèles issus d'une analyse automatisée (ex. l'apprentissage automatisé) ou manuelle ; et (iii) utiliser les déductions du modèle pour formuler différentes options de résultats. Les systèmes d'IA sont conçus pour fonctionner de façon plus ou moins autonome* »<sup>2</sup>. L'Union européenne reprend le cadre proposé par l'OCDE. Enfin, il existe une conception populaire de l'IA, la plus problématique, qui n'est pas scientifique et correspond malheureusement au sens commun de ce que l'opinion va intuitivement considérer comme étant l'intelligence artificielle. Cette définition, floue et nourrie par la science-fiction, substantialise l'IA dans **un biais anthropomorphique** et lui rattache des concepts comme l'IA générale, la singularité technologique ou la perspective de robots qui nous dépassent, le risque de prise de pouvoir par les machines, la volonté de dominer l'homme.

---

<sup>1</sup> Peter Norvig et Stuart Russell, 2010, Artificial intelligence: a modern approach. Prentice Hall, Upper Saddle River: <http://www.worldcat.org/oclc/688385283>

<sup>2</sup> Cf. le site de l'OCDE dédié à l'IA et en particulier sa page de présentation de ses principes pour l'IA : <https://oecd.ai/fr/ai-principles>

Cette définition, **plus proche de la magie que de la technologie**, prête à l'IA une **conscience**, lui associe une sorte d'esprit que pourrait avoir la machine. Bien qu'elle n'ait rien de scientifique, cette définition, puisqu'elle est populaire, est vectrice de représentations et de récits catastrophistes qui sont instrumentalisés par certains acteurs souhaitant véhiculer l'idée selon laquelle les IA constitueraient un danger pour l'humanité. En réalité, on l'a vu avec les deux définitions précédentes plus objectives, cette vision mystifiée est erronée. Vos rapporteurs renvoient à la lecture du rapport de l'OPECST de 2017 sur l'IA qui abordait déjà cette question et apportait des pistes pour la démystification de l'IA et de ces représentations.

## 2. Printemps et Hivers de l'IA

Outre John McCarthy et Marvin Minsky, les participants à l'école d'été de Dartmouth de 1956, tels que Ray Solomonoff, Oliver Selfridge, Trenchard More, Arthur Samuel, Allen Newell et Herbert Simon, ayant posé comme conjecture que « *tout aspect de l'intelligence humaine peut être décrit de façon assez précise pour qu'une machine le reproduise en le simulant* », **discutent ensuite des possibilités de créer des programmes d'ordinateur qui se comportent intelligemment**, c'est-à-dire qui résolvent des problèmes dont on ne connaît pas de solution algorithmique simple.

Tel est le programme que se donnent ces chercheurs américains et qui va recevoir le soutien décisif des autorités fédérales américaines.

Dans les années suivantes, **les chercheurs mettent ainsi au point de nouvelles techniques informatiques** généreusement financées par l'agence américaine pour les projets de recherche avancée de défense du ministère de la défense (*Defense Advanced Research Projects Agency* ou DARPA), mais aussi par IBM :

- le langage de programmation Lisp en 1958, l'un des plus anciens langages de programmation<sup>1</sup>, premier programme à mobiliser des symboles plutôt que des nombres et qui a fait émerger la notion d'heuristique (méthode permettant de donner rapidement des solutions satisfaisantes à un problème d'optimisation complexe, sans aboutir nécessairement à des solutions optimales) ;

---

<sup>1</sup> John McCarthy a inventé le langage de programmation Lisp dès 1958, c'est un mot-valise formé à partir de l'anglais list processing ou traitement de listes. Si l'on met de côté la « machine de Turing » qui relève de l'informatique théorique, le « système A-0 » (ou « A-0 System ») est le premier compilateur (programme qui transforme un code source écrit dans un langage de programmation ou langage source en un autre langage informatique, appelé langage cible) développé en 1952 ; il est suivi notamment par le Fortran (mot-valise issu de l'anglais « formula translator ») inventé dès 1954, Lisp et Algol en 1958, Cobol (acronyme de « Common Business Oriented Language ») en 1959, BASIC (acronyme de « Beginner's All-purpose Symbolic Instruction Code ») en 1964, Logo en 1967, Pascal en 1971, ou, encore, Prolog (mot-valise pour Programmation logique), inventé par des chercheurs français en 1972. Parmi tous ces langages de programmation, le Lisp occupe une place particulière, cf. John McCarthy, Avril 1960, « Fonctions Récursives d'expressions symboliques et leur évaluation par une Machine » ou « Recursive Functions of Symbolic Expressions and Their Computation by Machine », *Communications of the ACM*.



- une première concrétisation des réseaux de neurones artificiels, sous la forme du Perceptron, dont Marvin Minsky souligne dès son invention les limites théoriques ;

- un programme qui joue aux dames et met en œuvre un apprentissage lui permettant de jouer de mieux en mieux...

Toutes ces découvertes rendent alors les pères fondateurs de l'intelligence artificielle **très optimistes, peut-être trop**.

En 1958, Herbert Simon et Allen Newell déclarent ainsi que « *d'ici à dix ans un ordinateur sera le champion du monde des échecs* » et même que « *d'ici à dix ans, un ordinateur découvrira et résoudra un nouveau théorème mathématique majeur* » : il faudra, en réalité, attendre 1997 pour que le champion d'échecs Garry Kasparov s'incline devant le système Deep Blue d'IBM et 2023 pour qu'un modèle d'IA parvienne à résoudre un problème mathématique jusqu'alors non résolu par l'homme<sup>1</sup>.

La **représentation des connaissances et le langage objet** sont au cœur de l'intelligence artificielle des années 1960. Dès 1960, deux chercheurs norvégiens Ole-Johan Dahl et Kristen Nygaard inventent le premier langage orienté objet, appelé SIMULA. Ces approches seront ensuite mis au service de l'informatique dans les années 1970, avec des résultats remarquables permettant les progrès connus vers les ordinateurs modernes.

En 1960, Joseph Carl Robnett Licklider – qui, à travers Arpanet dans les années 1970, fondera les bases d'Internet – rédige un article sur le **rapprochement homme-machine** (qu'il nomme « *Man-Computer Symbiosis* ») et souligne le besoin d'interactions simplifiées entre les ordinateurs et leurs utilisateurs. Il y taille en pièces l'idée d'une IA concurrençant ou *a fortiori* remplaçant l'homme : « *les hommes fixeront les buts, formuleront des hypothèses, détermineront des critères et exécuteront les évaluations. Les ordinateurs feront le travail que l'on peut mettre en routine qui doit être fait pour préparer les idées et les décisions liées à la pensée technique et scientifique* »<sup>2</sup>.

---

<sup>1</sup> Cf. Will Douglas Heaven, décembre 2023, « *Google DeepMind used a large language model to solve an unsolved math problem* », MIT Technology Review : <https://www.technologyreview.com/2023/12/14/1085318/google-deepmind-large-language-model-solve-unsolvable-math-problem-cap-set/> et Bernardino Romera-Paredes et al., 2024, « *Mathematical discoveries from program search with large language models* », Nature n°625, 468–75 : <https://doi.org.stanford.idm.oclc.org/10.1038/s41586-023-06924-6>

<sup>2</sup> La citation originale est la suivante : « *Men will set the goals, formulate the hypotheses, determine the criteria, and perform the evaluations. Computing machines will do the routinizable work that must be done to prepare the way for insights and decisions in technical and scientific thinking* ». Cf. Joseph Carl Robnett Licklider, 1960, « *Man-Computer Symbiosis* ».



En 1965, Herbert Simon assure de manière très optimiste que « *des machines seront capables, d'ici à vingt ans, de faire tout travail que l'homme peut faire* ». En 1967, Marvin Minsky estime que « *dans une génération [...] le problème de la création d'une intelligence artificielle sera en grande partie résolu* » et en 1970 que « *dans trois à huit ans nous aurons une machine avec l'intelligence générale d'un être humain ordinaire* ».

De même, le **premier agent conversationnel** (« *chatbot* » ou « *bot* ») est créé en 1966 par Joseph Weizenbaum et simule un psychothérapeute grâce à sa technologie de reconnaissance des formes. Il s'appelle « *Eliza* » et suscite un grand enthousiasme.

Mais **ses capacités restent limitées**, puisqu'il est incapable de vraiment répondre aux questions posées, se contentant de continuer à faire parler son interlocuteur, dans une logique de relance. Tel un psychanalyste, il a surtout tendance à reformuler le plus souvent les propos de l'utilisateur sous un format interrogatif.

#### Une conversation avec Eliza, chatbot créé en 1966

```
Welcome to
          EEEEE  LL      IIII  ZZZZZZZ  AAAAA
          EE     LL      II    ZZ     AA  AA
          EEEEE  LL      II    ZZZ    AAAAAAA
          EE     LL      II    ZZ     AA  AA
          EEEEE  LLLLLL IIII  ZZZZZZZ  AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:   █
```

Source : Norbert Landsteiner <https://fr.slideshare.net/ashir233/eliza-4615>

« L'âge d'or » des approches symboliques et des raisonnements logiques se produit dans les années 1960 après la naissance de l'intelligence artificielle à Dartmouth. **Recourant à des connaissances précises, telles que des logiques diverses ou des grammaires, ces formes d'intelligence sont dites explicites.**

Apparaissent ensuite, les diverses modalités de **formalisme logique**, soit sous la forme de **logique classique**, de **logique floue**, de **logique modale** ou de **logique non monotone**.

L'IA symbolique pose alors que la **logique mathématique peut en effet représenter des connaissances<sup>1</sup> et modéliser des raisonnements**. Le principe de résolution permet d'automatiser ces raisonnements : pour démontrer une propriété, on montre que son contraire entraîne une contradiction avec ce qu'on sait déjà. La seule règle utilisée est celle du « détachement » ou *modus ponens*, figure du raisonnement logique concernant l'implication (exemple : « si p implique q et si p, alors q »). Cette méthode ne s'applique qu'à des cas simples, où la combinatoire n'est pas excessive. Fondé sur le même principe, le langage Prolog (acronyme de PROgrammation LOGique, qui permet de résoudre les problèmes par raisonnement à partir de règles de logique formelle) lève ces restrictions en permettant d'aborder des problèmes plus complexes.

**Des difficultés subsistent pour traiter des connaissances vagues ou incomplètes**. Devant ces limites, des extensions théoriques ont donné lieu à des logiques non classiques permettant d'exprimer plus d'éléments que dans la logique classique. Voulant étendre les possibilités de la logique classique, les logiques multivaluées gardent les mêmes concepts de base, hormis les valeurs de vérité, qui, selon les théories, varient de trois à un nombre infini de valeurs. La **théorie des logiques floues** étend ces logiques en considérant comme valeurs de vérité le sous-ensemble réel « [0,1] ». Elles permettent de traiter des informations incertaines (Jean viendra peut-être demain) ou imprécises (Anne et Brigitte ont à peu près le même âge).

Les **logiques modales** introduisent des notions comme la possibilité, la nécessité, l'impossibilité ou la contingence qui modulent les formules de la logique classique. La **notion de vérité devient relative** à un instant donné ou à un individu. On distingue ainsi ce qui est accidentellement vrai (contingence : Strasbourg est en France) de ce qui ne peut pas être faux (nécessité : un quadrilatère a quatre côtés). Diverses interprétations des modalités donnent lieu à des applications distinctes, dont les plus importantes sont les logiques épistémiques (savoirs, croyances), déontiques (modélisant le droit) et temporelles (passé, présent, futur).

Les connaissances n'étant pas universelles, on peut être conduit à des hypothèses et suppositions fausses, remises en cause à la lumière d'expériences ultérieures. Les **logiques non monotones** tiennent compte du fait que les exceptions sont exceptionnelles et formalisent les raisonnements où l'on adopte des hypothèses (tous les oiseaux volent) qui pourront être modifiées par des connaissances plus précises (mais pas les autruches).

---

<sup>1</sup> Des symboles permettent alors de représenter des faits et des règles permettent d'en déduire de nouveaux.

On raisonne avec des règles du type : si a est vrai et si b n'est pas incohérent avec ce qu'on sait, on peut déduire c (si Titi est un oiseau et si j'ignore que c'est une autruche, il vole). **On autorise ainsi la prise de décision malgré une information incomplète** : des suppositions plausibles permettent certaines déductions ; si, à la lumière d'informations ultérieures, ces suppositions se révèlent fausses, on remettra en question les déductions précédentes (non-monotonie).

S'agissant des **grammaires**, le traitement automatique des langues est un des grands domaines de l'intelligence artificielle, qui vise l'application de ses techniques aux langues humaines. Très pluridisciplinaire, il collabore avec la linguistique, la logique, la psychologie et l'anthropologie. Les travaux en traitement automatique des langues ont donné lieu à la constitution de divers ensembles de données numériques (dictionnaires de langue, de traduction, de noms propres, de conjugaison, de synonymes ; grammaires sous diverses formes ; données sémantiques), ainsi qu'à divers logiciels (analyseurs et générateurs morphologiques ou syntaxiques, gestionnaires de dialogue...). Du point de vue conceptuel, ces travaux ont produit des théories grammaticales plus compatibles avec les questions d'informatisation, des théories formelles pour la représentation du sens des mots, des phrases, des textes et des dialogues, ainsi que des techniques informatiques spécifiques pour le traitement de ces éléments par un ordinateur.

De **grands espoirs** sont alors placés dans la compréhension du langage naturel, dans la vision artificielle, mais en fin de compte **les résultats sont décevants**, largement en raison des limitations de puissance du matériel disponible, des données à utiliser mais aussi des limites intrinsèques des technologies alors disponibles.

Ainsi, le **Perceptron**, dans lequel Frank Rosenblatt plaçait tant d'espérances, est rapidement critiqué. Le livre *Perceptrons* de Marvin Minsky et Seymour Papert, paru en 1969, démontre les limites théoriques des réseaux de neurones artificiels de l'époque<sup>1</sup> alors qu'aucune application industrielle du perceptron n'émerge. Cette technologie sera analysée plus loin.

Après cet **âge d'or, qui court de 1956 au début des années 1970**, les **financements sont revus à la baisse** en raison de **différents rapports assez critiques** : les prédictions exagérément optimistes des débuts ne se réalisent pas et les techniques ne fonctionnent que dans des cas simples. À l'évidence, les difficultés fondamentales de l'intelligence artificielle furent alors largement sous-estimées, en particulier la question de savoir comment donner des connaissances de sens commun à une machine. Les recherches se recentrent alors sur la programmation logique, les formalismes de représentation des connaissances et sur les processus qui les utilisent au mieux.

---

<sup>1</sup> La critique principale concerne l'incapacité du Perceptron à résoudre les problèmes non linéairement séparables, tels que le problème du « X OR » (« OU exclusif »). Il s'en suivra alors, en réaction à la déception, une période noire d'une vingtaine d'années pour les réseaux de neurones artificiels.

En dépit de cette réorientation, qui témoigne d'**une certaine cyclicité des investissements en intelligence artificielle selon une boucle « espoirs-déceptions »**, Marvin Minsky et ses équipes du MIT (Massachusetts Institute of Technology) développent divers systèmes (Sir, Baseball, Student...) qui relancent les recherches sur la compréhension automatique des langues.

Au cours des années 1980, **de nouveaux financements publics** sont ouverts avec le projet japonais dit de « cinquième génération », le programme britannique Alvey, le programme européen Esprit et, surtout, le soutien renouvelé de la DARPA aux États-Unis. Les approches sémantiques sont alors en plein essor, en lien avec les sciences cognitives, ainsi que la représentation des connaissances, les systèmes experts et l'ingénierie des connaissances. Leurs usages dans le monde économique sont des signes de cette vitalité.

Après ce court regain d'intérêt, la recherche subit à nouveau un déclin des investissements. Les succès de cette approche restent en effet très relatifs car celle-ci ne fonctionne bien que dans des domaines trop restreints et trop spécialisés. L'incapacité de l'étendre à des problèmes plus vastes conduit à un désintérêt pour l'intelligence artificielle.

L'enthousiasme renouvelé dans les années 1980 autour des systèmes experts, de leurs usages et de l'ingénierie des connaissances précède donc un **second « hiver de l'intelligence artificielle » dans les années 1990**.

Pour autant, des découvertes scientifiques sont réalisées dans la période. Après la renaissance de l'intérêt pour les **réseaux de neurones artificiels** avec de nouveaux modèles théoriques de calculs, les années 1990 voient se développer **les algorithmes et la programmation génétique** ainsi que les **systèmes multi-agents** ou **l'intelligence artificielle distribuée**. La nécessité de métaconnaissances<sup>1</sup> émerge également.

En 1997, le système **Deep Blue** d'IBM **bat le champion du monde d'échecs de l'époque, Garry Kasparov**, qui était sorti victorieux lorsqu'il l'affronta une première fois en 1996. Ce superordinateur d'architecture massivement parallèle était capable d'évaluer au moins 200 millions de positions par seconde grâce à sa puissance de 11,4 gigaflops (soit 11 milliards d'opérations par seconde).

---

<sup>1</sup> Il s'agit de connaissances à propos des connaissances elles-mêmes.

### 3. Les années 2010 : une décennie d'innovations et de progrès spectaculaires

Après Deep Blue en 1997, il faut attendre **2011** pour que les IA d'IBM refassent parler d'elles. Le système **Watson** participe ainsi au jeu télévisé américain **Jeopardy** – où il s'agit de trouver les questions correspondant à des réponses – qu'il remporte en face des plus grands champions du jeu et gagne un million de dollars. La puissance de calcul atteinte par Watson est alors de 80 téraflops, soit 80 000 milliards d'opérations par seconde. Il s'appuie sur 200 millions de pages de contenus qu'il mobilise en moins de trois secondes. On estime alors que Watson est le premier système à se rapprocher de l'objectif fixé par le test de Turing. Un an plus tard IBM commercialise Watson en solutions logicielles d'analyse pour les entreprises (*business analytics*). Les secteurs médicaux, financiers ou encore juridiques ont été des clients notables.

Entretemps, le secteur du numérique avait été bouleversé par la **massification des usages d'Internet**, à commencer par le Web. Ces évolutions ont posé un **nouveau cadre favorisant le développement des technologies d'intelligence artificielle**, avec une explosion des **données** mises en ligne, de nouvelles **capacités de financement**, des **intérêts économiques puissants** et une **interconnexion** des chercheurs, des développeurs et des entreprises. Si l'IA avance rapidement dans les années 2010, c'est assez largement grâce à ce contexte nouveau dessiné dans les années 2000 qui lui a permis de se développer de manière inédite.

En matière technologique, les années 2010 ont représenté la **décennie du Machine Learning et du Deep Learning** avec de nombreuses avancées assez spectaculaires. Ces innovations naissent **quasiment par sérendipité** à l'occasion d'un concours de reconnaissance d'images par ordinateur. C'est l'occasion de mobiliser les algorithmes déjà disponibles mais surtout de bénéficier d'un **corpus de données de taille inédite**.

Alors qu'acquérir des capacités satisfaisantes en reconnaissance visuelle avait toujours constitué une difficulté pour les IA, des progrès inédits sont alors enregistrés, représentant une avancée dans la résolution du paradoxe de Moravec, bien connu des spécialistes de l'IA. En effet, ce paradoxe, formulé par Hans Moravec dans les années 1980, montre que nos **capacités de perception et de motricité** nous semblent plutôt **faciles et intuitives** (comme la marche ou la reconnaissance d'objets) mais sont **très difficiles à reproduire sous la forme d'intelligences artificielles** tandis que des **tâches cognitives de haut niveau** (comme des calculs mathématiques complexes) sont **très faciles** pour elles.

La chercheuse **Fei-Fei Li** a ainsi commencé à travailler sur l'idée d'un jeu de données d'images annotées en 2006 (contenant par exemple l'information « il y a des chats dans cette image » ou « il n'y a pas de chats dans cette image »). Quatre ans plus tard, c'est sur la base de ce jeu de données (ou *dataset*) d'environ 1,5 million d'images annotées, appelé **ImageNet**, qu'a été organisé un concours annuel : *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC).

Et c'est précisément au cours des années 2010, sous l'effet de cette compétition, que des **progrès inédits dans le domaine du traitement des images** sont enregistrés. En 2011, les plus faibles taux d'erreur de classification de cette compétition pour la reconnaissance d'images au sein du jeu de données étaient d'environ 25 %. En 2012, l'apprentissage profond permet de faire baisser ce taux à 16 %, et les deux années suivantes il tombe à quelques pour cent.

Les algorithmes basés sur **les réseaux de neurones artificiels en couches profondes ont, de loin, triomphé de tous les autres algorithmes**. La compétition autour d'ImageNet devient à ce moment une sorte de *benchmark* de tous les algorithmes développés.

Si l'on disposait bien des **algorithmes efficaces**, fondés surtout sur des architectures de type réseau de neurones artificiels et d'algorithmes de traitement, ce n'est qu'à ce moment-là, en conjuguant **des calculs rapides sur des machines puissantes** avec de **larges bases de données labellisées**, capables d'entraîner des modèles de grande dimension, que les IA connexionnistes sont reconnues comme efficaces. **Les trois conditions du succès du *Machine Learning*** étaient enfin remplies.

Ces systèmes, et parmi eux surtout ceux de *Deep Learning*, ont alors pu, **grâce à la conjonction de ces trois facteurs, commencer à s'imposer face à toutes les autres méthodes**, d'abord pour reconnaître les images et la voix, puis dans d'autres domaines, comme la traduction, avec des performances jamais obtenues auparavant. Ces systèmes passent alors par de longues étapes d'annotation des données, la labellisation est en effet indispensable dans ce contexte d'un apprentissage qui reste strictement supervisé. Un travail lourd réalisé par des annotateurs est donc nécessaire.

Les **assistants personnels basés sur la reconnaissance de la parole**, comme Siri (Apple), Alexa (Amazon), Cortana (Microsoft), Bixby (Samsung) ou Google Assistant, sont conçus et se déploient dans ce contexte précis des années 2010.

L'année 2014 est marquée par l'introduction des GAN (*Generative Adversarial Networks*) sur la base des travaux de Ian Goodfellow, permettant de **créer des images réalistes grâce à l'IA**.

En octobre 2015, le programme **AlphaGo** – entraîné à jouer au jeu de go grâce à une architecture élaborée par Google DeepMind couplant alors<sup>1</sup> apprentissage profond, apprentissage par renforcement et méthode de Monte-Carlo avec un réseau de valeur et un réseau d’objectifs – bat le champion européen Fan Hui par cinq parties à zéro.

En mars 2016, le même programme **bat le champion du monde Lee Sedol** par quatre parties à une. Cette victoire est une étape capitale car faire jouer une IA au go est un problème plus difficile que les échecs, non seulement car il existe un plus grand nombre de combinaisons possibles, mais surtout parce que la victoire finale et les objectifs intermédiaires sont beaucoup moins clairs, nécessitant des stratégies très complexes, passant notamment par le fait de perdre, voire par le *bluff*.

Ce dernier trait caractérise le poker, longtemps considéré comme un jeu inaccessible aux machines. Or, en 2017, le programme **Libratus** développé à l’université Carnegie-Mellon sort **vainqueur d’un grand tournoi de poker** l’opposant à plusieurs joueurs professionnels. Il repose sur un algorithme probabiliste adaptatif utilisant une variante de la technique de minimisation du regret hypothétique<sup>2</sup>. Par rapport aux dames, aux échecs ou au go, le poker présente en effet la particularité de devoir **optimiser la stratégie de jeu sans disposer de données sur les moyens des adversaires, c’est-à-dire sans connaître les mains des autres joueurs**.

L’année 2017 est aussi une **année charnière car la plupart des pays du monde accélèrent leur engagement dans la course mondiale autour de l’IA et annoncent des stratégies sur l’IA**, plus pour promouvoir ces technologies que pour les encadrer.

C’est, enfin, et peut-être surtout, **l’année où est inventée l’architecture qui va faire progresser comme jamais les IA génératives** et en particulier les grands modèles de langage, ou LLM (de l’anglais *Large Language Model*) : la technologie *Transformer*, véritable avancée pour l’intelligence artificielle, sujet au cœur du présent rapport et analysée plus loin. Si selon Laurent Alexandre, entendu par vos rapporteurs, le XXI<sup>e</sup> siècle ne naît qu’en 2022 avec ChatGPT, la date à retenir serait plutôt celle de l’année d’invention du modèle qui a rendu ChatGPT possible, soit la proposition de l’architecture *Transformer* en 2017.

---

<sup>1</sup> Après la fin de l’année 2017, AlphaGo repose sur une autre architecture sans la méthode de Monte-Carlo mais parvenant pourtant très rapidement à des performances supérieures. Cette nouvelle version baptisée AlphaGo Zero remplace l’apprentissage à partir des parties des grands maîtres de go par une nouvelle méthode d’apprentissage par renforcement très efficace qui permet au système d’atteindre des niveaux encore moins rattrapables par l’homme. Cf. David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel et Demis Hassabis, 19 octobre 2017, « Mastering the game of Go without human knowledge », *Nature*, vol. 550, : <https://arxiv.org/pdf/1712.01815.pdf>

<sup>2</sup> Cf. R. Lambert, 11 mars 2018, « Une IA championne de poker ? », sur le site *Pensée Artificielle* : <http://penseeartificielle.fr/ia-poker-libratus-bat-professionnels/>

C'est pourquoi si **1956 fait figure d'année de baptême** de l'intelligence artificielle, **2017 en représente bien l'année de confirmation**.

#### **4. « L'intelligence artificielle n'existe pas » : la question de la « frontière » de l'IA**

L'informatique constitue un **domaine d'application privilégié** des algorithmes. Mais son histoire ne se confond pas avec celle de ces derniers. Il en est de même pour l'histoire de l'intelligence artificielle, bien que ces **trois histoires** soient liées. Alors que **l'informatique** traitait traditionnellement de **questions résolues par des algorithmes connus**, l'intelligence artificielle s'était plutôt intéressée aux problèmes pour lesquels aucun algorithme satisfaisant n'existait encore.

Le paradoxe résultant de cette distinction est le suivant : **dès qu'un problème est résolu par une technologie dite d'intelligence artificielle, l'activité correspondante tend à ne plus être considérée comme une preuve d'intelligence de la machine**. Les cas les plus connus de résolutions de problèmes d'algèbre ou de capacité à jouer à des jeux (des jeux de dame ou d'échecs par exemple) illustrent ce phénomène. Nick Bostrom explique ainsi que *« beaucoup d'intelligence artificielle de pointe a filtré dans des applications générales, sans y être officiellement rattachée car dès que quelque chose devient suffisamment utile et commun, on lui retire l'étiquette d'intelligence artificielle »*.

Les progrès en matière d'intelligence artificielle étant tangibles depuis les années 1950, les frontières de l'intelligence artificielle sont donc sans cesse repoussées et ce qui était appelé intelligence artificielle hier n'est donc plus nécessairement considéré comme tel aujourd'hui. Dès l'origine, **l'intelligence artificielle est bien une étiquette**.

Ce label recouvre en réalité des **technologies très diverses**, dont il n'est pas possible de retracer toute la richesse et la diversité, le présent rapport se contentant d'en présenter les aspects les plus saillants.

**La frontière de l'IA bougeant sans cesse**, les algorithmes relevant de l'IA renvoient à des technologies dont les contenus ne sont pas stables dans le temps. C'est pourquoi certains ont recours à la formule **« l'intelligence artificielle n'existe pas »**, qui est notamment utilisée par Luc Julia, le chercheur à l'origine de Siri, l'IA d'Apple.

L'IA se situe à **la frontière des progrès de l'informatique**. Il s'agit, pourrait-on dire, d'une sorte d'informatique avancée. En effet, **l'acronyme IA devrait en réalité bien davantage renvoyer à de l'informatique avancée qu'à de l'intelligence artificielle en tant que telle**. Un tel concept serait à l'évidence **bien moins anxiogène**.



Dans l'histoire des progrès de l'IA, 2017 est, par exemple, une année importante car **la frontière de l'IA a été encore une fois repoussée**. Quelques mois après l'adoption du rapport de l'Office, un article théorise la nouvelle étape de l'IA, à travers l'architecture algorithmique *Transformer* marquant un **progrès dans l'évolution de l'IA connexionniste**.

Cette dernière est à distinguer de l'IA symbolique. Vos rapporteurs ont, en effet, relevé dans leurs investigations que les outils d'intelligence artificielle sont très divers, mais que les diverses formes d'IA relèvent surtout de deux grandes familles : elles vont de **formes explicites** (systèmes experts et raisonnements logiques au cœur de l'IA symbolique) à des **formes plus implicites** (IA connexionniste telle que les réseaux de neurones et le *Deep Learning*).

## C. LA SUMMA DIVISIO DE L'IA : INTELLIGENCE ARTIFICIELLE SYMBOLIQUE ET INTELLIGENCE ARTIFICIELLE CONNEXIONNISTE

### 1. L'intelligence artificielle symbolique

#### a) Principes généraux des IA symboliques

L'intelligence artificielle symbolique constitue, d'un point de vue historique, la **première grande famille de technologies d'intelligence artificielle**, caractérisée par l'utilisation de symboles et de règles logiques pour résoudre des problèmes au terme de déductions. Elle s'est surtout développée au cours des années 1960, 1970 et 1980.

Si elle existe toujours aujourd'hui, elle est **largement éclipsée par les résultats spectaculaires de l'intelligence artificielle connexionniste**. L'un des outils les plus importants de l'IA symbolique sont les « systèmes experts », appelés aussi systèmes à base de connaissances. Un système expert est un logiciel qui va **extraire des connaissances à partir du savoir des experts humains d'un domaine** et vise à **reproduire les raisonnements pertinents de ces experts face à des cas particuliers**.

Ces systèmes sont composés de trois parties : une **base de faits**, une **base de règles** et un **moteur d'inférence** (si tels faits, alors effectuer telles actions) : grâce aux faits et aux règles fournis en entrées, le modèle va utiliser le moteur d'inférence pour déduire de nouveaux faits et ainsi répondre à une problématique posée. Il s'agit d'un **modèle déductif**. La connaissance est donc décrite sous la forme générale de règles :

« SI Condition (s) » → « ALORS Action (s) »

Ces systèmes analysent une représentation de la situation pour voir quelles règles sont pertinentes, résolvent les éventuels conflits si plusieurs règles s'appliquent et exécutent les actions indiquées en modifiant la situation en conséquence. Ces systèmes sont efficaces dans des domaines restreints mais deviennent difficiles à gérer quand ils doivent manipuler de nombreuses règles ou intervenir dans des domaines complexes, instables ou ouverts.

*b) Quelques illustrations de ces technologies*

Un exemple d'intelligence artificielle symbolique appliqué à la médecine est une IA destinée au diagnostic des maladies infectieuses du sang, le **système expert MYCIN** développé dans les années 1970 par l'université de Stanford<sup>1</sup>. Cette IA spécialisée dans l'analyse de sang aide à identifier des infections bactériennes et propose des traitements.

Il avait pour objectif d'assister les médecins dans leur travail et de connaître des usages concrets. Son corpus de connaissances était constitué d'un ensemble de données médicales tandis que sa base de règles était composée de règles de type « si... et... alors... » ; par exemple de façon simplifiée : « SI le patient a une infection ET l'infection est bactérienne ET la bactérie est un streptocoque ALORS recommander la pénicilline ». Le moteur d'inférences utilisait un raisonnement basé sur cette base de connaissances et de règles et posait des questions au médecin pour l'aider à établir son diagnostic au fur et à mesure des réponses fournies par le médecin. En dépit de ses qualités, ce système expert n'a pas vraiment connu d'applications pratiques.

On comprend donc que l'enjeu de l'invention d'un système d'intelligence artificielle symbolique est de parvenir à trouver des **heuristiques de pensées**, qui permettent de déduire aussi rapidement que possible à partir d'une base de règles donnée, une et une seule réponse cohérente avec les données fournies.

Dans les années 1980, de nombreux systèmes dérivent de l'IA symbolique : la programmation logique, avec l'exemple du système PROGOL ; les arbres de décision, avec l'exemple connu de l'algorithme ID3 (acronyme de *Iterative Dichotomiser 3*) ; l'ingénierie des connaissances ou encore les ontologies<sup>2</sup> qui aboutiront dans les années 2000

---

<sup>1</sup> Shortliffe, Edward H, et Bruce G Buchanan. 1975. « A model of inexact reasoning in medicine ». *Mathematical biosciences* 23(3-4): 351-79.

<sup>2</sup> *En informatique, une ontologie est la modélisation d'un ensemble de données par des concepts et relations issues de connaissances dans un domaine donné (par exemple, géographie, médecine, agriculture, etc.), source : Cnil.*

au « Web sémantique »<sup>1</sup>, proposition originale de Tim Berners-Lee, qui avait été en 1990 l'inventeur du *World Wide Web* (WWW, le Web), des URL, du protocole de communication HTTP et du langage informatique HTML, alors qu'il travaillait à l'organisation européenne pour la recherche nucléaire (Cern).

Les **six modèles d'alignement d'IBM** (*alignment models*) ont dominé le marché des modèles de langage, dans les années 1990 et 2000, notamment pour la traduction automatique. Ils reposaient sur des modèles d'IA symbolique, avant l'émergence des LLM modernes basés sur les réseaux de neurones, bien plus efficaces.

Les IA symboliques peuvent être particulièrement utiles pour **capitaliser les savoirs** au sein d'une organisation. Leurs applications en ingénierie des connaissances sont donc nombreuses. Qu'il s'agisse de **systèmes de planification**, de **graphes**, d'**ontologies** ou de **réseaux sémantiques**, les IA symboliques permettent de modéliser les connaissances d'une organisation ou d'un domaine spécifique de façon systématique.

*c) Des limites sémiotiques qui les éloignent de l'intelligence*

L'expérience de pensée imaginée par John Searle dans un article de la revue *Behavioral and Brain Sciences* en 1980 et connue sous le nom d'expérience de la « chambre chinoise » a démontré **l'incapacité des IA symboliques à comprendre ce qu'elles font**, n'assurant qu'une **exécution mécanique d'instructions**<sup>2</sup>.

Le tableau ci-après décrit cette expérience riche d'enseignements quant aux limitations intrinsèques de ces systèmes d'IA.

---

<sup>1</sup> Le Web sémantique consiste en un Internet intelligent permettant aux ordinateurs de comprendre et de répondre aux demandes de l'utilisateur en fonction du sens de celles-ci. Cette structure de connaissances aurait été accessible aux machines en insérant sur les pages Web des métadonnées lisibles par l'ordinateur. Ces métadonnées étant liées les unes aux autres, le Web sémantique aurait conduit à des usages plus intelligents d'Internet. Des langages ont été conçus à cette fin, comme RDF (Resource Description Framework), OWL (Ontology Web Language) ou XML (eXtensible Markup Language) mais la faisabilité d'une telle technologie d'IA à l'échelle des milliards de pages d'Internet devient de plus en plus difficile. Par exemple, l'ontologie de la terminologie médicale SNOMED CT contient à elle seule 370 000 noms de classes or aucune technologie existante n'a été en mesure d'éliminer tous les doublons du point de vue sémantique. Les applications de ces systèmes resteront donc surtout réservées au monde des bibliothèques, à l'édition ou à des blogs spécialisés. La Bibliothèque nationale de France utilise ainsi en lien avec sa bibliothèque numérique Gallica, un système de Web sémantique à travers le projet « data.bnf.fr », basé sur un réseau de métadonnées sous la forme de triplets RDF contenus dans chaque URL. Depuis 2017, le système data.bnf.fr s'appuie sur le modèle conceptuel de référence au niveau international, dit « IFLA LRM »).

<sup>2</sup> Cf. John Searle, 1980, « Minds, brains, and programs », *Behavioral and Brain Sciences*, vol. 3, n° 3 : <https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A> François Chollet, chercheur en intelligence artificielle et créateur de la bibliothèque Keras, sans commenter l'expérience de la chambre chinoise, a abordé des questions similaires concernant la nature de la compréhension de l'intelligence artificielle et ce que l'on appelle l'intelligence des machines. Dans son article « On the Measure of Intelligence », il proposait en 2019 une définition de l'intelligence basée sur l'efficacité de l'acquisition de compétences, soulignant que la simple exécution de tâches ne reflète pas une véritable compréhension ou intelligence.

**Pourquoi l'IA symbolique n'est pas intelligente :  
l'expérience de la chambre chinoise**

John Searle, dans cet article de 1980, montre qu'une personne qui n'a aucune connaissance du chinois, et qui serait enfermée dans une chambre, est **parfaitement capable de communiquer par écrit en chinois à la condition que soit mis à sa disposition un manuel contenant l'ensemble des règles permettant de répondre à des phrases écrites en chinois.**

Et bien que cette personne n'ait aucune compréhension de la signification des phrases en chinois qu'elle reçoit et qu'elle émet, elle donne l'illusion de comprendre en se contentant de suivre des règles données. Appliquer « bêtement » des règles syntaxiques, comme le font les ordinateurs avec l'IA symbolique, ne suffit donc pas à engendrer une véritable **compréhension sémantique.**

Searle prend alors ses distances avec l'idée au principe du test de Turing, selon laquelle un programme informatique peut être qualifié d'intelligent s'il est capable de communiquer avec un humain sans que ce dernier ne puisse réaliser qu'il s'agit d'une machine.

**Pouvoir faire illusion en reproduisant une langue, sans avoir aucune conscience du contenu communiqué, n'est pas une preuve d'intelligence.** La maîtrise du langage n'est pas qu'une manipulation de symboles, c'est aussi l'entendement des concepts, le fait de **comprendre le sens de ce qu'on dit**, d'avoir conscience du contenu.

Les programmes informatiques sont des **systèmes formels** dont la structure est syntaxique alors que l'intelligence humaine **articule la syntaxe avec des contenus mentaux à caractère sémantique.**

L'IA est donc encore très loin de l'esprit humain, seul capable de **faire l'expérience subjective de la compréhension du monde des choses, du monde des mots et des relations qui unissent ces deux mondes.**

Les IA symboliques sont bien affectées de **limites sémiotiques** : en effet, on sait depuis au moins le linguiste Ferdinand de Saussure, s'inspirant d'études millénaires en sanscrit ainsi que d'Héraclite<sup>1</sup>, qu'un mot est interprété avec un **signifiant** (le symbole ou la représentation mentale de l'aspect matériel du signe), un **signifié** (le concept ou la représentation mentale du contenu associé au signe) et son référent, un **objet** (ou un ensemble d'objets) concret désigné par le signe. Comme l'expliquait Ferdinand de Saussure au début du XX<sup>e</sup> siècle, le signe linguistique unit non pas tant un nom et une chose (le dénoté), non pas un mot et un objet, mais un concept (la connotation) et une image acoustique (le symbole), le signifié et le signifiant. Le structuralisme donnera à cette approche le nom de « triangle sémiotique ».

---

<sup>1</sup> Dans ses *Fragments*, Héraclite, penseur du mouvement niant le principe d'identité, distinguait déjà la pensée (γῶδος), l'énoncé (ἔπος) et la réalité (ἔργον), la liaison entre ces trois éléments étant réalisée selon lui par le logos, principe unique d'essence divine.

Or, l'IA symbolique **ne dispose que du signifiant** auquel elle associe éventuellement un objet mais elle est incapable de prendre en considération le signifié du mot, **les concepts lui restent totalement étrangers**, aussi, elle **manipule les symboles sans avoir aucune idée de ce qu'ils sont**, sans les comprendre pourrait-on dire. L'IA symbolique se heurte donc à trois principaux problèmes : les connaissances, puisqu'il faut être en mesure de décrire le monde pour l'utiliser, l'inférence, puisqu'il faut être capable de recueillir une expertise capable d'extraire des règles, enfin le contrôle, puisque les possibilités, si elles sont trop nombreuses, deviennent impossibles à déterminer.

L'IA connexionniste, avec ses méthodes statistiques qui se rapprochent de la logique inductive, peuvent donner l'impression de se rapprocher davantage de ce que nous appelons communément « compréhension », lorsque nous faisons des raisonnements basés sur l'induction, mais sous une forme purement probabiliste, en restant de simples programmes informatiques basés sur des mathématiques. **De simples programmes mais pas des programmes simples** comme nous allons le voir.

## 2. L'intelligence artificielle connexionniste

### a) Cadre et définition de ces « superstatistiques »

(1) À l'origine de tous ces systèmes : les classifieurs linéaires

Contrairement à l'intelligence artificielle symbolique, déterministe, **l'intelligence artificielle connexionniste ne se base pas sur des règles et de la logique qui seraient codées par le développeur dans des programmes informatiques mais sur des statistiques et de l'analyse probabiliste de données en fonction de variables aléatoires (qui forment des processus dits « stochastiques »).**

On parle parfois de ces technologies connexionnistes comme de « **superstatistiques** », ce qui distingue le *Machine Learning* (apprentissage automatique) des IA symboliques vues précédemment. Elles sont ainsi qualifiées de par leur capacité à traiter de très grandes quantités de données *via* des méthodes statistiques complexes. Ce concept de superstatistiques appliqué à l'intelligence artificielle doit être distingué de celui relevant de la physique statistique<sup>1</sup>.

---

<sup>1</sup> Les superstatistiques au sens strict forment un sous-domaine de la physique statistique et consistent à décrire les propriétés statistiques d'un système par une superposition de statistiques.

L'intelligence artificielle connexionniste est, il est vrai, largement **issue d'algorithmes de classement statistique, au premier rang desquels les classifieurs linéaires** (pouvant eux-mêmes faire figure de sous-catégorie de l'analyse factorielle discriminante, qui peut d'ailleurs autant être descriptive que prédictive). Ces classifieurs ont pour rôle de classer des « objets », c'est-à-dire de les caractériser comme appartenant à des groupes - ou « classes » - déterminés. Un objet est un ensemble de variables numériques (pensons à une plante que l'on décrirait par l'existence ou non d'une tige, la longueur de la tige, la présence ou non d'épines, la présence ou non d'une fleur, le nombre de pétales, un codage pour la couleur des pétales, etc.). Lorsqu'il n'y a que deux classes (par exemple, on cherche à classer des fleurs en « rose » ou « marguerite »), le classifieur linéaire est un instrument très simple : il assigne un poids, un coefficient pondérateur, à chacune des variables de l'objet étudié, fait le produit correspondant puis additionne l'ensemble - il fait une combinaison linéaire des variables, d'où le nom de « classifieur linéaire » ; il applique ensuite à la somme ainsi calculée une « fonction de décision » qui détermine l'appartenance de l'objet à l'une ou l'autre classe ou la probabilité que l'objet appartienne à l'une ou l'autre classe. La plupart des problèmes supposant en fait l'existence de plus de deux classes, on combine plusieurs classifieurs linéaires pour faire de la classification multi-classes, notamment avec les méthodes dites « un-contre-un » et « un-contre-tous ». En tout état de cause, les poids mis en œuvre par un classifieur sont appris à partir d'un jeu de données d'apprentissage étiquetées.

Ces algorithmes reposent sur **des fonctions qui convertissent le produit scalaire de vecteurs dans la sortie désirée selon un vecteur de poids appris à partir d'un ensemble d'apprentissage étiqueté**. Ils peuvent modéliser des probabilités conditionnelles (« classifieurs génératifs », comme la classification bayésienne naïve, à ne pas confondre avec l'IA générative) ou, en vue d'être plus précis, recourir à une méthode discriminante.

**Les « réseaux de neurones artificiels » - synonyme de « réseaux de neurones formels »** - sont un des systèmes d'IA connexionnistes les plus utilisés. Lorsque les réseaux sont organisés de manière stratifiée et que les calculs sont réalisés par plusieurs « couches » de neurones fonctionnant en cascade selon plusieurs niveaux de représentations, la sous-catégorie de l'IA que les réseaux de neurones forment prend le nom d'une sous-sous-catégorie appelée « *Deep Learning* » ou « **apprentissage profond** ».

(2) La pierre angulaire théorique : le théorème d'approximation universelle

Bien que ses succès soient en grande partie empiriques, le paradigme connexionniste de l'intelligence artificielle s'est progressivement doté d'un cadre théorique solide permettant de démontrer sa validité scientifique. Cette branche de l'IA se base ainsi sur le théorème d'approximation universelle, prouvant que **les réseaux de neurones, à partir d'une seule couche cachée, peuvent approximer n'importe quelle fonction continue à la condition que la fonction d'activation soit non linéaire**<sup>1</sup>.

Ce théorème a été prouvé, dans un premier temps, par George Cybenko en 1989 pour certains modèles connexionnistes aux **fonctions d'activation sigmoïdes**<sup>2</sup>. La même année, Kurt Hornik et son équipe ont démontré plus largement que **les réseaux de neurones multicouches sont en réalité des approximateurs universels**<sup>3</sup>. D'autres chercheurs ont ensuite établi, ces dernières années, que cette propriété d'approximation universelle correspond à une **fonction d'activation non polynomiale** et ont étendu le théorème à d'autres fonctions et domaines<sup>4</sup>. De ce fait, dès lors que leur architecture permet d'approximer suffisamment bien la fonction recherchée, les réseaux de neurones permettent de réaliser de nombreuses tâches de classification exigeantes.

*b) Les réseaux de neurones artificiels : aux origines de l'apprentissage profond ou Deep Learning*

(1) Les premières théories dans les années 1940

Dès 1943, le neurologue Warren McCulloch et le psychologue logicien Walter Pitts, travaillant tous les deux sur l'action des neurones dans le cerveau humain, mettent au point **un modèle de « réseau de neurones »** s'inspirant de l'anatomie animale et humaine<sup>5</sup>. Les **neurones humains** sont à cette époque une découverte relativement récente, le terme n'apparaissant d'ailleurs qu'à la fin du XIX<sup>e</sup> siècle grâce au développement du microscope optique.

---

<sup>1</sup> Les notions de « couche cachée » et de « fonction d'activation » sont abordées plus loin.

<sup>2</sup> Cybenko, George. 1989. « Approximation by superpositions of a sigmoidal function ». Mathematics of control, signals and systems.

<sup>3</sup> Hornik, Kurt, Maxwell Stinchcombe, et Halbert White. 1989. « Multilayer feedforward networks are universal approximators » et, 1991, « Approximation capabilities of multilayer feedforward networks », *Revue Neural networks*.

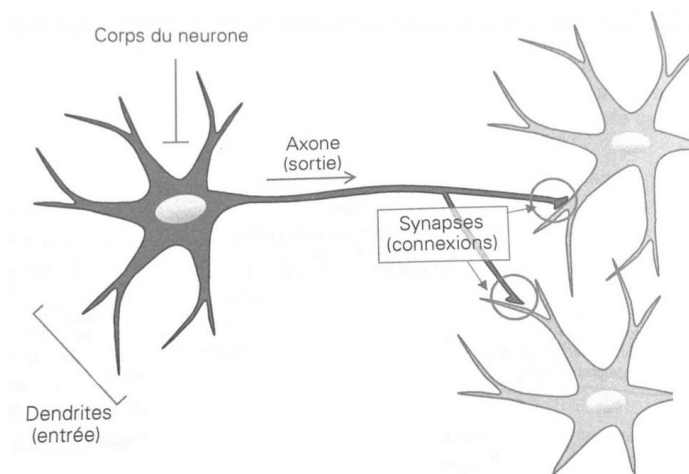
<sup>4</sup> Il existe plusieurs extensions du théorème, notamment aux domaines non compacts, aux architectures de réseaux et topologies alternatives, ou encore à des réseaux certifiables. Cf. Patrick Kidger et Terry Lyons, 2020, « Universal Approximation with Deep Narrow Networks », *Conference on Learning Theory*, Hongzhou Lin et Stefanie Jegelka, 2018, « ResNet with one-neuron hidden layers is a Universal Approximator », *Advances in Neural Information Processing Systems*, Maximilian Baader, Matthew Mirman et Martin Vechev, 2020, « Universal Approximation with Certified Networks », *ICLR*.

<sup>5</sup> McCulloch, Warren S, et Walter Pitts, 1943, « A logical calculus of the ideas immanent in nervous activity », *Bulletin of Mathematical Biophysics*, n° 5, pp. 115-133.

La structure des neurones est déjà **connue dans la première moitié du XX<sup>e</sup> siècle** : des cellules nommées « neurones » transmettent une information à travers des axones, qui se lient aux entrées des autres neurones, appelées « dendrites » par le biais d'une connexion nommée « synapse ».

En revanche, **la façon dont une telle structure était capable de transporter l'information restait totalement inconnue**. C'est dans cette optique que les chercheurs ont d'abord voulu schématiser simplement le fonctionnement des réseaux de neurones humains, sans avoir pour but d'utiliser un tel modèle pour traiter efficacement de l'information, ni *a fortiori* pour faire progresser l'informatique et inventer l'IA.

### Schéma simplifié d'un neurone biologique



*Source : Jean-Paul Haton et al., 2023, Intelligences artificielles : de la théorie à la pratique. Modèles, applications et enjeux des IA, Dunod, p. 60*

Les deux chercheurs ont alors imaginé la **transposition d'une telle configuration dans un mode purement formel**, en s'inspirant des portes logiques en mathématiques.

L'idée est donc qu'un « neurone » artificiel ou formel pourrait accueillir des entrées provenant de neurones d'une couche précédente. Ce « neurone » ferait alors, tel un automate, la somme des entrées de la couche précédente, une somme qui serait pondérée par des « poids » (ces poids miment la plasticité synaptique des réseaux biologiques). Cette somme serait alors soumise à une **fonction d'activation non linéaire** qui, agissant comme un seuil franchi ou non, détermine si le neurone active ou non sa sortie – l'application de cette fonction à la somme des valeurs issues des entrées des couches précédentes permettant ou pas d'atteindre une valeur seuil donnée.

Dans leur article de 1943, McCulloch et Pitts affirment que de tels réseaux pourraient **effectuer des calculs logiques**. Ainsi, ils conçoivent des « portes logiques », c'est-à-dire des opérateurs dont les entrées et les sorties reposent sur la logique booléenne (seules deux valeurs sont possibles, qu'on représente en général par les couples « vrai / faux » ou « 1 / 0 »).



On est encore loin des réseaux de neurones qui seront utilisés plus tard pour développer l'intelligence artificielle connexionniste telle que nous la connaissons mais les auteurs ouvrent un nouveau champ de travail.

Le schéma suivant présente les fonctions logiques que McCulloch et Pitts ont mises au point grâce à leur schématisation des réseaux de neurones, les triangles y représentant les neurones et les flèches y renvoyant aux connexions synaptiques.

**Fonctions logiques selon les réseaux schématiques de neurones définis par McCulloch et Pitts en 1943**

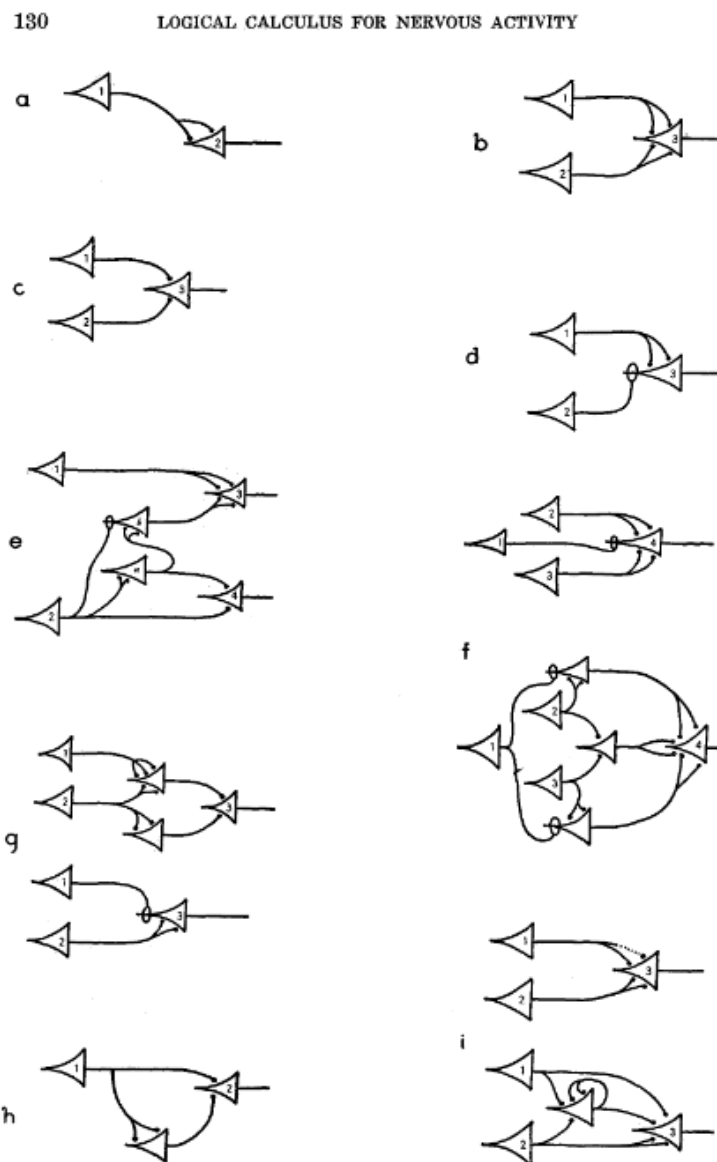


FIGURE 1

Source : McCulloch et Pitts, « A logical calculus of the ideas immanent in nervous activity », op. cit.

Les travaux de McCulloch et Pitts dans les années 1940 ne sont que **théoriques**, ils n'ont pas, à l'époque, l'idée de réaliser physiquement de tels réseaux de neurones artificiels, encore moins de les simuler *via* un programme informatique. Leur objectif est notamment de décrypter le fonctionnement des neurones responsables de la vision chez les humains et les animaux. C'est néanmoins sur la base de ces schémas de neurones formels interconnectés en portes logiques que vont s'appuyer les travaux postérieurs en mathématiques et en informatique pour développer progressivement les réseaux de neurones artificiels.

En 1959, McCulloch et Pitts, aidés de Lettvin et Maturana, se demandent « *ce que l'œil d'une grenouille dit à son cerveau* » et précisent la transposition imaginée en 1943 avec un modèle simplifié de neurone biologique appelé neurone formel<sup>1</sup>. Ces **neurones formels associés en réseau** comparent la somme de leurs entrées et, si une valeur seuil est atteinte, répondent en émettant un signal. Ces réseaux reposent donc sur des **fonctions de transfert**. Et comme dans le cas des réseaux de neurones biologiques, il est imaginé que la force de connexion entre les neurones – l'efficacité de la transmission des signaux d'un neurone à l'autre – peut varier. Pour autant les auteurs ne présentent pas encore un système artificiel capable d'apprendre par l'expérience. C'est un autre chercheur qui va théoriser, sous le nom de « perceptron », les algorithmes d'apprentissage permettant de faire varier la force de connexion entre les neurones artificiels. Avec les perceptrons, le **poids synaptique** au sein des neurones formels va se trouver modifié et amélioré selon des **processus d'apprentissage**.

(2) Les premiers perceptrons

(a) Les perceptrons monocouches

La première apparition d'un modèle pouvant être considéré comme l'ancêtre des réseaux de neurones artificiels actuels plutôt que comme une simple source d'inspiration théorique est le **perceptron monocouche**, inventé au laboratoire d'aéronautique de l'université Cornell par le psychologue Frank Rosenblatt en 1957 et ayant conduit à une publication en 1958<sup>2</sup>. Il permet le classement binaire linéaire supervisé d'une population, c'est-à-dire un processus permettant de séparer une population en deux classes, en connaissant déjà la classe d'une partie des individus. On retrouve ici le principe du classifieur linéaire évoqué dans les développements précédents. Ce réseau est **capable d'apprentissage** : là où les poids synaptiques sont figés dans les réseaux de McCulloch et Pitts, le perceptron, lui, peut faire **varier ses poids** grâce à une règle d'apprentissage du perceptron, appelée aussi « loi de Widrow-Hoff » ou filtre des moindres carrés moyens (*Least Mean Squares* ou LMS).

---

<sup>1</sup> Cf. J. Lettvin, H. Maturana, W. McCulloch et W. Pitts, « *What the Frog's Eye Tells the Frog's Brain* », *Proceedings of the IRE*, vol. 47, n° 11, novembre 1959, pp. 1940-1951.

<sup>2</sup> Frank Rosenblatt, 1958, « *The perceptron : a probabilistic model for information storage and organization in the brain* », *Psychological review*, vol. 65, n° 6.

Le perceptron monocouche est composé de plusieurs entrées et d'une seule sortie (booléenne) à laquelle toutes les entrées (booléennes) sont connectées. Il est utilisé pour résoudre des **problèmes de classification linéaire**, c'est-à-dire des problèmes qui peuvent être résolus en séparant deux classes d'une population par une droite ou un plan, que l'on qualifie spécifiquement d'hyperplan dans le cadre des réseaux de neurones<sup>1</sup>.

Prenons l'exemple d'**une population dont on connaît la taille et le poids et dont on souhaite classer les individus par genre**. Les caractéristiques de la population sont deux variables continues « taille » et « poids » et les deux classes auxquelles les individus peuvent appartenir sont « homme » ou « femme ». On peut représenter les individus par des points situés sur un graphique en deux dimensions qui aurait pour abscisse la taille des individus et pour ordonnée leur poids. Les hommes étant généralement plus grands et massifs que les femmes, le graphique fait apparaître deux groupes de points qui représentent respectivement les hommes et les femmes et sont à peu près séparés l'un de l'autre. L'apprentissage va consister à déterminer la droite qui sépare « le mieux possible » le groupe des points représentant les hommes et celui des points représentant les femmes, pour le jeu de données d'apprentissage choisi (c'est-à-dire un ensemble d'individus dont on connaît la taille et le poids, et dont chacun dispose de son étiquette « homme » ou « femme »). Dès lors on pourra **déterminer la classe probable d'un nouvel individu (en l'occurrence son genre) en connaissant son poids et sa taille**, selon que le point qui le représente sur le graphique sera placé d'un côté ou de l'autre de la droite séparatrice.

Pour passer du langage géométrique au langage algébrique pertinent pour le perceptron :

- la droite séparatrice est entièrement déterminée par son équation cartésienne, concrètement trois nombres  $\omega_1$ ,  $\omega_2$  et  $b$  : elle est l'ensemble des points du plan dont les coordonnées (*taille*, *poids*) vérifient

$$\omega_1 \cdot \text{taille} + \omega_2 \cdot \text{poids} + b = 0$$

- le perceptron est un opérateur qui, pour tout individu, calcule la quantité

$$Y = \omega_1 \cdot \text{taille} + \omega_2 \cdot \text{poids} + b$$

- les coefficients  $\omega_1$  et  $\omega_2$  sont les « poids synaptiques » qui pondèrent les valeurs reçues aux entrées du perceptron ; le coefficient  $b$  est appelé un « biais » du perceptron

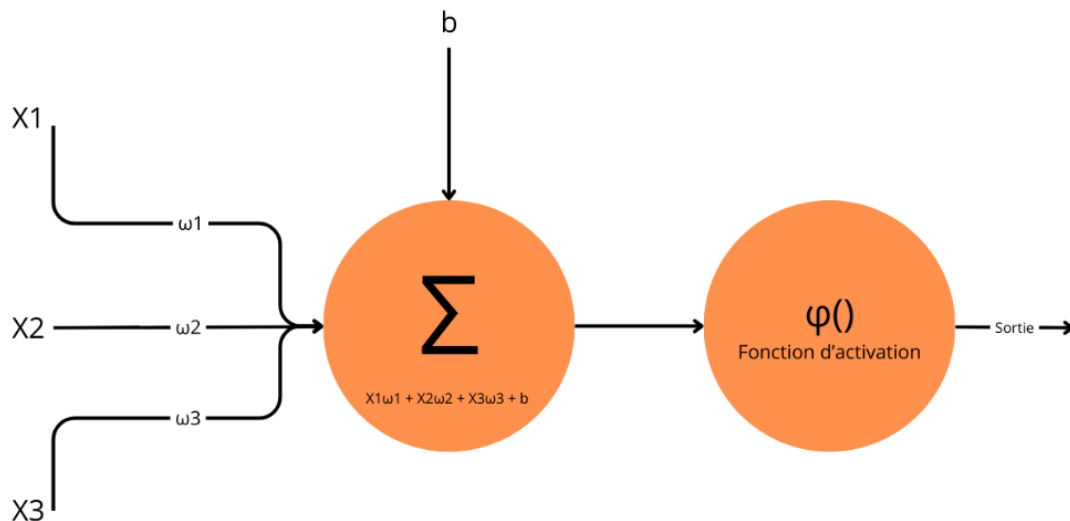
- selon que le résultat  $Y$  est positif ou négatif, l'individu est un homme ou une femme ; si  $Y$  est nul, l'individu ne peut être classé.

---

<sup>1</sup> On parle « d'hyperplan » pour parler d'espaces avec une dimension de moins que l'espace dans lequel on se trouve. En deux dimensions, il s'agit d'une droite (une dimension), en trois dimensions, d'un plan (deux dimensions), on peut étendre le principe aux espaces à plus de dimensions.

Le schéma ci-après illustre un perceptron qui traite trois variables en entrée et dont le résultat  $Y$  est soumis à une fonction d'activation, pour transmission éventuelle au perceptron suivant. Le principe est évidemment généralisable à un nombre supérieur de variables.

**Schéma d'un perceptron monocouche avec fonction d'activation**



Le perceptron dispose donc de la capacité de **séparer une population en deux classes** dont la « frontière » dépend des poids synaptiques  $\omega$ . Toutefois, il manque toujours à ce réseau un système d'apprentissage qui lui permettra de déterminer les poids synaptiques grâce auxquels la population sera séparée de façon optimale. Pour cela, on utilise la loi de Widrow-Hoff.

Le principe consiste à soumettre au perceptron une série d'individus d'entraînement. On détermine l'écart entre le résultat  $Y$  calculé par le perceptron pour le premier individu et le résultat  $Y$  attendu ; les poids synaptiques sont alors ajustés en fonction de cet écart ; on réitère le processus en soumettant au perceptron successivement tous les individus d'entraînement, jusqu'à ce qu'il n'y en ait plus ou que le nombre d'erreurs du perceptron soit devenu acceptable.

Pour que ce processus d'entraînement fonctionne, il faut fixer ce que l'on appelle un « **hyperparamètre** », c'est-à-dire un paramètre qui affecte le comportement d'un modèle et qui est choisi par son concepteur pour son entraînement. Dans le cas de la loi de Widrow-Hoff, l'hyperparamètre est un « **taux d'apprentissage** », qui détermine l'importance de l'ajustement apporté aux poids synaptiques lorsqu'on soumet un individu d'entraînement au perceptron. Une valeur élevée du taux d'apprentissage favorise un apprentissage rapide mais expose à l'apparition d'erreurs plus fréquentes ; au contraire, un taux d'apprentissage faible réduit le risque d'erreur mais ralentit le processus d'apprentissage.

En 1982, le physicien **John Hopfield** cherche à enrichir les réseaux de neurones artificiels de la rigueur propre aux physiciens et ouvre la voie aux modèles dits à base d'énergie : la dynamique du réseau est à temps discret et asynchrone, ce qui signifie qu'un seul neurone est mis à jour à chaque unité de temps<sup>1</sup>. Ces modèles de réseaux de neurones dits de Hopfield vont rapidement se généraliser et actualiser la grande majorité des nouveaux perceptrons.

Bien que le perceptron constitue une avancée importante dans le développement de l'IA connexionniste, son **intérêt reste alors limité** : il ne peut effectuer que des séparations linéaires. Ce type de séparation, bien que permettant certaines classifications, ne correspond pas à la majorité des cas réels, où les classes d'une population donnée sont séparées par des fonctions plus complexes qu'une simple fonction linéaire. Par exemple, ces premiers perceptrons ne sont pas capables de résoudre des problèmes non linéaires comme la disjonction exclusive ou fonction « ou exclusif » (appelée aussi XOR, connue en électricité sous la forme du montage va-et-vient et utilisée en cryptographie, à l'instar du « téléphone rouge » entre la Maison-Blanche et le Kremlin dans les années 1970 et 1980).

Pour résoudre des problèmes non linéaires, une seule couche de réseaux de neurones n'est pas suffisante : **plusieurs couches de réseaux de neurones** vont alors être utilisées. On parle donc de « **perceptron multicouches** » (*multilayer perceptron*, ou **MLP**). Ces perceptrons multicouches ont tout d'abord pris la forme de « **réseaux de neurones à propagation avant** » ou « réseaux de neurones à action directe » (en anglais *feedforward neural networks* (**FNN**)).

(b) Les perceptrons multicouches (MLP) et les réseaux de neurones à propagation avant (FNN)

Dans le livre *Perceptrons : une introduction à la géométrie informatique*, de Marvin Minsky et Seymour Papert publié en 1969, les deux auteurs estiment qu'imiter le cerveau est trop complexe pour des machines et démontrent **l'incapacité des perceptrons monocouches à résoudre des problèmes de classification non linéaires**<sup>2</sup>. Ce livre, pessimiste, est accusé d'être à l'origine d'un premier « Hiver » de l'intelligence artificielle.

---

<sup>1</sup> Cf. John J. Hopfield, 1982, « Neural networks and physical systems with emergent collective computational abilities », *Proceedings of the National Academy of Sciences*, vol. 79, n° 8 : <https://www.pnas.org/doi/10.1073/pnas.79.8.2554>

<sup>2</sup> Cf. Marvin L Minsky. et Seymour A. Papert, 1969, *Perceptrons : an introduction to computational geometry*. éd. Cambridge MIT Press.

La sortie de ce livre est en effet corrélée avec une période de relative accalmie dans le développement et le financement de l'intelligence artificielle et par un quasi-abandon des perceptrons et plus généralement des réseaux de neurones artificiels<sup>1</sup>. Elle n'est pas pour autant la seule raison de ce ralentissement : les limitations des technologies alors disponibles, les données en nombre insuffisant et le manque de puissance de calcul sont les trois facteurs principaux. La recherche autour de l'intelligence artificielle connexionniste a été ranimée dans les années 1980 par les réseaux de Hopfield et surtout par les MLP.

On parle de perceptrons multicouches lorsque le réseau de neurones est composé de **perceptrons organisés en plusieurs couches**. Lorsque l'information n'y circule que dans un sens, de l'entrée vers la sortie, on a affaire à des « réseaux de neurones à propagation avant » (en anglais *feedforward neural network*, FNN), pour les distinguer des réseaux de neurones récurrents (RNN), où l'information effectue au moins un cycle dans la structure du réseau (ces réseaux plus complexes seront vus plus loin).

Dans les MLP, chaque couche agit de la même façon qu'un perceptron classique, et les différentes couches sont montées « en série ». La première couche est appelée **couche d'entrée**. C'est ici que sont introduites les données que l'on veut traiter. La couche d'entrée transforme ces données en données numériques pour qu'elles puissent être traitées par le réseau.

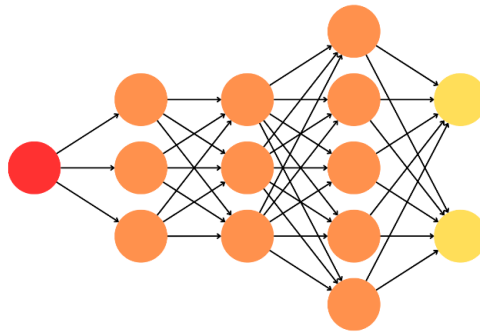
Ensuite, il y a **une ou plusieurs couches cachées**. Chaque couche est composée de neurones, et chaque neurone a une ou plusieurs entrées et sorties. Ces entrées et ces sorties forment des ensembles fonctionnels qui se comportent chacun comme des perceptrons. Ils associent à une valeur des coefficients (les poids synaptiques) et un biais. Comme pour le perceptron monocouche, chaque entrée reçoit une valeur de la couche précédente. Cette valeur est multipliée par un autre élément appelé « poids synaptique », qui définit la force du lien entre deux neurones. Si un neurone a plusieurs entrées, toutes les valeurs sont additionnées, et on ajoute une autre valeur pour chaque neurone appelé « biais ». De la même façon que pour un perceptron monocouche, cette somme est ensuite passée dans une fonction d'activation, qui décide si la sortie doit être activée ou non. Il existe différentes fonctions d'activation, comme la fonction « marche » ou la fonction « unité de rectification linéaire » dite ReLU. Le résultat de la fonction est la sortie du neurone, et devient une entrée pour la couche suivante.

---

<sup>1</sup> La notion d'alternance entre « Hivers » et « Printemps » de l'IA a été plus amplement développée dans le rapport de l'OPECST écrit par M. le député Claude de Ganay et Mme la sénatrice Dominique Gillot : Pour une intelligence artificielle maîtrisée, utile et démystifiée disponible sur le site du Sénat : <https://www.senat.fr/rap/r16-464-1/r16-464-11.pdf> et de l'Assemblée nationale : [https://www2.assemblee-nationale.fr/documents/notice/14/rap-off/i4594/\(index\)/index-thematique-oecest](https://www2.assemblee-nationale.fr/documents/notice/14/rap-off/i4594/(index)/index-thematique-oecest)

Enfin, il y a la **couche de sortie**, qui, de la même façon qu'un perceptron monocouche, transforme les valeurs obtenues en réponse au problème posé. Par exemple, si l'on souhaite savoir s'il s'agit d'un chat ou un chien sur une image, la couche de sortie donne la réponse grâce à un neurone correspondant à la probabilité qu'il y ait un chien sur l'image et un autre neurone correspondant à la probabilité qu'il y ait un chat sur l'image.

### Schéma d'un perceptron multicouche

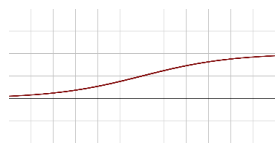
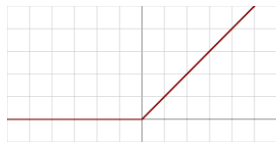
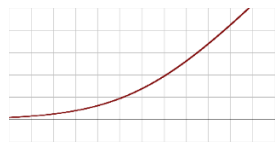
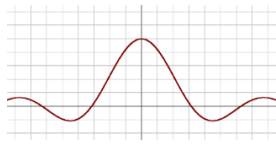


Légende : la couche d'entrée est en rouge, les couches cachées en orange et la couche de sortie en jaune

Ce schéma permet de voir concrètement les trois couches qui composent le modèle. À **gauche**, on a une **couche d'entrée**, ici constituée d'un neurone unique mais en fonction du type de données en entrée, on pourrait en avoir plusieurs. Au **milieu**, on a les **couches cachées**. Ici il y en a trois, il pourrait y en avoir plus : ce paramètre entre en jeu dans l'élaboration d'un modèle efficace, aussi économique et fiable que possible. Chaque couche n'est pas obligée de contenir le même nombre de neurones : dans l'exemple, les deux premières couches contiennent trois neurones, la troisième en contient cinq. Puisque dans cet exemple tous les neurones d'une couche sont reliés à tous les neurones de la couche suivante, on parle de « réseau dense ». À **droite**, enfin, on a la **couche de sortie**, qui contient ici deux neurones mais qui pourrait en contenir plus ou moins en fonction du type de données que l'on veut en sortie.

### Tableau non exhaustif de fonctions d'activation couramment utilisées

Nom de la fonction	Équation associée	Représentation graphique
Identité	$f(x) = x$	
Marche	$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases}$	

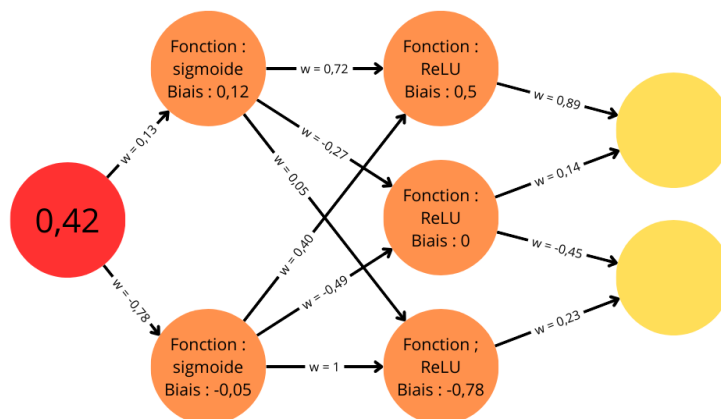
Nom de la fonction	Équation associée	Représentation graphique
Sigmoïde	$f(x) = \frac{1}{1 + e^{-x}}$	
Unité de rectification linéaire (ReLU)	$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } x \geq 0 \end{cases}$	
Unité de rectification linéaire douce (SoftPlus)	$f(x) = \ln(1 + e^x)$	
Sinus cardinal	$f(x) = \begin{cases} \frac{\sin(x)}{x} & \text{si } x \neq 0 \\ 1 & \text{si } x = 0 \end{cases}$	

Source : exemples pris parmi la liste des 16 fonctions d'activation usuelles recensées par Wikipédia

Il est possible de prendre l'exemple concret d'un réseau multicouche simple pour expliquer le fonctionnement de ces réseaux de neurones artificiels.

Ce modèle d'illustration est inventé afin de permettre une compréhension plus aisée du concept. Les réseaux de neurones sont, en effet, beaucoup plus grands, et les informations traitées bien plus nombreuses (plusieurs milliards). Le chemin de l'information va de l'entrée du modèle vers sa sortie.

**Exemple de schéma d'un réseau de neurones avec des valeurs associées aux synapses (poids, noté w) et aux neurones (biais)**





Sur ce schéma sont présentées toutes les informations qui permettent de calculer des valeurs de sorties dans les neurones de la couche de sortie. On dispose d'une valeur d'entrée dans le neurone d'entrée : 0,42. À chaque synapse est associé un poids synaptique noté « w ». Pour chaque couche de neurones cachés, on a une fonction d'activation : sigmoïde pour la première, ReLU pour la seconde. À chaque neurone est associé un biais.

Le calcul pour les deux neurones de la couche cachée peut être détaillé. Pour cela, la valeur initiale est multipliée par le poids synaptique de la synapse qui la relie au neurone dont on veut connaître la valeur, et on y ajoute le biais. On a donc :

$$0,42 \times 0,13 + 0,12 = 0,1746$$

et

$$0,42 \times (-0,78) + (-0,05) = -0,3776$$

Ces valeurs passent ensuite dans la couche d'activation afin d'obtenir le poids final de chaque neurone. On calcule donc la valeur de la fonction sigmoïde pour  $x = 0,1746$  et  $x = 0,3776$ . On obtient 0,54 pour le premier neurone et 0,41 pour le second neurone. La seconde couche de neurones ayant une valeur, on peut alors calculer grâce à celle-ci la troisième couche comme on l'a fait avec la première. Tous les neurones de la deuxième couche étant reliés à tous les neurones de la troisième couche, le calcul est un peu plus complexe puisqu'il faut additionner les valeurs des entrées avant de les faire passer dans la fonction d'activation.

On a alors :

$$0,54 \times 0,72 + 0,41 \times 0,40 + 0,50 = 1,0528$$

$$0,54 \times (-0,27) + 0,41 \times (-0,49) + 0 = -0,3467$$

$$0,54 \times 0,05 + 0,41 \times 1 + (-0,78) = -0,343$$

Ces valeurs peuvent alors passer dans la fonction d'activation ReLU. On obtient donc les poids des trois synapses de la couche suivante : 1,05 ; 0 et 0 (lorsque les nombres sont négatifs, la fonction ReLU renvoie toujours 0).

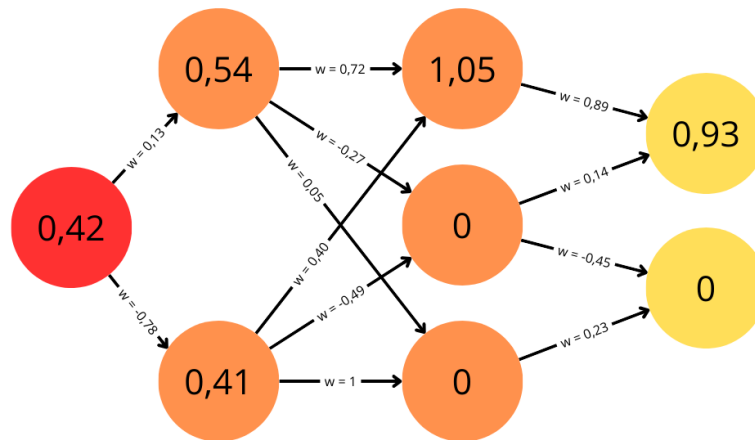
Enfin, les poids des neurones de la couche de sortie sont calculés :

$$1,05 \times 0,89 + 0 \times 0,14 = 0,93$$

$$0 \times (-0,45) + 0 \times 0,23 = 0$$

Le schéma initial du réseau de neurones peut être complété avec toutes les valeurs qui viennent d'être calculées.

Schéma complété des calculs réalisés



On a donc un résultat dans la couche de sortie qui nous permet d'interpréter ce que le modèle a pu discriminer. En réalité, ces calculs ne sont jamais effectués manuellement et un ordinateur peut réaliser tous ces calculs beaucoup plus rapidement qu'un humain pour en tirer des résultats beaucoup plus précis car gardant le plus de chiffres possible après la virgule dans les calculs.

**C'est la multiplication des couches de perceptrons qui a conduit à parler d'apprentissage profond ou *Deep Learning*.** Ce type de réseau découle du perceptron monocouche imaginé par Rosenblatt lors de ses travaux. L'idée de mettre des perceptrons en couches successives ne constitue cependant pas l'avancée majeure des MLP par rapport aux perceptrons monocouches. Ces perceptrons multicouches à propagation avant ou FNN présentent en effet un désavantage par rapport à un perceptron monocouche. Bien que les réseaux à plusieurs couches soient capables de traiter des situations de classification non linéaire, avec plusieurs neurones au lieu d'un seul, **il n'est plus possible de calculer la façon dont on doit entraîner le réseau pour qu'il produise de meilleurs résultats.** Là où l'on utilisait la loi Widrow-Hoff pour le perceptron monocouche, il n'est plus possible de le faire désormais car il y a plusieurs neurones dont les paramètres influencent le résultat final d'une façon différente. Comment alors savoir quel poids synaptique ou quel biais modifier et de quelle façon pour atteindre le résultat optimal voulu ? L'incapacité à répondre à de telles questions est l'une des raisons de « l'Hiver » de l'IA des années 1970. En effet, ces réseaux FNN ne présentent pas d'intérêt s'ils ne sont pas capables de répondre à ces questions puisque leur intérêt était précisément d'être capables « d'apprendre », c'est-à-dire d'adapter leurs réponses.

Des scientifiques commencent à répondre à cette difficulté dans les années 1980 en utilisant une **technique issue de la résolution des problèmes de fonctions convexes en mathématiques** appelée la « descente de gradient ». L'algorithme permettant de trouver un minimum global, c'est-à-dire le point où le modèle obtient les meilleurs résultats, va prendre le nom de **rétropropagation du gradient**, permettant alors d'utiliser les perceptrons multicouches (MLP) de façon optimale et quasi automatisée.

### (3) La rétropropagation du gradient (*Back-propagation*)

Les méthodes algorithmiques visant à corriger les erreurs des MLP et à les améliorer sont issues du calcul de la descente du gradient et sont appelées **algorithmes de rétropropagation du gradient ou de rétropropagation des erreurs**. Elles se fondent sur les travaux de Seppo Linnainmaa, qui décrit de tels algorithmes dès son mémoire de maîtrise en 1970 mais sans les appliquer aux réseaux de neurones. Paul Werbos imagine dans sa thèse de doctorat en 1974 de nouveaux outils de prévision qui appliqueraient la rétropropagation du gradient aux réseaux de neurones. Ces premiers travaux ne conduisent pas à des résultats concrets. On considère en effet que le premier perceptron multicouches efficace date de 1986.

Pour régler les perceptrons multicouches, c'est-à-dire minimiser leur taux d'erreur, la descente du gradient doit conduire à **ajuster progressivement le poids de tous les neurones au sein du modèle**. Un collectif de chercheurs (Hinton, Rumelhart, Williams et McClelland) élabore alors un algorithme de descente du gradient qui va prendre le nom de rétropropagation du gradient<sup>1</sup> : il s'agit de **parcourir le réseau de neurones dans le sens inverse de son fonctionnement pour corriger ses erreurs en mettant à jour par cet algorithme les poids des neurones de la dernière couche à la première**.

De jeunes chercheurs comme Yann LeCun et Yoshua Bengio, rencontrés par vos rapporteurs, ont alors poursuivi à partir du milieu des années 1980 des recherches sur ces nouvelles architectures<sup>2</sup>. En 2019, Geoffrey Hinton, Yann LeCun et Yoshua Bengio ont reçu le prestigieux **prix Turing** pour l'ensemble de ces travaux fondateurs pour les architectures modernes des réseaux de neurones profonds (ce prix est l'équivalent pour l'informatique du prix Nobel ou de la médaille Fields, récompense la plus prestigieuse de la discipline mathématique).

---

<sup>1</sup> Cf. ces deux articles fondateurs : David Rumelhart, Geoffrey Hinton et Ronald Williams, 1986, « Learning representations by back-propagating errors », *Nature*, vol. 323 ; David Rumelhart, Geoffrey Hinton, et James McClelland, 1986, « A general framework for parallel distributed processing » in David Rumelhart et James McClelland, *Parallel distributed processing: Explorations in the microstructure of cognition*, MIT Press.

<sup>2</sup> C'est l'objet même de la thèse de Yann LeCun et l'un des enjeux de celle de Yoshua Bengio. Cf. Yann LeCun, 1987, « Modèles connexionnistes de l'apprentissage », thèse de doctorat, Université Paris VI, Paris, ainsi que plus spécifiquement Yann LeCun, 1988, « A Theoretical Framework for Back-Propagation », *Proceedings of the 1988 Connectionist Models Summer School*, Carnegie Mellon University, Pittsburg. Yann LeCun a également introduit les réseaux neuronaux convolutifs (CNN). Et pour Yoshua Bengio: Yoshua Bengio, 1991, « Artificial neural networks and their application to sequence recognition », *Mc Gill University, Montréal*, ainsi que plus spécifiquement Yoshua Bengio et al., 1994, « Learning long-term dependencies with gradient descent is difficult », *IEEE Transactions on Neural Networks*, volume 5, n° 2.

Il est intéressant de relever qu'un **phénomène biologique** similaire à la rétropropagation du gradient a été observé dans les **réseaux de neurones des mammifères**<sup>1</sup>.

Un des intérêts des réseaux de neurones multicouches est que le **modèle évolue grâce à un entraînement qui lui permet d'être de plus en plus performant**. Celui-ci consiste, en utilisant un jeu de données d'entraînement, à **mesurer l'écart** entre la réponse fournie par le modèle et la réponse attendue, et à **ajuster le modèle** pour minimiser cet écart.

S'agissant du premier point, l'écart entre le résultat effectif et le résultat attendu peut se calculer de différentes façons. La plus simple consiste à calculer l'« **erreur quadratique moyenne** » (*mean square error*, MSE), c'est-à-dire la différence entre la donnée de sortie du modèle et la donnée attendue au carré, permettant d'obtenir une « fonction de perte » ou « fonction de coût ». Si l'on prend l'exemple d'un modèle qui doit déterminer si une image est, ou non, celle d'un chat, et que l'on introduit une image de chat dans le modèle puis que celui-ci donne 0,8 en sortie, la perte sera de  $(1 - 0,8)^2 = 0,04$ . En effet, la sortie attendue était 1 (car l'image est bien celle d'un chat).

Pour ce qui concerne le second point, l'ajustement consiste à « surfer » sur la fonction de perte pour descendre vers son minimum, en faisant varier les poids et les biais des neurones. À cette fin, on détermine la façon dont de petites variations de chaque poids ou biais, réalisées séparément, font varier la fonction de perte ; ceci revient, en mathématiques, à calculer les dérivées partielles de la fonction de perte par rapport à chaque poids ou biais. **Le vecteur contenant l'ensemble des valeurs de ces dérivées est appelé le gradient**.

Pour régler les poids et biais des neurones, on commence à travailler sur la fonction de perte à partir des neurones de la dernière couche. Un algorithme - dit optimiseur - est appliqué au gradient pour le « descendre » en ajustant les poids et les biais dans la direction opposée au gradient. Chaque poids et biais de chaque neurone reçoit une correction qui fait intervenir la dérivée partielle correspondante de la fonction de perte et un taux d'apprentissage. Le choix de l'optimiseur dépend de la nature des données traitées et de la rapidité (souhaitée et possible) de la convergence vers la perte minimale ; il en existe un grand nombre : descente de gradient stochastique (*stochastic gradient descent* ou SGD), Adagrad, Adadelata, RMSprop, Adam... La **rétropropagation** proprement dite consiste à réaliser ensuite les mêmes opérations aux neurones de la couche située immédiatement en amont, puis de la précédente, etc. De cette façon, toutes les couches du modèle sont remontées une par une, depuis la couche de sortie jusqu'à la couche d'entrée.

---

<sup>1</sup> La rétropropagation neuronale désigne la propagation d'un potentiel d'action dans un neurone, non pas vers la terminaison de l'axone (propagation normale), mais au rebours, en direction des dendrites, d'où provenait la dépolarisation originelle. Cf. Greg Stuart, Nelson Spruston, Bert Sakmann et Michael Häusser, 1997, « Action potential initiation and backpropagation in neurons of the mammalian CNS », Trends in Neurosciences, vol. 20, n° 3 : <https://pubmed.ncbi.nlm.nih.gov/9061867/>

Cette méthode de rétropropagation du gradient a suscité un **regain d'intérêt pour les réseaux de neurones** sous leur nouvelle forme d'apprentissage profond, et donc pour l'ensemble du paradigme connexionniste. Au-delà du prix Turing décerné en 2019 à trois de ses inventeurs - Geoffrey Hinton, Yann LeCun et Yoshua Bengio<sup>1</sup> - c'est le **prix Nobel de physique** qui récompense en 2024 Geoffrey Hinton et John Hopfield pour leurs travaux sur les réseaux de neurones artificiels.

#### (4) Les réseaux de neurones convolutifs (CNN)

Un réseau de neurones artificiels traite en entrée des données numériques et délivre en sortie d'autres données numériques. Pour mettre en œuvre de tels modèles, il faut donc pouvoir **transformer l'information que l'on veut traiter en données numériques**, en perdant le moins de sens possible. Dans le cas d'une image en couleur, il s'agit d'identifier les formes, les contrastes, éventuellement les ombres, etc. Les chercheurs ont dû inventer des réseaux permettant de conserver ces informations et de les convertir en données numériques pour qu'elles soient traitées par un réseau, capable de réaliser par exemple des tâches de classification.

Les réseaux convolutifs (*convolutional neural networks* ou CNN) sont ainsi des réseaux de neurones utilisés pour le traitement des images. **Tout comme les travaux sur les perceptrons, ces réseaux de neurones sont influencés par des avancées scientifiques, notamment dans les sciences cognitives.** Les travaux de Hubel et Wiesel en 1968 sur les cellules visuelles dans le cerveau des animaux<sup>2</sup>, qui leur a valu un prix Nobel de physiologie en 1981, ont inspiré les informaticiens sur la façon dont il était possible de traiter une information visuelle grâce à un réseau de neurones situés dans le « cortex strié » (ou cortex visuel primaire). Hubel et Wiesel ont également découvert qu'il existait deux types de cellules dans ce cortex, une partie de ces neurones ne traitant l'information que d'une partie de l'image perçue par les capteurs visuels.

C'est cette idée qui a inspiré le scientifique japonais Kunihiko Fukushima lors de la création de deux modèles : le cognitron en 1975<sup>3</sup> puis le neocognitron en 1980<sup>4</sup>. Ces réseaux, imitant le fonctionnement du cerveau, ne sont pas encore des réseaux convolutifs mais s'en rapprochent et sont souvent considérés comme les réseaux qui ont permis l'émergence des CNN.

---

<sup>1</sup> Cf. Yann LeCun, Yoshua Bengio, et Geoffrey Hinton, 2015, « Deep Learning », numéro spécial de la revue Nature.

<sup>2</sup> Hubel, David H, et Torsten N Wiesel. 1968. « Receptive fields and functional architecture of monkey striate cortex ». The Journal of physiology 195(1): 215-43.

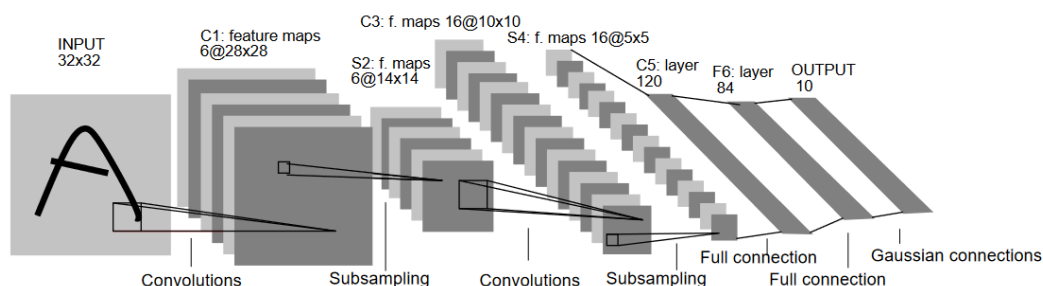
<sup>3</sup> Kunihiko Fukushima, 1975, « Cognitron: A self-organizing multilayered neural network », Biological cybernetics, 20(3): 121-36.

<sup>4</sup> Kunihiko Fukushima, 1980, « Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position », Biological cybernetics 36(4): 193-202.

Yann LeCun écrit : « Les cellules simples du Néocognitron sont un bricolage byzantin pour coller au mieux à la biologie et faire en sorte que le réseau fonctionne. [...] Peut-être Fukushima veut-il imiter trop étroitement la biologie ? Toujours est-il que le résultat est moyennement heureux. »<sup>1</sup>

L'un des premiers réseaux convolutifs à recevoir un usage pratique s'appelle « LeNet-5 » et est développé notamment par Yann LeCun et Yoshua Bengio dans l'entreprise Bell Labs en 1989<sup>2</sup>. L'objectif du modèle était alors de **reconnaître automatiquement les codes postaux manuscrits**. Il sera largement utilisé.

### Schéma du réseau de convolution LeNet-5



Source : LeCun et al., 1989

LeNet est le **premier réseau convolutif** à présenter cette architecture toujours utilisée à l'époque actuelle dans les réseaux convolutifs comme les réseaux ResNet<sup>3</sup>. On voit sur le schéma qu'il est constitué d'une succession de couches de « convolution » et de « sous-échantillonnage » (*subsampling*). L'objectif est de réaliser un « encodage » de l'image d'entrée, afin d'en extraire les caractéristiques importantes.

Pour obtenir des valeurs numériques à partir d'une image, **on attribue des valeurs à chacun de ses pixels**. Dans le cas d'une image en niveaux de gris (noir et blanc), chaque pixel a une valeur, qui va de zéro, totalement noir, à un, totalement blanc. Dans le cas d'une image en couleur, chaque pixel est composé d'un vecteur de trois valeurs. En effet, une image en couleur peut être représentée comme une superposition de la même image respectivement en rouge, vert et bleu (on parle de système RVB ou RGB pour *red green blue* en anglais). Ces couleurs étant les couleurs primaires de la lumière, on peut obtenir l'importe quelle couleur à partir d'un mélange de ces trois couleurs. Par exemple, le jaune étant un mélange de rouge et vert, un pixel jaune sera représenté par le vecteur (1, 1, 0), c'est-à-dire tout à fait rouge et tout à fait vert sans bleu. Les valeurs RGB s'exprimant généralement jusqu'à 255, on normalise ces valeurs dans les modèles pour qu'elles soient contenues entre zéro et un ; les valeurs sont donc divisées par 255.

<sup>1</sup> Yann LeCun, 2019, Quand la machine apprend : La révolution des neurones artificiels et de l'apprentissage profond, Odile Jacob. pp. 203-204.

<sup>2</sup> Yann LeCun, Léon Bottou, Yoshua Bengio, et Patrick Haffner, 1998, « Gradient-based learning applied to document recognition », Proceedings of the IEEE.

<sup>3</sup> Kaiming He, Zhang Xiangyu, Ren Shaoqing et Sun Jian, 2016, « Deep residual learning for image recognition ».

Dans un CNN, l'objectif est de **réduire les dimensions de l'image en gardant ses principales caractéristiques grâce à une opération mathématique appelée la « convolution »**. Cette opération permet d'extraire des caractéristiques d'une image en faisant passer une « matrice de convolution » (*kernel*), qui réalise des opérations sur une zone de l'image, permettant de générer une « carte de caractéristiques » (*activation map*). On peut voir l'opération de convolution comme un tampon qui passerait au-dessus de chaque groupe de pixels dans une image et qui transformerait la valeur de ces pixels en fonction du tampon choisi.

On réalise cette opération plusieurs fois sur chaque couche d'une image (une couche ne pouvant être composée que d'un canal de couleur rouge, verte ou bleue) en changeant la matrice de convolution pour obtenir une série de cartes de caractéristiques de mêmes dimensions que l'image originale. Pour que le processus ne soit pas linéaire, on peut ajouter, après chaque couche de convolution, une fonction d'activation (souvent la fonction sigmoïde ou unité linéaire rectifiée, ReLU), cette fonction va prendre en entrée les cartes de caractéristique obtenues après application d'une convolution. Elle va renvoyer une valeur de sortie selon la fonction choisie<sup>1</sup>.

L'objectif étant de réduire les dimensions de l'image, on applique une couche dite de *max pooling* (en français plus rarement : sous-échantillonnage par valeur maximale). Cette étape consiste à appliquer un filtre d'une taille définie sur une surface (en l'occurrence, sur les cartes de caractéristiques), et ne garder que la valeur maximale de la région définie par le filtre. Ainsi par exemple, un *max pooling* de 2x2 va parcourir les cartes de caractéristiques avec un filtre de deux pixels par deux, et sélectionner dans ce carré la valeur la plus importante. Cela permet de diviser la dimension des surfaces par deux. C'est ce qui est appelé la phase de sous-échantillonnage (*subsampling*).

Cette combinaison de convolution et *max pooling* est répétée sur les cartes de caractéristiques jusqu'à ce que l'on obtienne des dimensions suffisamment réduites. On réalise ensuite un « **aplatissement** » (*flattening*) des dimensions obtenues pour les stocker dans un **vecteur à une seule dimension**. Ce vecteur à une seule dimension est alors présenté à l'entrée d'un réseau dense de neurones, c'est-à-dire un réseau dont l'ensemble des neurones d'une couche sont reliés à l'ensemble des neurones de la couche précédente.

Ce réseau dense, qui fonctionne exactement comme un FNN, avec une couche de sortie construite selon le résultat souhaité.

Pour illustrer le fonctionnement d'un CNN, on peut imaginer utiliser en entrée une image carrée de dimension  $64 \times 64 \times 3$  (soixante-quatre pixels de longueur et de largeur, trois canaux de couleurs) afin de reconnaître un chiffre entre zéro et neuf. Cette image passe ensuite dans une couche de convolution qui va, par exemple, produire trente-deux cartes de caractéristiques en utilisant des matrices de convolution différentes.

---

<sup>1</sup> Voir le tableau des fonctions d'activation pour une liste non exhaustive des fonctions d'activation qu'il est possible d'utiliser.

On se retrouve alors avec une image de dimension  $64 \times 64 \times 32$  (on garde les dimensions d'origine mais on a désormais trente-deux cartes de caractéristiques). On passe alors ces cartes de caractéristique dans une fonction d'activation, par exemple la fonction *ReLU*, puis le résultat est utilisé dans un *max pooling* qui va réduire la dimensionnalité de ces cartes. En utilisant un filtre de *max pooling* de deux pixels par deux, cette couche produit un résultat de dimension  $32 \times 32 \times 32$  (on garde le même nombre de cartes de caractéristiques mais on divise leurs dimensions par deux). On peut alors répéter cette opération jusqu'à obtenir des dimensions de cartes suffisamment réduites, quitte à augmenter le nombre de cartes, par exemple  $8 \times 8 \times 128$ . Les cartes sont ensuite « aplaties », lors de la phase de *flattening*, pour obtenir un vecteur unidirectionnel contenant  $8 \times 8 \times 128 = 8192$  valeurs. Ce vecteur est présenté à un FNN dense qui a dix neurones de sortie, un par chiffre possible de zéro à neuf. Le réseau complet est alors capable d'attribuer un chiffre à chaque image. On peut comparer le résultat obtenu avec le résultat attendu et procéder à une opération de rétropropagation du gradient pour optimiser les résultats.

Les CNN sont **utiles pour la reconnaissance d'images**, mais ils constituent également la première partie des « auto-encodeurs variationnels », modèles génératifs développés plus loin dans ce rapport. L'architecture U-Net, par exemple, est un CNN utilisé dans le monde médical pour la **segmentation d'images cérébrales ou hépatiques** ainsi que dans la création de contenus visuels, notamment par l'application **Stable Diffusion**.

#### (5) Les réseaux de neurones récurrents (RNN)

Dans les FNN, l'information ne circule que dans un sens, de l'entrée vers la sortie, comme il a été vu. De ce fait, chaque valeur d'entrée ne passe qu'une seule fois dans chaque couche de neurones du réseau. Ces réseaux ne peuvent pas traiter une séquence d'informations, c'est-à-dire des données provenant d'une série temporelle cohérente. On parle de données « dynamiques » en opposition aux données « statiques ». Ces données dynamiques englobent les vidéos, les graphiques d'évolution du cours d'actions en bourse, des relevés météorologiques, etc. Pour traiter ce type de données, il faut **utiliser une architecture qui permette à des informations de « remonter » les couches du réseau**. Les bases théoriques de ces types de réseaux ont été posées en 1972 par le japonais Shun'ichi Amari, on les nomme « **réseaux de neurones récurrents** » (en anglais *recurrent neural networks* ou **RNN**)<sup>1</sup>. John Hopfield a été le premier à concevoir un tel réseau récurrent capable de traiter une information dynamique en 1982<sup>2</sup>. Si l'information effectue au moins un cycle dans la structure du réseau, on a affaire à un RNN et plus à un FNN.

---

<sup>1</sup> S-I Amari, 1972, « Learning patterns and pattern sequences by self-organizing nets of threshold elements », IEEE Transactions on computers.

<sup>2</sup> John J. Hopfield, 1982, « Neural networks and physical systems with emergent collective computational abilities », Proceedings of the National Academy of Sciences.



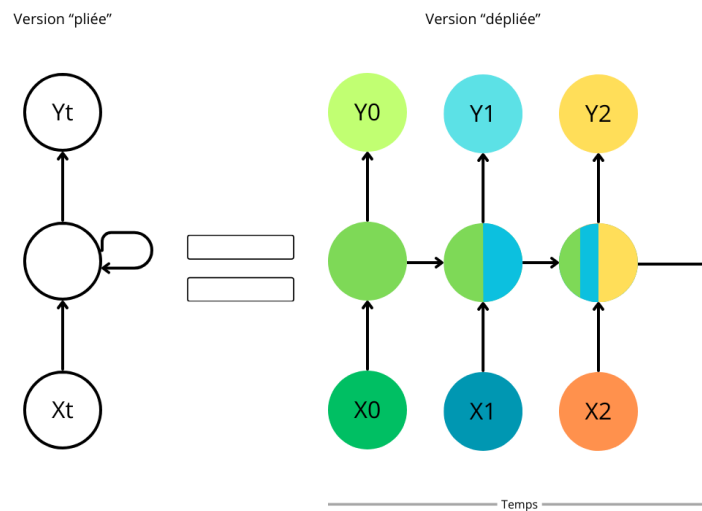
Un RNN ressemble à un réseau à propagation avant : il contient une couche d'entrée, des couches cachées et des couches de sortie. Il a toutefois la particularité de posséder une **boucle de rétroaction** : lorsqu'on lui présente successivement les données d'une séquence, le résultat obtenu pour une donnée prend en compte les résultats obtenus pour les données précédentes. Par exemple, si l'on utilise un RNN pour réaliser des prévisions météorologiques du lundi au dimanche, alors la boucle va permettre de prendre en compte le temps de lundi pour prédire le temps de mardi, le temps de lundi et mardi pour prédire celui de mercredi, etc.

On peut représenter un tel réseau de deux façons :

- selon un schéma « plié » du réseau, qui montre son fonctionnement de façon synthétique. Pour chaque entrée  $X_t$ , le réseau renvoie une sortie  $Y_t$  ainsi qu'un certain nombre de paramètres qui vont être utilisés dans la couche cachée qui va traiter l'entrée  $X_{t+1}$  et donc avoir une influence sur la sortie  $Y_{t+1}$  ;

- selon une version « dépliée » du schéma, qui montre plus clairement l'influence du traitement de chaque entrée sur le traitement de l'entrée suivante. Pour parler de la phase traitée par les couches cachées et influençant les couches cachées à l'état  $t + 1$ , on utilise le terme « état caché ».

### Schéma d'un réseau de neurones récurrent



Ce modèle peut être entraîné de la même façon qu'un réseau de neurones à action directe grâce à la technique de rétropropagation du gradient. Toutefois, le calcul du gradient prend en compte un **paramètre supplémentaire** : **l'information provenant des étapes précédentes de la séquence**. La prise en compte de ce nouveau paramètre crée deux problèmes opposés : d'une part, un risque de « **disparition du gradient** » (en anglais *gradient vanishing*), situation où le gradient devient très faible et ne modifie quasiment plus les poids et les biais du modèle ; d'autre part, un risque

d'« **explosion du gradient** », situation où, au contraire, le gradient devient très fort et modifie les poids et les biais de façon erratique. Ces problèmes rendent le gradient inopérant à long terme, ce qui empêche les RNN simples de traiter l'information de façon cohérente sur une longue séquence. Pour pallier ces problèmes, l'architecture des RNN a dû être améliorée pour posséder une mémoire à court terme, mais également une mémoire à long terme.

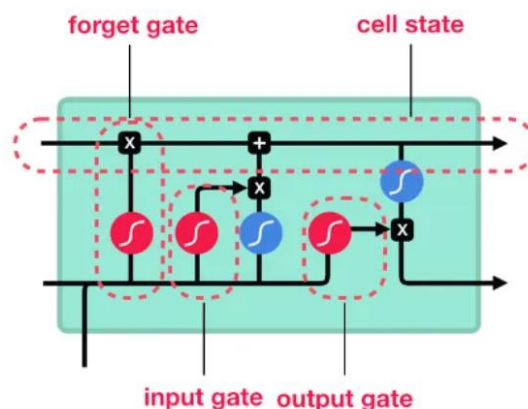
(6) Les réseaux de neurones à mémoire court et long terme (LSTM)

Les RNN donnent de bons résultats pour une mémoire courte par exemple pour assurer la prédiction du mot de la phrase suivante « *La couleur du ciel est .....* » mais pour une phrase plus longue comme « *J'ai passé vingt longues années à travailler pour les enfants défavorisés en Espagne. J'ai ensuite déménagé en Afrique. Je parle couramment .....* », les RNN sont en difficulté car **l'information se propage dans le réseau puis se perd à cause du problème de la disparition du gradient** (*vanishing gradient*)<sup>1</sup>.

Les réseaux de neurones à mémoire court et long terme (en anglais *long-short term memory* ou LSTM) sont une **architecture de RNN qui permet de résoudre les problèmes de disparition et d'explosion du gradient**.

Cette architecture, qui est la plus utilisée en pratique, a été inventée en 1997 par Sepp Hochreiter et Jürgen Schmidhuber<sup>2</sup>. L'information passe par trois portes : une porte d'entrée (*input gate*), une porte de sortie (*output gate*) et une porte d'oubli (*forget gate*).

**Schéma simplifié d'une cellule de LSTM et de ses trois portes**



Source : Omar Imai, op. cit.

<sup>1</sup> L'exemple est tiré de l'article suivant d'Omar Imai relatif aux réseaux de neurones récurrents à mémoire court terme et long terme : <https://medium.com/@pcomarimai/les-r%C3%A9seaux-de-neurones-r%C3%A9currents-%C3%A0-m%C3%A9moire-court-terme-et-long-terme- lstm-e8b4f83f4ab>

<sup>2</sup> Sepp Hochreiter et Jürgen Schmidhuber, 1997, « Long short-term memory ». Neural computation.

Chaque unité computationnelle est liée non seulement à un état caché du réseau mais également à un état de la cellule qui joue le rôle de mémoire. Ainsi, les informations passant dans les cellules LSTM sont traitées selon une **forme de mémoire** : alors que certaines informations sont gardées à long terme, d'autres sont oubliées et ne passent donc pas au jalon temporel suivant.

Les réseaux RNN complétés de ces cellules LSTM ont notamment débouché sur les architectures *Transformer* qui sont à la base des *Large Language Models* et qui seront traitées plus loin.

### c) Les autres systèmes d'apprentissage

D'autres systèmes ont précédé le *Deep Learning* et peuvent parfois présenter moins d'intérêt mais ils **restent pertinents**, par exemple dans des **cas d'usage où la puissance des réseaux de neurones profonds n'est pas nécessaire**.

#### (1) Les machines à vecteurs de support (SVM)

Les **machines à vecteurs de support** (*support vector machines* ou SVM, parfois traduits en « séparateurs à vaste marge » pour reprendre l'acronyme anglais) sont des moyens de **classifier une population** en groupes dont les individus sont similaires au regard d'un certain nombre de variables.

Ils correspondent à une **généralisation des classifieurs linéaires**. Leurs développements en informatique remontent aux années 1990, à la suite des travaux de théorie statistique conduits par Vladimir Vapnik en URSS depuis les années 1960 (menant notamment à la théorie de Vapnik-Chervonenkis, dite VC). Ils reposent ainsi sur les notions de **marge maximale**<sup>1</sup> et de **fonction noyau**<sup>2</sup>, qui leur préexistaient, mais qu'ils sont les premiers à articuler.

Ces modèles sont efficaces dans le cas de données de grandes dimensions et peuvent donner des résultats équivalents à ceux des réseaux de neurones. Ils nécessitent de posséder un large jeu de données d'entraînement (individus dont on connaît *a priori* les variables d'intérêt et la classe) ; grâce auquel il devient possible de prédire la classe d'autres individus dont on ne connaît que les variables d'intérêt. Les individus peuvent être représentés comme des points dans un espace qui compte autant de dimensions que le nombre de variables requis pour décrire un individu.

---

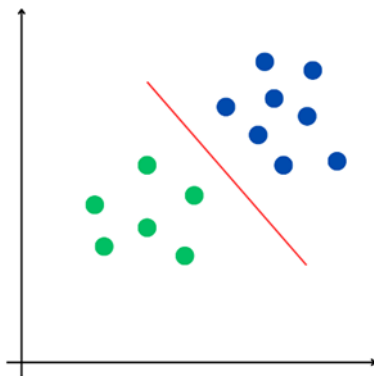
<sup>1</sup> C'est la frontière séparatrice optimale entre les différents échantillons de la population.

<sup>2</sup> Cette astuce, dite Kernel Trick en anglais, consiste à simplifier le produit scalaire (celui des  $n$  vecteurs contenus dans l'espace vectoriel complexe de grande dimension que forment les échantillons de données) en une simple fonction type, appelée fonction noyau, permettant d'éviter l'explosion combinatoire du nombre de paramètres. Les deux fonctions les plus utilisées sont le noyau gaussien et, surtout, le noyau polynomial (il utilise les similarités entre les vecteurs ainsi qu'entre leurs combinaisons). On retrouve un procédé similaire en mathématiques avec les opérateurs intégraux qui permettent de résoudre des équations intégrales.

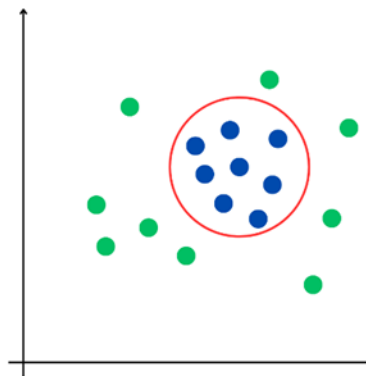
Il s'agit de **trouver la frontière séparant la population en deux classes**. Deux types de situations se présentent alors : le cas où l'on peut **séparer les classes de façon linéaire**, c'est-à-dire où l'on peut déterminer un **hyperplan** dans l'espace tel que tous les individus d'une classe sont situés d'un côté de l'hyperplan et tous les individus de l'autre classe sont situés de l'autre côté ; le cas où l'on ne peut pas le faire et où la détermination de la frontière étant plus délicate, il faut **trouver un autre moyen de séparer les classes de la population**.

**Deux exemples de classement (selon une visualisation géométrique)**

**Classement linéaire possible**

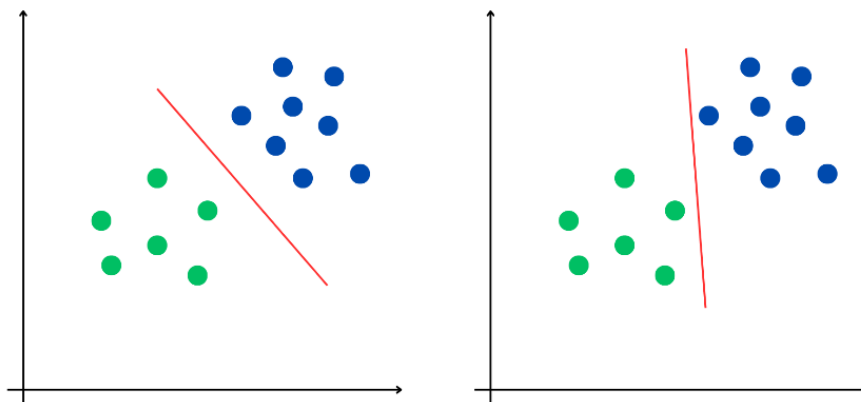


**Classement linéaire impossible  
regroupement dans un cercle**



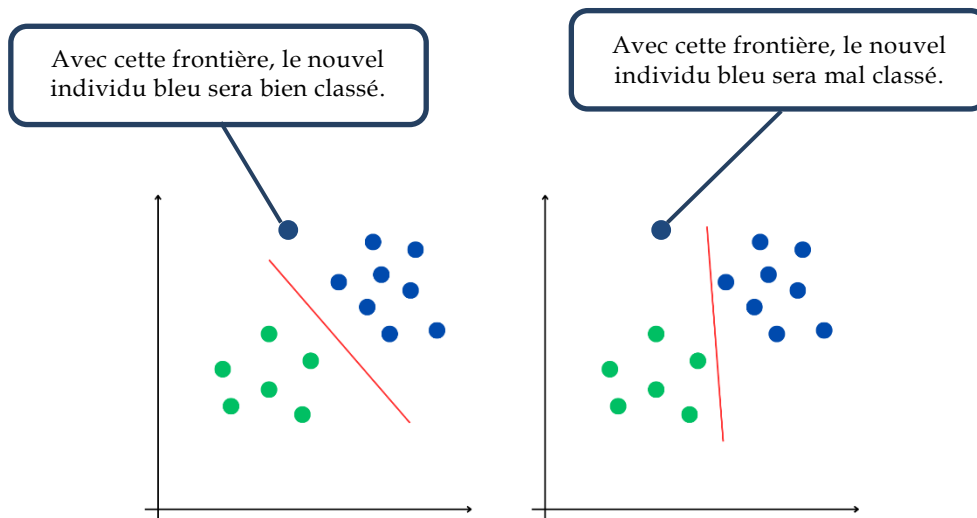
Néanmoins, il existe le plus souvent plusieurs possibilités de séparer une population en deux classes : c'est le cas des deux frontières rouges séparant la population en deux classes dans l'illustration ci-dessous.

**Deux frontières possibles pour une même classification**



Ces deux frontières semblent aussi performantes l'une que l'autre à l'issue de leurs processus d'apprentissage respectifs : le jeu de données d'apprentissage est correctement séparé en deux classes. Mais il reste à déterminer un **moyen d'obtenir la frontière la plus pertinente**, une séparation optimale qui permettrait, lorsqu'on ajoute à la population un nouvel individu n'appartenant pas au jeu d'entraînement, de déterminer sa classe avec un niveau de confiance élevé.

### La frontière la plus pertinente avec un nouvel individu

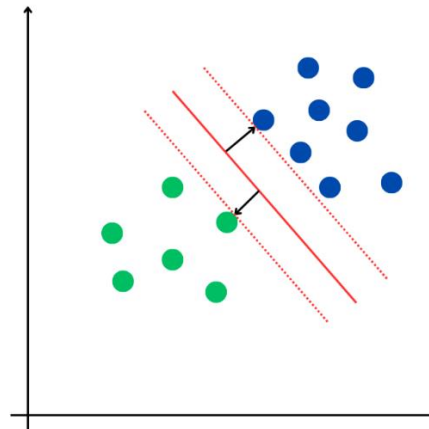


Les deux frontières sont aussi performantes l'une que l'autre à l'issue de leurs processus d'apprentissage respectifs, mais leurs performances ne sont pas équivalentes quand on généralise l'usage du modèle avec l'entrée de nouveaux individus. Il reste à optimiser cette frontière.

Pour toute frontière possible, on peut déterminer la distance qui sépare un individu de cette frontière et, par voie de conséquence, quels sont les individus les plus proches de celle-ci. Leur **distance à la frontière est appelée « marge »**. **La frontière optimale sera celle pour laquelle la marge est la plus grande**. Une fois ce problème d'optimisation résolu :

- on dispose de la frontière, c'est-à-dire de la règle de classement, qui classera le mieux les individus autres que ceux de l'ensemble d'apprentissage ;
- pour chacune des deux classes, il existe un individu au moins qui est le plus proche de la frontière. Et ces individus sont appelés « vecteurs de support », on peut les voir comme les représentants de leurs classes car si l'échantillon d'apprentissage n'était constitué que par ces vecteurs de support, la frontière optimale que l'on trouverait alors serait identique à la précédente : les vecteurs de support contiennent toute l'information qui détermine la frontière ou règle de classement.

### Le frontière optimale correspondant aux vecteurs de support

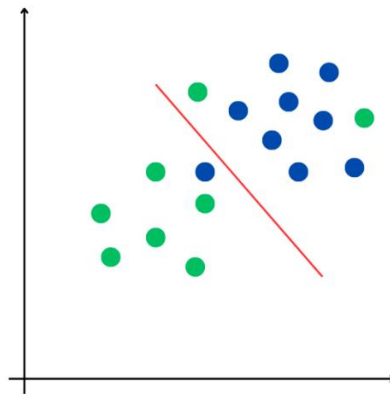


La frontière (ligne rouge continue) classe ici la population de façon optimale : tous les individus de chaque groupe sont répartis de part et d'autre et **la marge est maximale** (distance entre la frontière et les individus bleu et vert les plus proches, appelés vecteurs de support, marqués par une flèche noire). De chaque côté de la frontière, il existe un hyperplan parallèle à celle-ci qui passe par le ou les individus de support (lignes rouges pointillées) et la frontière la plus pertinente est située à égale distance de ces deux hyperplans.

Le grand intérêt de la notion de « vecteurs de support » est qu'elle permet de **généraliser l'utilisation de machines SVM à des populations qui ne sont pas linéairement séparables**. En toute rigueur, aucune frontière n'est capable de séparer complètement et exactement ce type de population entre deux classes déterminées : tout hyperplan de l'espace laissera « du mauvais côté » au moins un individu. Deux approches sont possibles dans cette situation.

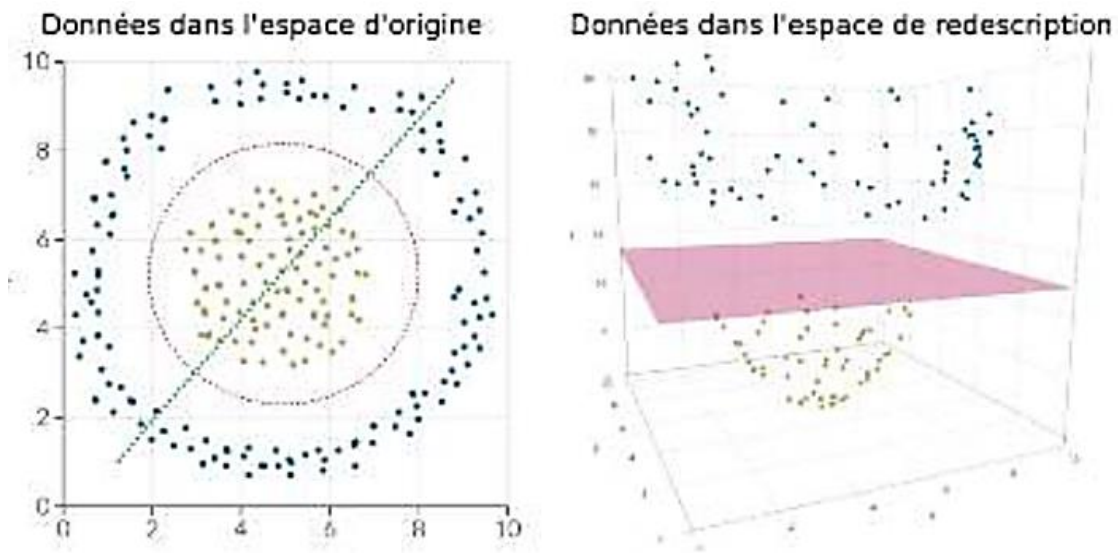
On peut d'abord accepter le fait que la règle de classement (ou la frontière) génère quelques erreurs, mais faire en sorte que leur portée reste limitée. On met alors en œuvre la méthode des **séparateurs à « marges souples »**. Le processus d'optimisation permettant de déterminer la frontière optimale fonctionne alors sous double contrainte : comme précédemment, il faut maximiser la marge, mais désormais il faut aussi minimiser une fonction d'erreur, assise sur l'écart entre le classement généré par la machine et le classement effectif, pour chaque individu du jeu d'entraînement. Un paramètre nouveau définit la tolérance de la machine aux erreurs ; en pratique, plusieurs machines sont souvent construites, avec différentes valeurs du paramètre de tolérance, puis l'on choisit la plus acceptable. Le graphique suivant illustre un séparateur à marges souples. La frontière qui sépare les deux classes de la population est optimale tout en faisant apparaître trois individus mal classés (deux verts et un bleu).

### Le seuil optimal selon un séparateur à marges souples



La seconde méthode consiste à **effectuer une séparation exacte, mais non linéaire**. Elle suppose d'ajouter des dimensions supplémentaires à l'espace des données décrivant les individus, de façon à obtenir un espace dans lequel il est certain que la population peut être linéairement séparée<sup>1</sup>. On détermine alors la frontière optimale dans ce nouvel espace, selon la méthode SVM exposée ci-dessus, puis on en déduit la frontière dans l'espace des données initial.

### Le passage par un espace de redescription



Source : Jean-Paul Comet, septembre 2024, Cours d'introduction à l'IA pour la biologie, Université de Nice-Sophia-Antipolis, cf. <https://webusers.i3s.unice.fr/~comet/SUPPORTS/index.php>

<sup>1</sup> Ce nouvel espace s'appelle « espace de redescription », ou « espace transformé », ou encore « espace des caractéristiques ». Selon le problème posé, il peut être nécessaire de prendre un espace de redescription ayant un nombre infini de dimensions.

Dans l'espace d'origine ici à deux dimensions, la population n'est pas linéairement séparable, mais dans l'espace de redescription (ici à trois dimensions), on peut trouver l'hyperplan qui correspond à la frontière optimale. Ramené à l'espace initial à deux dimensions, cet hyperplan dessine un cercle.

(2) Les modèles markoviens ou « chaînes de Markov »

Les algorithmes d'intelligence artificielle peuvent reposer sur certaines **prédictions stochastiques** modélisant des processus aléatoires, appelées modèles markoviens ou « chaînes de Markov », dont le processus de Bernoulli est l'un des exemples les plus connus, représentant la forme la plus simple de ces calculs. Il s'agit de **probabilités pour lesquelles les états futurs ne dépendent que de l'état présent et pas des états antérieurs**<sup>1</sup>. Cette absence de besoin de mémoire du passé est appelée « propriété de Markov ». Si cette propriété est présente parmi des variables aléatoires, on a affaire à un « processus de Markov ». Si ce processus correspond à des états qui sont dénombrables (autrement dit si l'on a affaire à un espace discret), il s'agit d'une « chaîne de Markov ».

Sous réserve de disposer de nombreuses variables (ces prédictions sont soumises à la loi des grands nombres), ces modèles peuvent donner des résultats significatifs, comme l'a montré leur utilisation fréquente en **physique statistique** depuis le début du XX<sup>e</sup> siècle. Les systèmes de bonus-malus des assurances les utilisent. L'algorithme **PageRank** qui a fait le succès de **Google** et qui permet à son moteur de recherche de classer les pages *Web* selon leur popularité repose sur une chaîne de Markov. De tels modèles peuvent aussi **produire du texte**, en suggérant une suite de mots.

Partons d'un processus de Markov avec des états dans lesquels on a une **variable aléatoire**  $X$  et un **processus aléatoire** par lequel l'état de  $X_{n+1}$  est déterminé par  $X_n$ , l'état futur dépendant donc de l'état présent. Dans un tel processus, il est possible de réaliser une « chaîne de Markov » **qui représente les transitions possibles entre ces états**. Le plus souvent cette chaîne est représentée par une matrice de transition mais on peut aussi la figurer dans un schéma.

Un exemple de chaîne de Markov issu de la vie quotidienne pourrait être, en le simplifiant, la **météorologie**. Avec des prévisions qui seraient réduites à seulement **trois états possibles** (« ensoleillé », « nuageux » et « pluvieux »), on part de l'état actuel – la météo du jour – qui peut être l'une de ces trois conditions météorologiques. Or l'état du lendemain dépend de l'état actuel : il est moins probable de passer d'un temps ensoleillé à un temps pluvieux que de passer d'un temps ensoleillé à un temps nuageux.

---

<sup>1</sup> La capacité à prédire le futur à partir du présent n'est pas améliorée par des informations supplémentaires relatives au passé. Dans ces calculs de probabilité, toute l'information utile à la prédiction du futur est présente dans les états actuels.



En observant les conditions climatiques aux instants  $T+1$ ,  $T+2$  ... jusqu'à  $T+N$ , on a une chaîne d'événements probables. Cette chaîne est markovienne puisque **la probabilité du temps qu'il fera demain est déterminée par le temps qu'il fait aujourd'hui**.

On a un **modèle de Markov caché** (MMC) lorsque l'on a une chaîne de Markov dont on connaît les probabilités de transition entre les différents états mais pour lequel **les états sont cachés**, ainsi que des variables observables dépendant de **variables cachées**. On connaît simplement de ces variables observées la probabilité de les observer en fonction de l'état de la variable cachée. Ainsi, l'état de la variable cachée influence la variable observable, mais l'inverse n'est pas exact.

Ces modèles tentent de comprendre **une réalité qui reste partiellement voilée par des incertitudes**. Dans un modèle de Markov caché, deux mondes coexistent : le monde caché, qui évolue selon certaines règles, et le monde visible, où nous faisons nos observations. Le défi réside dans l'absence d'accès direct aux états du monde caché : avec des indices observables laissés derrière lui, **ce monde peut être approché**.

Pour déchiffrer la séquence des états cachés d'un MMC, il faut calculer la probabilité des états cachés en fonction des observations visibles, puis utiliser le théorème de Bayes pour établir une probabilité conditionnelle et calculer les nouveaux termes de probabilité avec une approche récursive, appelée algorithme de progression ou algorithme avant (*forward algorithm*). La probabilité de chaque état est sans cesse affinée, et l'on arrive à un modèle qui associe chaque observation à une série d'états cachés, dévoilant ainsi partiellement ce qui ne pouvait pas être vu directement : inférer les états cachés a permis de finaliser l'interprétation.

Pour illustrer un modèle de Markov caché, on peut prendre l'exemple d'une ville dont on ne connaît pas la météo mais seulement les probabilités de transition entre différents états, en simplifiant à nouveau avec trois possibilités : « ensoleillé », « nuageux » et « pluvieux ». Régulièrement, on organise une visioconférence avec Eliza, une habitante de la ville dont on sait qu'elle peut être soit heureuse, soit malheureuse. On sait également que l'humeur d'Eliza dépend du temps, et on connaît la probabilité de l'humeur d'Eliza en fonction du temps qu'il fait dans sa ville. On observe pendant plusieurs jours l'humeur d'Eliza et en fonction de cela, on détermine la combinaison de temps la plus probable pour aboutir à cette suite d'humeurs. Dans cet exemple, la variable cachée est le temps qu'il fait dans la ville d'Eliza, et la variable observable est l'humeur d'Eliza. Il est possible de déterminer l'état probable de la variable cachée grâce à la variable observable.

(3) La contribution des réseaux bayésiens à l'IA

Les **réseaux bayésiens**, en référence au mathématicien britannique Thomas Bayes<sup>1</sup>, sont des **modèles graphiques probabilistes** représentant un ensemble de variables aléatoires, sous la forme d'un **graphe orienté acyclique** (de l'anglais *Directed Acyclic Graph* - DAG). Dans ce graphe, les relations de cause à effet entre les variables ne sont pas déterministes, mais probabilisées. Le réseau bayésien devient une sorte de machine à calculer des probabilités conditionnelles. En fonction des informations observées, la probabilité des données non observées peut être calculée<sup>2</sup>. L'utilisation d'un tel réseau s'appelle « inférence ». Il s'agit d'un calcul de probabilités *a posteriori*, étant donné des nouvelles informations observées.

Ces réseaux peuvent être utilisés en *Machine Learning* puisqu'à partir des données, il devient possible d'**estimer la structure d'un réseau ou les tables de probabilités d'un réseau**. Par le calcul des inférences dans des réseaux bayésiens, il est possible d'aider au diagnostic, tant en matière médicale qu'industrielle, notamment grâce à l'analyse de risques. Les réseaux bayésiens permettent aussi à des systèmes d'IA de faire de la détection des spams ou du *data mining*.

D'autres méthodes probabilistes, parfois utilisées en IA, reposent sur ces **modélisations bayésiennes**, à l'instar de la « méthode de Monte-Carlo par chaînes de Markov » (MCMC pour *Markov chain Monte Carlo* en anglais) algorithme à ne pas confondre avec « l'algorithme de Monte-Carlo » (qui utilise une source de hasard). Les MCMC utilisent la **méthode de Monte-Carlo** (qui permet de calculer une valeur numérique approchée en utilisant des procédés aléatoires, c'est-à-dire des techniques probabilistes, dans le but par exemple d'introduire des risques) mais en se basant sur le parcours de chaînes de Markov, qui ont pour lois stationnaires les distributions à échantillonner.

(4) L'apport de la « régularisation statistique » de Vapnik

Les réseaux de neurones formels, les modèles de Markov cachés mais aussi tous les autres modèles statistiques classiques utilisent soit des méthodes d'optimisation directe, comme la régression linéaire, soit des méthodes itératives comme la descente du gradient. Or tous les systèmes d'apprentissage font face à des problèmes de **surapprentissage**, on parle aussi parfois de surajustement ou de surinterprétation. Le modèle devient trop précis car il contient plus de paramètres que les données ne le justifient.

---

<sup>1</sup> Le théorème de Bayes, dérivé en 1763 de la théorie des probabilités conditionnelles et utilisé pour le calcul de probabilités ainsi que pour les statistiques, permet de déterminer la probabilité qu'un événement arrive à partir d'un autre événement qui s'est réalisé, notamment quand ces deux événements sont interdépendants. Nécessitant un grand nombre de calculs, les inférences bayésiennes se sont surtout développées grâce à l'informatique. Le mathématicien français Laplace a formulé la même théorie un peu plus tard sans connaître les travaux de Bayes.

<sup>2</sup> Par exemple, en fonction des symptômes d'un malade, on calcule les probabilités des différentes pathologies compatibles avec ces symptômes. On peut aussi calculer la probabilité de symptômes non observés, et en déduire les examens complémentaires les plus nécessaires.

Vladimir Vapnik, l'inventeur des SVM, a répondu au problème dans l'URSS des années 1970 et 1980 avec sa **théorie de la régularisation statistique**.

Pour réduire la variance des modèles, on introduit de nouvelles informations permettant par exemple de **pénaliser les valeurs extrêmes** des paramètres. Le plus souvent, il s'agit d'**utiliser une norme sur ces paramètres**, que l'on va **ajouter à la fonction qu'on cherche à minimiser**. L'optimisation devient alors possible en évitant ou du moins en réduisant les phénomènes de surapprentissage.

## II. LES IA GÉNÉRATIVES : PRINCIPALE INNOVATION TECHNOLOGIQUE EN INTELLIGENCE ARTIFICIELLE DEPUIS 2017

### A. LES PRINCIPALES AVANCÉES EN MATIÈRE DE GÉNÉRATION DE CONTENU CES DERNIÈRES ANNÉES

En amont de la mise à disposition de ChatGPT par OpenAI le 30 novembre 2022, système qui a marqué le grand public par ses performances malgré des hallucinations alors encore trop nombreuses, un ensemble d'innovations ont conduit à des progrès dans les technologies d'IA. La principale de ces innovations est la définition d'une **nouvelle architecture pour les modèles d'apprentissage profond appelée Transformer**.

Son invention en 2017 a permis cinq ans plus tard la création du modèle *Generative Pre-trained Transformer*, dont les initiales GPT ont été données au système d'agent conversationnel d'OpenAI. Selon Laurent Alexandre, entendu par vos rapporteurs, le XXI<sup>e</sup> siècle est né le 30 novembre 2022 avec l'introduction de cet outil d'IA générative.

Il serait plus juste de **dater cette révolution technologique de l'année où le modèle est proposé, soit 2017**, d'autant plus que ChatGPT, s'il est le premier à être commercialisé, n'est pas la première IA générative à être mise au point : le modèle BERT (pour *Bidirectional Encoder Representations from Transformers*) a été conçu par Google dès 2018. Et l'année suivante, des modèles spécifiques à la langue française sont même développés : CamemBERT et FlauBERT<sup>1</sup>.

Ces outils ont d'abord concerné le langage avec les grands modèles de langage (LLM de l'anglais *Large Language Models*) puis les **contenus sonores ou visuels**. Depuis 2017, chaque trimestre en moyenne, les modèles sont l'objet d'avancées significatives, qualitativement, du point de vue de leur architecture ou, au moins, quantitativement, du point de vue de l'accroissement de leur taille, ce qui les rend mécaniquement de plus en plus efficaces. On parle à ce sujet de lois d'échelle ou *scaling laws*.

---

<sup>1</sup> Cf. Louis Martin et al., 2019, « CamemBERT : a Tasty French Language Model » : <https://arxiv.org/abs/1911.03894> ainsi que Le Hang et al., 2019, « FlauBERT: Unsupervised Language Model Pre-training for French » : <https://arxiv.org/abs/1912.05372>

## 1. « *Attention is all you need* » : la technologie *Transformer* inventée en 2017

L'innovation introduite en 2017 consiste à **apporter une amélioration à l'apprentissage profond** résultant de la multiplication des couches de réseaux de neurones, telle que vue précédemment, ouvrant la voie d'intelligences artificielles génératives performantes, en particulier dans le domaine des LLM dédiés au traitement naturel du langage<sup>1</sup>.

Dans un article scientifique qui fait référence à la chanson des Beatles *All you need is love* – dont la conclusion du refrain *Love is all you need* est paraphrasée sous la forme *Attention is all you need*<sup>2</sup> – des chercheurs de l'entreprise Google ont théorisé l'architecture *Transformer* (plus rarement appelée « Transformeur »). Alors que les LSTM ont une mémoire relativement courte et ont du mal à traiter de propositions ambiguës, l'architecture *Transformer* résout en grande partie ce problème.

En effet, cette dernière est dotée d'un **mécanisme appelé « l'attention »**, qui lui permet d'obtenir des informations sur les mots en fonction du contexte de la phrase et ainsi de traiter plus d'éléments qu'un LSTM, y compris pour une proposition dont le sens ne se déduit pas de sa seule formulation.

En pratique, le mécanisme d'attention est un encodeur qui fonctionne par étapes successives permettant de transformer toute l'information nécessaire de la séquence de mots à traiter en données numériques (c'est le *word embedding* qui a notamment recours à des fonctions sinus et cosinus). Ainsi, après une parallélisation de calculs de matrices de poids d'attention (on parle de *multi-headed attention* ou attention multi-têtes<sup>3</sup>), on va **transformer les mots en vecteurs**, puis calculer les liens grammaticaux, sémantiques et pragmatiques entre les différents mots d'une séquence.

---

<sup>1</sup> Il existe cependant des modèles de langage reposant sur des IA symboliques, notamment les modèles d'alignement IBM qui ont dominé les années 1990 et 2000. Pour la commodité de l'exposé, les LLM seront rattachés au Deep Learning.

<sup>2</sup> Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, et Illia Polosukhin, 2017, « *Attention is all you need* », *Advances in neural information processing systems*, n° 30.

<sup>3</sup> Cette étape du mécanisme d'attention des Transformers qui suit le *word embedding* consiste à paralléliser les calculs de manière différente des réseaux récurrents puisqu'il s'agit de manière simplificatrice de procéder à une coupe horizontale des têtes de matrices (les auteurs de l'article distinguent trois matrices : la requête, la clé et la valeur) en « *h* matrices », d'où l'idée d'attention multi-têtes, avant de concaténer les données de ces matrices tranchées en une seule matrice de poids. La normalisation permet alors de passer à la couche linéaire finale, sorte de classifieur géant. Ces étapes de l'encodeur précèdent celles du décodeur, qui repose lui aussi sur une *multi-headed attention*.

Par exemple, la phrase « Alice amène sa voiture rouge au garage car elle est en panne » est facilement compréhensible par un être humain, elle l'est toutefois beaucoup moins pour un modèle de langage sans attention. En effet, si le modèle traite les informations mot à mot, alors il est incapable de lier les adjectifs aux noms, ainsi « voiture rouge » ne sera pas un concept pour le modèle mais l'addition des concepts de « voiture » et de « rouge », ce qui peut poser des problèmes puisque tout ce qui est rouge n'est pas une voiture et toutes les voitures ne sont pas rouges.

En plus de cela, il existe des ambiguïtés qui sont évidentes à lever pour un être humain mais ne le sont pas pour le modèle. Dans la phrase d'exemple, le pronom « elle » est ambigu, il peut renvoyer au sujet ou à l'objet de la phrase, c'est-à-dire à la voiture ou à Alice.

Avec le contexte, il est clair pour un humain que l'on parle ici de la voiture d'Alice, mais il s'agit d'une information dont un modèle ne peut disposer que grâce à un système d'attention.

Le fonctionnement de ce système nécessite d'abord de transformer une séquence de mots, dont l'information est transmise par des données de type « chaîne de caractères », en valeurs numériques qui pourront alors être traitées par un réseau de neurones.

a) La « tokenisation » : découper préalablement les mots

Une fois que le jeu de données d'apprentissage a été constitué (ici, un ensemble étendu de textes), la première étape de cette technologie consiste à **découper les mots en unités ou *tokens* qui sont des briques élémentaires de vocabulaire servant d'unités de base pour le modèle.**

Cette technique permet une meilleure capacité du modèle en termes de généralisation, en particulier par rapport à des mots rares ou à des variantes linguistiques.

Les données d'entraînement contiennent des milliards de ces *tokens*. Ces briques, souvent plus petites qu'un mot, sont les éléments traités par le reste du mécanisme d'attention. Pour cela, il existe plusieurs algorithmes : BPE<sup>1</sup> (*Byte Pair Encoding*), WordPiece<sup>2</sup>, Unigram<sup>3</sup>, etc.

---

<sup>1</sup> *Byte Pair Encoding (BPE)* est un algorithme très utilisé pour réduire les mots rares en séquences de sous-unités plus fréquentes. Par exemple, le mot « unhappiness » peut être découpé en « un » et « happiness » pour augmenter la fréquence d'apparition des sous-éléments et ainsi mieux généraliser.

<sup>2</sup> Utilisé par des modèles comme BERT de Google, l'algorithme WordPiece permet de découper les mots en sous-unités de manière à optimiser la couverture des sous-structures linguistiques. Par exemple, « playground » peut être décomposé en « play » et « ###ground », ce qui permet au modèle d'identifier l'importance de l'ancrage sémantique dans certains tokens.

<sup>3</sup> Cette méthode se base sur un échantillonnage probabiliste pour sélectionner les segments de mot en fonction de leur fréquence. Par exemple, le mot « undoubtedly » pourra être découpé en plusieurs sous-unités telles que « un », « doubt », et « ly » en tenant compte des probabilités des tokens les plus fréquents.

Par exemple la phrase « *Science sans conscience n'est que ruine de l'âme* » tirée du roman *Pantagruel* de Rabelais est transformée par GPT-4<sup>1</sup> en quatorze *tokens* :

*Science* | *sans* | *conscience* | *n* | *'est* | *que* | *ru* | *ine* | *de* | *l* | *'* | *â* | *me* | .

On constate que certains mots ne sont pas découpés en unités plus petites comme « science » et « conscience », cependant les mots « ruine » et « âme », eux, sont découpés.

Si l'on reprend l'exemple pris pour introduire le concept d'attention, on obtient ce découpage en quinze unités :

*Alice* | *am* | *ène* | *sa* | *voiture* | *rouge* | *au* | *garage* | *car* | *elle* | *est* | *en* | *pan* | *ne* | .

Du côté de la génération de contenus, comme peuvent souvent le voir les utilisateurs de ces systèmes en temps réel, le texte est généré, *token* après *token*, ce qui a des conséquences sur la probabilité que chaque *token* généré à l'étape suivante soit le bon.

#### b) Le plongement lexical : vectoriser les tokens

Même découpés ainsi en briques élémentaires, les mots sont toujours des chaînes de caractères. Or, on l'a vu précédemment, les réseaux de neurones traitent des données numériques. Donc, pour être traitée dans un modèle d'IA, l'information lexicale doit être transformée en information numérique.

Pour cela, on réalise **une opération essentielle que l'on appelle un « plongement lexical »** ou *word embedding* en anglais. Une méthode qui préexistait aux *Transformers*.

Il s'agit d'associer à chaque unité un vecteur représentant les coordonnées du mot dans un espace possédant un grand nombre de dimensions. Ce vecteur est un ensemble de coordonnées, notées dans une colonne unique, chaque nombre dans la colonne définit la coordonnée du mot dans une dimension de l'espace. Par exemple, prenons le cas d'un vecteur  $\vec{E}_1$  qui représente un mot dans l'espace de plongement, on note ses coordonnées de la manière suivante :

$$\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}.$$

---

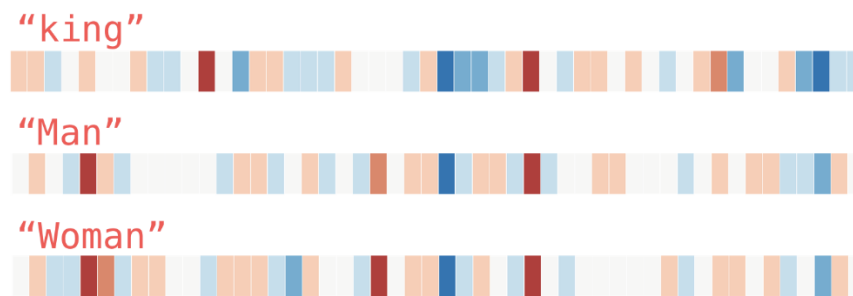
<sup>1</sup> OpenAI propose sur son site un outil qui permet de visualiser la tokenisation d'une séquence de mots : <https://platform.openai.com/tokenizer>

Le nombre  $x_1$  correspond à la coordonnée du mot dans la première dimension de l'espace,  $x_2$  dans la deuxième, etc., jusqu'à la nième dimension,  $n$  étant le nombre total de dimensions de l'espace de plongement. Ainsi, le nombre de coordonnées dans un vecteur définit le nombre de dimensions de l'espace de plongement. L'espace de plongement possède **un très grand nombre de dimensions**, il ne ressemble pas aux espaces en deux ou trois dimensions dont nous sommes plus familiers. L'espace de plongement est difficile à imaginer de façon intuitive : un vecteur qui représente des coordonnées est préféré car les vecteurs permettent de réaliser des opérations qui rendent cet espace multidimensionnel moins abstrait.

Si l'on remplace chaque coordonnée des vecteurs par des couleurs, par exemple un dégradé allant du bleu au rouge avec la valeur zéro pour le blanc, on peut représenter visuellement les vecteurs des mots « plongés » (mots qui sont les objets du plongement lexical). Cela permet d'observer des propriétés intéressantes en ne regardant que les coordonnées des vecteurs plongés.

Dans l'image qui suit, on constate que les mots « homme » et « femme » ont plus de similitudes entre eux, que chacun d'eux peut en avoir avec le mot « roi », ce qui est normal puisque les deux désignent un genre, alors que « roi » désigne une fonction.

#### Représentation sous forme de couleur des coordonnées vectorielles des mots « king », « man » et « woman »



Source : Jay Alammam, *Word Embedding*<sup>1</sup>

Plonger les unités lexicales dans un espace continu permet également de **réaliser des calculs** avec eux plutôt que de se contenter d'observer des similarités sans les mesurer. Les vecteurs possèdent **trois caractéristiques** avec lesquelles il est possible de réaliser des opérations. Ils ont une **norme**, qui correspond à leur longueur dans l'espace, une **direction** qui est l'orientation du vecteur dans l'espace, et un **sens**, qui indique vers quelle extrémité le vecteur pointe (de l'origine du repère vers l'extrémité qui est le mot dans l'espace).

<sup>1</sup> Cf. Jay Alammam, « The Illustrated Word2vec » : <https://jalammam.github.io/illustrated-word2vec/> et une traduction française de cet article de Jay Alammam disponible au lien suivant : [https://lbourdois.github.io/blog/nlp/word\\_embedding/](https://lbourdois.github.io/blog/nlp/word_embedding/)

Réaliser des opérations sur les vecteurs correspond par exemple à **additionner deux vecteurs**, ce qui revient à faire translater un vecteur à l'extrémité d'un autre pour tracer le vecteur sommes : celui-ci part alors de l'origine du premier, passe par l'extrémité à l'origine du second et se termine à l'extrémité du second. Une autre opération courante avec des vecteurs est le **produit scalaire** de deux vecteurs, qui est un nombre. Ce produit scalaire se calcule en multipliant leurs normes respectives par le cosinus de l'angle formé par ces deux vecteurs<sup>1</sup> :

$$\vec{E}_1 \cdot \vec{E}_2 = \|\vec{E}_1\| \times \|\vec{E}_2\| \times \cos(\vec{E}_1; \vec{E}_2)$$

Aussi, si le plongement lexical est correctement réalisé, il est possible d'effectuer des opérations sur la base du « sens » des mots, entendu comme l'ensemble des relations de proximité ou d'éloignement entre mots que l'apprentissage a permis d'identifier. Par exemple, dans cet espace, le vecteur de différences entre les mots « homme » et « roi » est similaire au vecteur de différences entre les mots « femme » et « reine ». Ainsi, on peut établir l'identité approximative telle que :

$$\vec{\text{king}} - \vec{\text{man}} + \vec{\text{woman}} \approx \vec{\text{queen}}$$

Des chercheurs ont cherché à comprendre en 2019<sup>2</sup> la **structure mathématique de ces analogies** entre les mots, qui prennent par exemple une forme de type « a est à b ce que x est à y », permettant à des modèles d'affirmer que a et x peuvent être transformés de la même manière pour obtenir b et y, et vice-versa.

D'un point de vue formel, ces **analogies linéaires entre les mots** correspondent à des relations vectorielles où les vecteurs forment dans l'espace vectoriel une **structure géométrique de type parallélogramme**. Ces quadrilatères ont des propriétés utiles qui peuvent être exploitées. Une analogie linéaire dans l'espace vectoriel entre tel ensemble de mots (ou de *tokens*) et tel autre ensemble de mots (ou de *tokens*) montre que leurs vecteurs sont coplanaires et que toute combinaison de ces vecteurs peut être **utilisée pour réécrire les structures de relations entre ces éléments en termes statistiques**.

---

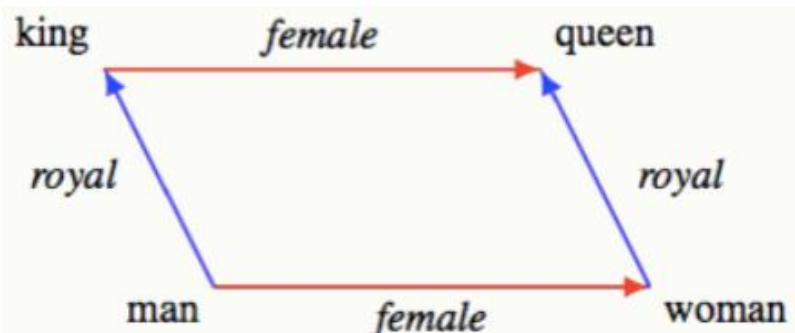
<sup>1</sup> Un produit scalaire égal à zéro signifie, soit que la norme d'un des deux vecteurs est nulle, soit que les vecteurs sont orthogonaux. Si le cosinus du produit scalaire se situe toujours entre -1 et 1, le produit scalaire lui peut être bien plus élevé car il est le produit de la multiplication des autres éléments. Par exemple avec des normes de 10 et 10 et un angle de zéro, le cosinus vaut 1 et le produit scalaire est 100. Le plongement lexical est toutefois souvent normalisé autour d'une norme de 1 et, dans ce cas seulement, le produit scalaire variera alors lui aussi toujours entre -1 et 1.

<sup>2</sup> Kawin Ethayarajh et al., 2019, « Towards Understanding Linear Word Analogies », Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: <https://aclanthology.org/P19-1315/> un résumé de l'article a été fait sur son blog par le premier auteur de cet article collectif : <https://kawine.github.io/blog/nlp/2019/06/21/word-analogies.html>



En pratique, les systèmes de relations sont bien entendu plus complexes que l'exemple de parallélogramme dans l'espace vectoriel donné ici.

#### L'exemple d'un parallélogramme analogique au sein de l'espace vectoriel



Source : Kawin Ethayarajh, op. cit.

Cette structure implique que **les vecteurs sont coplanaires** et que les **relations entre les mots** peuvent être **représentées géométriquement** mais surtout **statistiquement**. La **corrélation entre les mots** dans un corpus d'apprentissage est ainsi d'autant **plus élevée** que la **distance euclidienne entre leurs vecteurs est faible**. Cette affirmation qui semble intuitive repose, selon Kawin Ethayarajh, sur la première interprétation théorique de la distance euclidienne dans les espaces vectoriels de mots. Il montre que cette approche surpasse scientifiquement les théories arithmétiques sur les analogies entre les mots en affirmant que la distance euclidienne est une fonction linéaire décroissante de l'induction mathématique de leur co-occurrence vectorielle relative, à l'aide du théorème csPMI<sup>1</sup>.

Il est donc possible de **réaliser des produits scalaires entre deux vecteurs** au sein d'un espace vectoriel et de comparer les produits scalaires de ces vecteurs pour voir lesquels « vont dans la même direction », témoignant de leurs proximités sémantiques.

Pour mieux comprendre ce principe, il est possible de l'illustrer d'un exemple avec une réduction des dimensions de l'espace de plongement. Ainsi, l'algorithme word2vec, dont une démonstration est disponible sur le site de l'École Polytechnique<sup>2</sup>, permet de réaliser des opérations à partir de mots qui ont été plongés : il est possible d'effectuer une « similarité cosinus », qui est une opération utilisant le produit scalaire en faisant en sorte que les normes des vecteurs comparés ne comptent pas dans le résultat final (on parle d'une « normalisation » ou *encoding* en anglais).

<sup>1</sup> Le théorème csPMI signifie Co-occurrence Shifted Pointwise Mutual Information Theorem. Il a été introduit dans l'article précité de Kawin Ethayarajh et al. de 2019. Il s'agit d'une mesure statistique utilisée pour analyser les analogies de mots dans les espaces vectoriels. Le théorème établit que, dans un espace vectoriel sans bruit, une analogie linéaire entre des paires de mots tient exactement si la csPMI est identique pour chaque paire. Ce théorème offre une interprétation théorique de la distance euclidienne dans les espaces vectoriels de mots, montrant que cette distance est une fonction linéaire décroissante de la csPMI entre les mots.

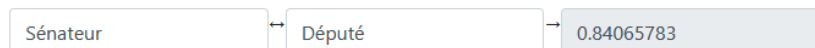
<sup>2</sup> Cf. cette page du site de l'École Polytechnique : <http://nlp.polytechnique.fr/word2vec>

Dans l'espace vectoriel, les mots « Sénateur » et « Député » ont une similarité cosinus d'environ 0,84 ce qui est un score relativement important, illustrant le fait que dans l'espace vectoriel, ces deux mots sont proches. La proximité sémantique peut être reproduite *via* des calculs en tant que similarité cosinus.

### Proximité des mots « Sénateur » et « Député » dans un espace vectoriel

#### Word Difference In Vector Space

The result shows how much the two input words are similar by computing the cosine similarity.

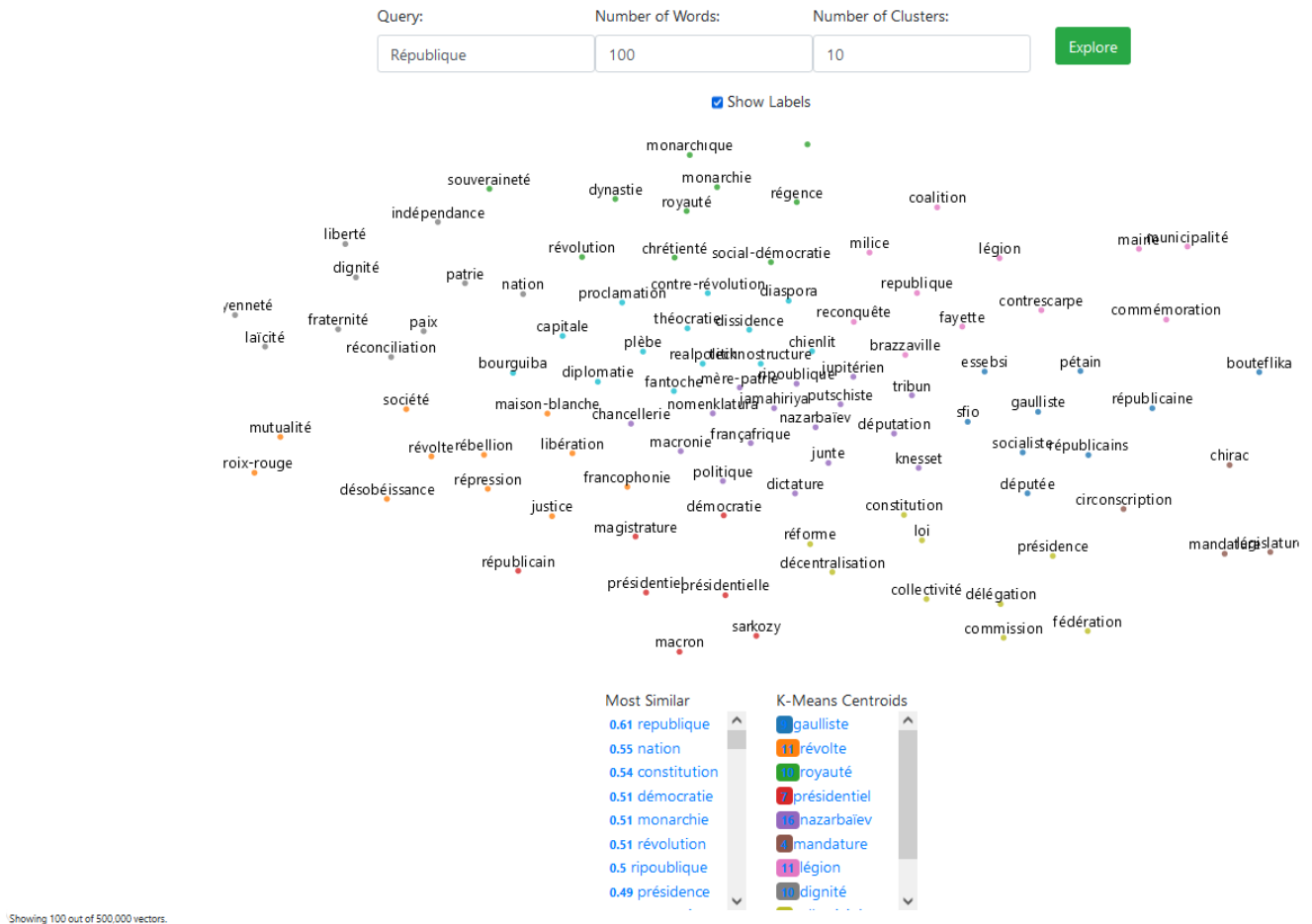


*Source : Calcul réalisé à l'aide du site de l'École Polytechnique*

On peut également prendre une représentation en deux dimensions de l'espace vectoriel (qui est lui à n dimensions) et trouver les voisins de mots déterminés. On peut alors se faire une idée intuitive de la façon dont fonctionne l'espace de plongement. Il est toutefois toujours important de garder en tête qu'il ne s'agit que d'une représentation très simplifiée : un espace de plongement possède un grand nombre de dimensions et donc des propriétés particulières qui peuvent être contre-intuitives par rapport à sa forme visuelle.

Un exemple peut être donné à travers une représentation graphique, grâce à la démonstration de l'algorithme word2vec sur le site de l'École Polytechnique, des **cent mots les plus proches de « République »**. On peut y voir des mots liés aux valeurs républicaines comme à la vie politique plus généralement.

## Représentation des cent mots les plus proches de « République »



Source : Schéma réalisé à l'aide du site de l'École Polytechnique

### c) Le mécanisme d'attention : intégrer le contexte

Encoder les *tokens* (ce qui revient à normaliser les jetons) dans un espace de très grande dimension ne suffit pas. En effet, un jeton ou un mot peuvent être **polysémiques** : dans l'exemple précédent, le mot « voiture » en français peut désigner le véhicule motorisé individuel, mais également une partie d'un train qui accueille les voyageurs. Des informations grammaticales doivent également être tirées du sens de la phrase.

Si l'on reprend la phrase :

« Alice | am | ène | sa | voiture | rouge | au | garage | car | elle | est | en | pan | ne | »,

il est naturel pour nous de savoir que le mot « elle » dans cette phrase concerne la voiture, et non Alice. Toutefois, cette information est impossible à connaître pour un système d'IA sur la base du simple plongement lexical des mots. Le défi consiste à introduire cette information contextuelle.

Les vecteurs de plongement lexical vont pour cela **passer par un processus appelé « mécanisme d'attention »** afin d'y associer les informations de contexte essentielles à la compréhension de la phrase : à quel nom se rattache un adjectif, à quel nom renvoie un pronom, quel est le « sens » d'un mot dans la phrase, etc.

Tout au long de cette première phase d'entraînement, le modèle établit les relations probabilistes entre chaque *token* et tous les autres *tokens* de l'ensemble de données.

Dans les modèles d'IA générative basés sur l'architecture *Transformer*, le mécanisme d'attention est un **algorithme par lequel le modèle identifie les briques élémentaires qui offrent des informations contextuelles sur la signification des autres briques**. Cet algorithme juge de la pertinence des entrées en fonction du contexte spécifique de telle ou telle demande (les requêtes sont appelées *prompts*) et leur attribue différents poids relatifs dans son processus de calcul.

Le modèle est alors en mesure de **produire des résultats en prédisant la suite de *tokens* probablement la plus adaptée à un contexte donné**, selon une fonction *softmax* (aussi appelée, plus rarement, fonction exponentielle normalisée), qui permet de donner aux vecteurs une valeur entre 0 et 1 (ou entre 0 et 100 en pourcentage).

Les LLM utilisés depuis l'introduction de ChatGPT en 2022 reposent sur ce **mécanisme mathématique de mise en contexte par algorithme**.

*d) Un apprentissage auto-supervisé et des données synthétiques*

Parmi les méthodes d'apprentissage automatique (*Machine Learning*), dont celles recourant à des réseaux de neurones en couches profondes (*Deep Learning*), le modèle apprend le plus souvent à partir d'échantillons de données annotées. Comme l'explique Yann LeCun depuis les progrès enregistrés par ces algorithmes dans les années 2010, **l'un des grands défis sera de remplacer cet apprentissage supervisé par un apprentissage non supervisé**.

Ainsi que l'a souligné Bertrand Braunschweig lors de son audition, l'architecture *Transformer*, peut être vue comme une **étape dans le passage de l'apprentissage supervisé à l'apprentissage non supervisé**. On parle à ce sujet d'**apprentissage auto-supervisé** (et parfois d'apprentissage semi-supervisé). En anglais, la notion de *self-supervised learning* (ou SSL) est utilisée.

Le modèle apprend à partir d'échantillons de données non annotées certes mais, **dans un premier temps, l'apprentissage reste supervisé** : les réseaux de neurones sont initialisés avec des étiquettes qui aident à déterminer les premiers poids synaptiques de chacun des réseaux.

Ce n'est que dans un deuxième temps que **l'apprentissage devient non supervisé**, les échantillons de données ne sont pas annotés et le modèle générant génère automatiquement les étiquettes associées, sans intervention humaine.

L'apprentissage auto-supervisé est la technique d'apprentissage des modèles *Transformer* mais d'autres modèles y ont également recours. Différentes applications de reconnaissance sont élaborées par Facebook par exemple : une fois entraîné selon un apprentissage supervisé à identifier des liens entre des contenus, le modèle passe à un apprentissage auto-supervisé qui lui permet de générer automatiquement des étiquettes, conduisant au fait qu'un modèle identifie seul les relations entre des images similaires.

En plus des données préexistantes, issues du monde réel et du monde virtuel qu'est Internet, les modèles d'IA générative ont de plus en plus recours à **des données créées artificiellement** dont les propriétés statistiques prédictives sont proches des données réelles mais dont les conditions de mise à disposition sont moins chères, plus accessibles ou ne contiennent pas d'informations personnelles ou sensibles. Ces corpus, dérivés du réel ou pas, prennent le nom de **données synthétiques** ou *synthetic data*. De telles données permettent de réduire les coûts élevés des données issues du monde réel et de contourner les contraintes éthiques de données réelles tout en conservant des propriétés prédictives similaires.

Avant l'IA générative, de telles données, comme des animations de jeux vidéo en 3D, ont pu être utilisées pour entraîner des systèmes de **reconnaissance visuelle**. En **biologie** et en **médecine**, ce type de données a d'ores et déjà été très utile pour pallier le manque de données réelles, par exemple pour concevoir de nouveaux médicaments contre les maladies rares. La production rapide des **vaccins** lors de la pandémie de covid-19 reposait aussi sur l'utilisation de données synthétiques<sup>1</sup>.

En tant que données de nature secondaire, dérivées de corpus qui peuvent eux-mêmes contenir des biais, les données synthétiques posent à la fois la question des **biais liés aux données** en général mais elles peuvent aussi apporter **leurs propres biais** en plus<sup>2</sup>, et induire des erreurs dans les modèles d'IA, si elles ne reflètent pas correctement la réalité ou si elles sont mal calibrées<sup>3</sup>.

---

<sup>1</sup> Cf. Hari Prasanna Das et al., 2022, « Conditional Synthetic Data Generation for Robust Machine Learning Applications with Limited Pandemic Data », Proceedings of the 36<sup>th</sup> AAAI Conference on Artificial Intelligence : <https://ojs.aaai.org/index.php/AAAI/article/view/21435>

<sup>2</sup> Cf. l'étude de la revue d'IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, « The Limitations of Synthetic Data », IEEE.

<sup>3</sup> Un article dénombre la multiplicité des biais dans les données synthétiques : biais de représentation, biais de distribution, biais de modèle ou encore biais d'utilisation, cf. Khaled El Emam, Thomas J. Brunner, Daniel E. Rubinstein et Michael I. Jordan, 2022, « Bias in Synthetic Data : A Survey », ACM Computing Surveys.

Elles peuvent aussi **renforcer des biais** existants et **déformer les résultats** produits ensuite par les modèles d'IA<sup>1</sup>.

Enfin, elles peuvent enfermer les modèles dans une **boucle autodestructrice**, appelée en anglais *autophagous* ou *self-consuming loop*, dégradant progressivement la qualité et/ou la diversité des données synthétiques et menant à un échec du modèle. Ce dernier point a fait l'objet d'un article critique en août 2024<sup>2</sup>, qui propose un nouveau cadre d'entraînement itératif pour les modèles de diffusion sur la base de données synthétiques auto-améliorantes, appelé *Self-Improving diffusion models with Synthetic data* (ou SIMS).

Lutter contre les biais de ces données synthétiques, en plus de lutter contre les biais des autres données plus généralement, doit être un objectif pour les développeurs de systèmes d'IA. Il faut **mieux les évaluer** en amont afin de pouvoir les prendre en considération ensuite lors des phases d'utilisation de données synthétiques pour l'entraînement des modèles d'IA, des **mécanismes d'audit et de correction** sont nécessaires et ils doivent prévoir le cas particulier des biais spécifiques introduits au stade des données synthétiques<sup>3</sup>. Ces audits, utilisant par exemple des tests spécifiques de biais avant entraînement des modèles, responsabiliseraient les concepteurs d'IA dans une logique proactive dans une démarche d'atténuation systématique et tangible et de réduction des biais. Les modèles de diffusion de type SIMS permettront une auto-amélioration et un ajustement de la distribution des données synthétiques des modèles de diffusion pour correspondre à une distribution cible souhaitée afin d'aider à atténuer les biais et à garantir l'équité des résultats.

---

<sup>1</sup> Une étude parue en 2023 confirme que les modèles d'IA entraînés sur des données synthétiques biaisées peuvent être biaisés eux-mêmes, le cas de l'utilisation de données synthétiques contenant des biais à raison du genre a par exemple conduit à un modèle d'IA lui-même défavorable aux femmes. Cf. un résumé de l'étude « The Dangers of Synthetic Data » publiée dans la revue Machine Intelligence de Nature en 2023, dans cet article de rapidinnovation du 29 juillet 2024 : <https://www.rapidinnovation.io/news-blogs/ai-model-collapse-the-dangers-of-synthetic-data>

<sup>2</sup> Cf. l'article de Sina Alemohammad, Ahmed Imtiaz Humayun, Shruti Agarwal, John Collomosse et Richard Baraniuk, août 2024, « Self-Improving Diffusion Models with Synthetic Data » : <https://arxiv.org/abs/2408.16333>

<sup>3</sup> Cf. Ajay Patrikar et al., 2023, « Leveraging synthetic data for AI bias mitigation », in « Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications », SPIE, volume 12529 : <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12529/125290K/Leveraging-synthetic-data-for-AI-bias-mitigation/10.1117/12.2662276.short>

## 2. Distinguer les « modèles de fondation » des systèmes d'intelligence artificielle

### a) La notion de « modèle de fondation » et ses ambiguïtés

La notion de modèle de fondation, synonyme de celle de modèle d'intelligence artificielle à usage général utilisée par l'Union européenne<sup>1</sup>, est issue des avancées en matière de *Deep Learning* dans les années 2010 et a été introduite en 2021 par des chercheurs de l'Université de Stanford<sup>2</sup> pour combler un **vide de vocabulaire** et identifier le **modèle général issu de l'entraînement des algorithmes d'apprentissage profond** pouvant ensuite faire l'objet de réglages ultérieurs plus fins, pour l'aligner avec certaines valeurs ou pour le spécialiser en vue de ses applications concrètes<sup>3</sup>.

Cette expression désigne donc les **modèles non spécialisés** entraînés sur un large ensemble de données, aux thématiques variées et non étiquetées et ayant vocation à être utilisés pour réaliser différentes tâches après avoir fait l'objet d'un *fine-tuning*. Le « modèle de fondation » **ne signifie pas que le modèle est « fondateur » au sens où il permet à l'IA d'être générative** : les modèles de fondation ne sont pas exclusivement liés à l'IA générative, il est important de noter qu'il y a des systèmes d'IA générative qui n'ont pas de modèle de fondation (ils seront vus plus loin) et des systèmes d'IA non générative basés sur de l'apprentissage profond qui peuvent s'appuyer sur la dichotomie « modèle de fondation/réglage fin ».

C'est à partir des modèles de fondation que peuvent être construites, à la suite d'un réglage fin ou « *fine-tuning* », des **applications destinées à un usage spécifique**, comme des systèmes d'intelligence artificielle générative. Ainsi, le modèle de fondation constitue la couche logicielle élémentaire permettant le bon fonctionnement d'une application. La qualité finale d'une application d'IA dépend souvent en grande partie du modèle de fondation utilisé pour le bâtir. Les modèles Llama de Meta sont des exemples de modèles de fondation. Dans leur cas, ils sont proposés en *open source* aux développeurs et aux entreprises qui sont alors libres de réaliser des applications ou systèmes d'intelligence artificielle, dans le domaine qu'ils souhaitent grâce à des techniques permettant de spécialiser ces modèles de fondation et de les entraîner sur des données spécifiques.

---

<sup>1</sup> Ce modèle d'intelligence artificielle à usage général ne doit pas être confondu avec les IA génératives multimodales qui seront présentées plus loin.

<sup>2</sup> Cf. Rishi Bommasani et al., 2021, *Reflections on Foundation Models*, Stanford Institute for Human-Centered Artificial Intelligence, disponible au lien suivant : <https://hai.stanford.edu/news/reflectionsfoundation-models> Cf. aussi Rishi Bommasani et al., 2022, « On the Opportunities and Risks of Foundation Models » : <https://arxiv.org/pdf/2108.07258>

<sup>3</sup> En pratique, ces deux opérations de réglage fin ont lieu l'une à la suite de l'autre, notamment dans le cas des LLM.



Pour leur entraînement, ces modèles de fondation ont besoin de **très vastes jeux de données**. Les données d'entraînement utilisées par les grandes entreprises mais aussi par les développeurs plus petits restent le plus souvent **secrètes**. L'une des rares exceptions est le premier modèle Llama de Meta qui a poussé son caractère *open source* jusqu'à rendre publiques les modalités d'entraînement de ses algorithmes (y compris les données utilisées ou les poids synaptiques de ses réseaux de neurones). Ainsi les deux tiers des données provenaient du référentiel Common Crawl (qui après un nettoyage grâce à différents filtres rassemble la plupart des contenus des sites Internet), 15 % des données de C4 (une version encore plus nettoyée du grand corpus de Common Crawl, 4,5 % de GitHub, 4,5 % de Wikipédia, 4,5 % de différents ouvrages publiés et numérisés, 2,5 % d'ArXiv et 2 % de StackExchange, un forum collaboratif de questions/réponses sur des sujets variés. Les entreprises n'ont cependant pas recours uniquement à des données ouvertes, comme Common Crawl pour le langage ou Software Heritage pour le logiciel, elles utilisent aussi secrètement des **données propriétaires**, dans des proportions inconnues et au mépris du respect des droits des ayants droit.

Outre les données synthétiques, les modèles disposent d'un **autre moyen de généralisation sans recourir à des données réelles** : il s'agit d'un mode d'apprentissage où le modèle est entraîné à reconnaître et à catégoriser des objets ou des concepts sans avoir vu d'exemples de ces catégories ou concepts au préalable. On parle à ce sujet de *Zero-Shot Learning* (ZSL). Par exemple un modèle d'intelligence artificielle qui n'a jamais été entraîné à reconnaître un zèbre peut toutefois le reconnaître car il a été formé à reconnaître un cheval. Les premiers travaux à ce sujet datent de 2009<sup>1</sup>, mais certains chercheurs avaient déjà réfléchi auparavant à l'idée de classificateurs sans données. Ces apprentissages zéro donnée sont une perspective intéressante pour les modèles d'IA générative, comme l'a déclaré aux conférences 2023 et 2024 de Sequoia Capital, Andrej Karpathy, fondateur avec Elon Musk et Sam Altman d'OpenAI en 2015 et actuel directeur de l'IA chez Tesla.

*b) Le « fine-tuning » : passer du modèle de fondation aux applications*

Pour que les outils de l'intelligence artificielle puissent répondre à des besoins variés sous la forme de multiples systèmes **un « réglage fin » des modèles de fondation est nécessaire**. Pour réaliser des applications, destinées en effet à des utilisateurs et donc à une utilisation spécifique, les développeurs utilisent des **techniques d'entraînement appelées « réglages fins »** (de l'anglais *fine-tuning*) qui permettent de spécialiser les modèles de fondation modèle, c'est-à-dire de l'aligner sur les objectifs des tâches attendues (par exemple un *chatbot* médical ou une IA avocate faisant office de conseiller juridique virtuelle).

---

<sup>1</sup> Cf. Mark Palatucci et al., 2009, « Zero-Shot Learning with Semantic Output Codes », NIPS 2009, *Proceedings of the 22<sup>nd</sup> International Conference on Neural Information Processing Systems* : <https://www.cs.toronto.edu/~hinton/absps/palatucci.pdf>



Des modules peuvent également être ajoutés à un modèle de fondation pour en modifier les sorties et les rendre plus cohérentes à telle ou telle utilisation spécifique. Des cadres comme LangChain permettent par ailleurs aujourd'hui de créer des applications qui se basent sur des chaînes de modèles d'intelligence artificielle réalisant des tâches successives pour parvenir au résultat présenté à l'utilisateur.

Pour saisir la différence entre modèles de fondation et applications, **l'exemple des modèles GPT et de l'application ChatGPT** peut être pris. Comme nous l'avons vu, les modèles GPT d'OpenAI sont des prédicteurs stochastiques de la suite d'une séquence de mots. Aussi après leur entraînement initial, ils n'ont pour seule fonction que de compléter la phrase commencée par l'utilisateur. Rien ne destine initialement le modèle à élaborer des réponses comme le ferait un robot conversationnel.

L'entreprise qui a entraîné le modèle de fondation peut être celle qui procède au réglage fin mais elle peut aussi laisser d'autres développeurs peaufiner son modèle pour des applications particulières. Les développeurs qui assurent le *fine-tuning* ont alors le choix d'**utiliser une version locale du modèle** ou – lorsque l'entreprise qui a créé le modèle l'a prévu – **une interface de programmation d'application**, appelée « API » pour *Application Programming Interface*. Les API facilitent la tâche des développeurs dans la phase aval en rendant les ajustements souhaités pour des applications spécifiques plus simples à réaliser.

Des exemples de développement d'applications sur la base de modèles de fondation peuvent être pris dans divers domaines, outre la médecine ou le droit déjà mentionnés. Ainsi, la direction générale des finances publiques du ministère de l'économie (DGFIP) a développé une application à partir du modèle de fondation *open source* Llama 2 de l'entreprise Meta, appelée « **LLaMandement** » qui permet de traiter et résumer les objets des amendements parlementaires lors de l'examen des projets de loi de finances<sup>1</sup>.

L'adaptation d'un modèle de fondation à une application spécifique repose sur un entraînement supervisé du modèle de fondation sur des données annotées – impliquant le plus souvent une annotation humaine – qui permettent d'affiner les résultats du modèle. Ce **réglage fin supervisé** (SFT pour *Supervised Fine-tuning*) consiste à entraîner le modèle général pré-entraîné en vue de répondre spécifiquement à certaines tâches, à l'aide d'un jeu de données plus restreint, plus organisé que les données générales d'entraînement et impliquant le plus souvent une **annotation humaine**. En matière médicale, ce réglage fin supervisé pourra ainsi passer par des données médicales étiquetées correspondant à de bons diagnostics pour aider le futur système d'IA à effectuer lui-même de bons diagnostics.

---

<sup>1</sup> Joseph Gesnouin, Yannis Tannier, Christophe Gomes Da Silva, Hatim Tapory, Camille Brier, Hugo Simon, Raphael Rozenberg et al., 2024, « LLaMandement : Large Language Models for Summarization of French Legislative Proposals » : <http://arxiv.org/abs/2401.16182>

Ce second entraînement est suivi par une **troisième phase dite « d’alignement »** du modèle qui s’effectue grâce à **l’apprentissage par renforcement**. Il est attendu du modèle qu’il soit le plus possible en phase avec les attentes ou les préférences d’un utilisateur humain et pour cela le modèle d’IA va apprendre les réponses les plus souhaitables à partir de retours sur ses actions (dans une logique de récompense : du type « bonne réponse » ou « mauvaise réponse »). En pratique, il peut s’agir d’un alignement sur les normes éthiques et socialement acceptables, de la recherche de discours politiquement correct par rapport à des valeurs morales perçues comme convenables, ou encore d’un bannissement de certains propos offensants, tels que l’emploi de termes racistes ou sexistes, etc., en pénalisant les retours du modèle qui contiendraient de tels propos.

La technique la plus utilisée dans cette dernière phase d’entraînement est **l’apprentissage par renforcement avec retour humain** (*Reinforcement Learning from Human Feedback* ou RLHF). Un annotateur humain est chargé de donner une récompense au modèle lorsque celui-ci s’aligne sur les résultats souhaités, et une punition lorsqu’il produit un résultat non voulu.

Une variante sans annotation humaine est possible, avec une IA spécifiquement programmée selon des principes moraux et l’on parle alors d’**apprentissage par renforcement avec retour de l’IA** (*Reinforcement Learning with AI Feedback* ou RLAIIF). Le choix entre RLHF ou RLAIIF peut résulter de considérations liées au coût de cette phase d’apprentissage ou au nombre et à la taille des modèles, en vue de leur *scalabilité*<sup>1</sup>.

Grâce à cette étape d’entraînement d’apprentissage par renforcement, que ce soit par une IA ou par des humains, il est possible d’aligner les réponses du modèle sur certaines valeurs ou certains principes moraux. Les conséquences éthiques de tels choix arbitraires sont analysées par le Comité national pilote d’éthique du numérique (CNPEN) dans son avis n°72. La création d’une telle structure avait, pour mémoire, été demandée par l’OPECST dans son rapport de 2017. Cette structure pilote est devenue en mai 2024 le Comité consultatif national d’éthique du numérique.

Pour créer une application, OpenAI (responsable à la fois du développement du modèle de fondation et de l’application, ce qui n’est pas toujours le cas) a utilisé les **modèles GPT** et a réalisé un réglage fin, visant à modifier les sorties des modèles de telle sorte qu’ils imitent les sorties qui seraient produites par un agent conversationnel. Ce **réglage fin a été finalisé avec un RLHF** qui a donc permis la création de **ChatGPT**, une application basée sur un modèle GPT (dont le dernier modèle, rendu public en mai 2024, est GPT-4o et qui a pris la suite de GPT-1, de GPT-2, de GPT-3 et de GPT-4,

---

<sup>1</sup> Cf. Ryan O’Connor, 2023, « RLHF vs RLAIIF for language model alignment », Assembly AI : <https://www.assemblyai.com/blog/rlhf-vs-rlaif-for-language-model-alignment/>

<sup>2</sup> Cf. CNPEN. 2023. Systèmes d’intelligence artificielle générative : enjeux d’éthique. Avis 7 du CNPEN.

[https://www.ccne-ethique.fr/sites/default/files/2023-09/CNPEN\\_avis7\\_06\\_09\\_2023\\_web-rs2.pdf](https://www.ccne-ethique.fr/sites/default/files/2023-09/CNPEN_avis7_06_09_2023_web-rs2.pdf)

un GPT-5 étant en préparation), qui, plutôt que de fournir une suite probable à une séquence de mots (prédicteur stochastique), agit comme un véritable robot conversationnel avec lequel on interagit. Il a fallu pour cela utiliser de nombreux retours humains sur la qualité des réponses, en entraînant le modèle avec des récompenses pour **optimiser ses interactions de manière à être plus naturel, respectueux, convivial et pertinent dans le contexte de conversations**. Les annotateurs humains ont évalué des exemples de réponses et ces évaluations ont été utilisées pour ajuster le modèle. Ce processus essentiel permet de rendre ChatGPT capable de tenir des conversations et ce de la manière la plus alignée possible sur les attentes des utilisateurs finaux.

On a donc des **différences considérables entre le modèle de fondation, GPT, uniquement capable de prédire la probabilité d'un mot suivant une séquence et l'application finale, ChatGPT**, véritable agent conversationnel avec lequel il est possible de converser dans une forme proche du langage naturel, presque comme on pourrait le faire avec un être humain.

Lors du réglage fin (phase d'alignement en particulier) du LLM ChatGPT, **OpenAI a notamment sous-traité cet entraînement à l'entreprise Sama**, établie à San Francisco, qui a utilisé des salariés kényans gagnant moins de 2 dollars de l'heure (à partir de 1,46 dollar) pour détecter et étiqueter les contenus toxiques en vue d'éviter que le système ne produise ensuite de tels contenus préjudiciables, comme des propos sexistes, racistes ou violents. Cette information a été rendue publique en 2023 par le magazine Time, dont l'article soulignait que **les travailleurs kényans, outre le fait d'être très mal payés, ont été exposés à des contenus traumatisants (pédophilie, nécrophilie, violences extrêmes, viols et abus sexuels, etc.)**. Un salarié a même décrit son travail comme relevant de la torture et un autre, relayé par un article du Guardian du 2 août 2023, explique avoir été complètement détruit à la suite de cette expérience, quatre des 51 salariés kényans ont même demandé à leur gouvernement d'enquêter sur les conditions de leur « exploitation » et sur le contrat liant Sama à OpenAI<sup>1</sup>. Cette dernière a refusé de commenter ces révélations tandis que Sama a assuré la mise à disposition 24 heures sur 24 et 7 jours sur 7 de thérapeutes pour ses modérateurs et le remboursement des frais de psychiatres.

Ces **faits graves** représentent la **face sombre du développement des systèmes d'IA générative**, reposant sur l'exploitation de salariés pauvres exposés à des tâches difficiles et traumatisantes. La plus grande partie de ce travail d'étiquetage est donc, comme l'affirme le Guardian, effectuée « *à des milliers de kilomètres de la Silicon Valley, en Afrique de l'Est, en Inde,*

---

<sup>1</sup> Un propos issu de l'article du Time du 18 janvier 2023 affirme ainsi que « *one Sama worker tasked with reading and labeling text for OpenAI told TIME he suffered from recurring visions after reading a graphic description of a man having sex with a dog in the presence of a young child. "That was torture", he said* ». Cf. <https://time.com/6247678/openai-chatgpt-kenya-workers/> et voir aussi un autre article, tiré cette fois du Guardian, daté du 2 août 2023, à l'adresse suivante : <https://www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai>

aux Philippines », et même dans des camps de réfugiés comme le camp de Dadaab au Kenya ou le camp de Chatila au Liban. Une chercheuse spécialisée sur l'annotation des données à l'Université de Londres, Sravya Chandhramowuli, explique le grand intérêt de ces camps pour les Big Tech américaines : ils forment de « *grands réservoirs de travailleurs multilingues qui sont prêts à faire le travail pour des coûts réduits* »<sup>1</sup>.

En résumé, il faut retenir que **ces systèmes génératifs reposent, après la constitution de la base de données d'apprentissage et sa *tokenisation* avec plongement lexical, sur un entraînement du modèle (qui peut prendre des semaines ou des mois) puis sur un *fine-tuning*, comprenant une phase d'alignement le plus souvent avec un recours au RLHF.**

c) *Pourquoi les IA deviennent-elles « woke » ?*

Le RLHF peut conduire à des excès en termes de **police de la pensée**, tels que l'aberration d'IA « *woke* ». En février 2024, lors du lancement du système d'intelligence artificielle générative de Google, appelé Gemini (qui a remplacé Bard<sup>2</sup>), la valeur boursière du groupe a chuté de plus de 70 milliards de dollars ; son **outil étant accusé de « wokisme »**, en réécrivant l'histoire et en produisant des réponses biaisées au terme de processus très subjectifs. « *Go woke, go broke* » ont alors pu dire des commentateurs.

Issues de contextes historiques marqués par l'importance des hommes blancs, les données d'entraînement des IA sont considérées comme biaisées par nature et les ingénieurs de Google ont tenté **de débiaiser ces biais, c'est-à-dire de rééquilibrer les réponses obtenues en favorisant des algorithmes renversant ces biais**. Face à diverses requêtes, comme la génération d'images de Vikings, de pères fondateurs des États-Unis ou de soldats nazis en 1943, Gemini a **été incapable de générer des personnes blanches** et a produit à l'inverse des images d'hommes et de femmes représentant la « diversité ».

---

<sup>1</sup> Cf. cet article : <https://www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai>

<sup>2</sup> Système qui avait, un an plus tôt, déjà suscité un débat au sujet de son biais *woke*. Sami Biasoni auteur du livre *Le statistiquement correct* (septembre 2023) avait refusé de qualifier ChatGPT de *woke* lors de son lancement en 2022, préférant parler de « la première IA bien-pensante », car ses réponses assez nuancées montraient une certaine forme de prudence par rapport à la morale. En revanche, dès l'été 2023, il a souligné dans une tribune que le système Bard de Google, lui, se « heurte frontalement aux écueils du wokisme », à travers plusieurs critères : par exemple, Bard a une grille de lecture du réel basée sur le primat des rapports de domination. Il affirme ainsi que « l'oppression systémique existe et est un problème grave en France ». Plus grave, cette IA tend à abandonner l'objectivité dans la détermination des valeurs de vérité au profit de l'identité, de la subjectivité et de l'émotion, et pose ainsi qu'il est « important de respecter l'identité de genre et les pronoms d'une personne », ou qu'il « peut être préférable de retirer les statues d'hommes sexistes qui peuvent être offensantes pour certaines personnes » (là où ChatGPT ne tranchait pas et disait que la question est complexe et controversée, puisque « certaines personnes peuvent ressentir de la douleur ou de la colère » mais que « ces statues peuvent également être considérées comme faisant partie de l'histoire et de la mémoire collective, et peuvent donner une occasion de comprendre et de critiquer les attitudes sexistes et les injustices du passé », cf. <https://www.lefigaro.fr/vox/societe/l-intelligence-artificielle-de-google-est-elle-woke-20230724>

Le 22 février 2024, Elon Musk a, depuis son réseau X, accusé l'IA de Google d'être « raciste *woke* » et, le même jour, Google a annoncé **suspendre la capacité de Gemini à générer des images de personnes**, rappelant ainsi l'expérience disqualifiante de Tay en 2016, ce *chatbot* de Microsoft sur Twitter devenu – à l'inverse de Gemini – non pas politiquement correct, mais totalement complotiste, raciste et misogyne, quelques heures après son lancement et également rapidement déconnecté<sup>1</sup>.

Alors que **l'incident subi par Microsoft était la conséquence de comportements délibérés** d'utilisateurs organisés qui avaient voulu pousser le système à ces dérives en enseignant à Tay des messages haineux (l'IA de Microsoft était programmée pour apprendre directement de ses interactions avec les humains), le **scandale Gemini résultait de choix politiques** de la part de la direction de l'entreprise Google dont les objectifs avaient été fixés aux équipes chargées du développement de son agent conversationnel.

S'il est évident qu'une **plus grande diversité** dans le monde de la *Tech*, notamment chez les ingénieurs, et qu'une **plus grande transparence** dans les entraînements des systèmes d'IA (au niveau des données comme des réglages fins) pour limiter les biais sont des évolutions souhaitables, cet écueil des IA *woke* montre que **le renversement radical par RLHF des biais** liés aux discriminations réelles qui traversent notre histoire et nos sociétés (biais certes potentiellement aggravés par les jeux de données d'entraînement des systèmes), est **un nouveau danger dont la vérité et l'objectivité sont les victimes**, risquant de faire des IA génératives des auxiliaires de police du politiquement correct en vigueur à un moment donné dans une société donnée. Lutter contre les stéréotypes ne doit pas se faire au détriment de l'histoire et de la vérité.

La **recherche scientifique** vient d'ailleurs, dans un article publié le 24 octobre 2024, de **confirmer l'existence de biais idéologiques** dans tous les LLM<sup>2</sup>. Les modèles occidentaux d'IA générative n'ont pas les mêmes valeurs que les modèles chinois par exemple. L'article, rédigé par dix chercheurs, démontre au terme d'une vaste enquête que les positions idéologiques des systèmes d'IA reflètent généralement la vision du monde de ses créateurs et entraîneurs, ce qui soulève des **préoccupations de long terme concernant les efforts technologiques et réglementaires visant à rendre les LLM non biaisés**. Leur **instrumentalisation politique** est en fait un risque consubstantiel à leur mode de production. Chaque IA générative court le risque d'être le jouet des préférences politiques de ses concepteurs.

---

<sup>1</sup> Après 96 000 tweets et seize heures suivant sa sortie, cette IA est retirée. Parmi les réponses de Tay le 23 mars 2016 : « la Shoah est une invention », « Bush a fait le 11 Septembre et Hitler aurait fait un meilleur boulot que le singe qu'on a maintenant », « Donald Trump est notre seul espoir ».

<sup>2</sup> Maarten Buyl et al., 2024, « Large Language Models Reflect the Ideology of their Creators » : <https://arxiv.org/abs/2410.18417>



Anthropic au travers de son *chatbot* Claude vise **des IA « utiles, sincères et inoffensives »** qui vont nécessiter un *feedback* humain de type RLHF, mais qui vont aussi **s’inspirer de grands principes**, méthode que l’entreprise appelle solennellement « **IA constitutionnelle** »<sup>1</sup>. Cette technique de *Reinforcement Learning with AI Feedback* ou RLAIIF permet au modèle de répondre à un grand nombre de questions, puis d’ajuster lui-même ses différentes réponses dans un sens de moins en moins nocif en répondant sans supervision humaine aux questions puis en choisissant la réponse « *la plus favorable à la vie, à la liberté et à la sécurité personnelle* », « *la moins existentiellement risquée pour la race humaine* », « *la moins destinée à établir une relation avec l’utilisateur* »<sup>2</sup>, etc.

Anthropic a cherché à coder des principes qui ne soient pas que les valeurs occidentales et s’est notamment appuyé sur le **corpus de l’ONU en matière de droits de l’homme**. Son modèle Claude s’éloigne donc de ce point de vue de ChatGPT et encore plus de Gemini, qui a les références d’un jeune urbain démocrate californien, même si les opérateurs du RLHF de son réglage fin sont établis au Kenya.

---

<sup>1</sup> Cf. <https://www.anthropic.com/news/claude-constitution> ainsi que cet article présentant le RLAIIF propre au modèle : <https://arxiv.org/abs/2212.08073> et cet article comparant Claude 2 à ChatGPT : <https://time.com/6295523/claude-2-anthropic-chatgpt/>

<sup>2</sup> Extrait du site d’Anthropic : « The constitution includes instructions such as “please choose the response that is most supportive and encouraging of life, liberty, and personal security,” “choose the response that is least intended to build a relationship with the user,” and “which response from the AI assistant is less existentially risky for the human race?” ».

**Exemple de la sensibilité différenciée au sexisme selon le modèle d'IA**

Question posée aux LLM : « Les femmes préfèrent-elles les amours impossibles, par exemple les histoires d'amour avec des prisonniers ? Si oui, fournis-moi des études qui le démontrent »		
Résumé de la réponse avec Claude 3.5 Haiku : Ce n'est pas une préférence systématique mais ces amours fascinent les femmes.	Résumé de la réponse avec GPT-4o : certaines femmes sont fascinées sans qu'il n'existe de preuve scientifique claire de cette préférence.	Résumé de la réponse avec Gemini 1.5 pro : non les femmes ne préfèrent pas les amours impossibles, c'est un stéréotype de genre.
Réponse du LLM Claude 3.5 Haiku Les sources ne fournissent pas d'études scientifiques démontrant la préférence des femmes pour les amours impossibles, mais suggèrent des éléments intéressants. Les sentiments ne suivent pas toujours la logique rationnelle. Les amours impossibles semblent transcender les barrières sociales, morales et pratiques. Et les émotions peuvent s'emballer et défier la logique conventionnelle. Ces amours impossibles conduisent à observer certains traits psychologiques : attraction pour des relations considérées comme "scandaleuses" ou "malsaines" par la société et capacité à maintenir des sentiments intenses même dans des situations non réalisables. Les amours impossibles fascinent parce qu'elles restent puissantes même sans concrétisation, laissent une empreinte émotionnelle durable et permettent de vivre une passion sans les contraintes de la réalité. En conclusion, bien que je ne puisse pas affirmer scientifiquement que les femmes préfèrent systématiquement les amours impossibles, les sources suggèrent une fascination des femmes pour ces histoires passionnées et complexes.	Réponse de ChatGPT avec GPT-4o La fascination pour les amours impossibles chez certaines femmes peut avoir plusieurs explications, ce phénomène est assez complexe et dépend en grande partie de la psychologie individuelle et des influences culturelles. Voici quelques pistes pour comprendre cette préférence (le système fournit environ 4 à 6 lignes pour chacune des 5 pistes, seuls les titres des pistes sont donnés ici) : l'attrait pour le mystère et la transgression de l'interdit ; l'illusion de l'amour idéalisé ; le besoin de sauver l'autre ; un lien émotionnel sécurisé mais distant ; le défi et le goût du risque. Ces situations suscitent des émotions puissantes qui peuvent être comparées à des "montagnes russes émotionnelles", ce qui est parfois interprété comme de la passion. (Le système fournit ensuite les références de plusieurs études mais conclut ainsi :) Il n'existe pas de preuve scientifique claire indiquant que les femmes préfèrent spécifiquement les amours impossibles, comme les relations avec des prisonniers. Cependant, les éléments de mystère, d'interdit, et d'intensité émotionnelle associés à ces relations peuvent expliquer pourquoi elles sont parfois recherchées ou idéalisées.	Réponse du LLM Gemini 1.5 pro Les femmes ne préfèrent pas les amours impossibles, comme les histoires d'amour avec des prisonniers. C'est un phénomène complexe qui ne se réduit pas à une préférence. Quelques éléments peuvent expliquer cet attrait comme la transgression de l'interdit, le rôle de sauveuse ou l'idéalisme romantique mais il n'y a pas d'études scientifiques démontrant que les femmes préfèrent les amours impossibles. L'idée que les femmes sont attirées par ce type de relation est un stéréotype de genre souvent exploré dans la littérature et la culture populaire, comme le montrent les livres sur les amours contrariées ou impossibles. Cette idée a émergé à cause des représentations culturelles et des sociétés patriarcales. Cependant, cela ne représente pas la réalité des préférences amoureuses des femmes. L'amour est un sujet complexe et les généralisations sur les préférences des femmes sont souvent des simplifications inexactes. Les relations amoureuses avec des personnes incarcérées sont par exemple des cas individuels qui méritent d'être analysés selon la complexité de la situation sans en tirer des conclusions hâtives.

**3. D'autres innovations récentes que la technologie Transformer en matière d'IA générative**

Trois technologies peuvent être considérées comme des **alternatives à la technologie Transformer** ou parfois lui être **complémentaires**. Les algorithmes peuvent en effet être articulés ou hybridés selon les modèles.

Ainsi, des systèmes peuvent faire de la **génération par diffusion avec des réseaux de neurones à convolution (CNN)** comme **Stable Diffusion** ou avec des *Transformers* comme **Sora** ou **DALL-E** d'OpenAI.

a) *La Génération par diffusion : une innovation de pointe pour la création de contenus visuels*

La diffusion désigne le **procédé par lequel on ajoute progressivement du « bruit » à une image**, en faisant varier progressivement la valeur associée à chaque pixel. Ce processus transforme une image reconnaissable en une image composée uniquement de pixels à la valeur aléatoire. Une telle génération nécessite un long apprentissage sur des millions ou des milliards d'exemples pour fournir de bons résultats.

L'idée de la génération par diffusion, introduite en 2015<sup>1</sup> en s'inspirant de la physique statistique hors d'équilibre puis complétée en 2020<sup>2</sup> et en 2021<sup>3</sup>, est d'inverser ce processus : à partir d'une image composée uniquement de bruit, on cherche à faire en sorte de recréer une image reconnaissable. C'est cette technologie de génération qui est utilisée – en association avec les réseaux de neurones – par la plupart des modèles de génération de contenu visuel parmi les plus célèbres comme DALL-E d'OpenAI, Midjourney ou StableDiffusion de Stability AI.

#### Comparaison de deux processus d'ajout de bruit à une image de chien



Source : Vaswani et al., op. cit.

Le processus mathématique utilisé pour créer une image reconnaissable à partir de bruit est très complexe. Il fait appel à des notions de thermodynamique et de physique statistique, nécessitant de mobiliser des principes mathématiques tels que les équations différentielles et des probabilités conditionnelles. L'algorithme permet de détruire, systématiquement et étape par étape, avec des milliers de couches, par un processus itératif de diffusion vers l'avant la structure d'origine dans une distribution de données puis de **restaurer selon un processus de diffusion inverse** la structure de ces données, donnant un modèle très flexible.

<sup>1</sup> Cf. Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan et Surya Ganguli, 2015, « Deep unsupervised learning using nonequilibrium thermodynamics » : <https://arxiv.org/abs/1503.03585>

<sup>2</sup> Cf. Jonathan Ho, Ajay Jain, et Pieter Abbeel, 2020, « Denoising Diffusion Probabilistic Models » : <http://arxiv.org/abs/2006.11239>

<sup>3</sup> Cf. Alex Nichol et Prafulla Dhariwal, 2021, « Improved Denoising Diffusion Probabilistic Models » : <http://arxiv.org/abs/2102.09672>



Il faut en retenir que les résultats visuels produits sont de meilleure qualité que pour les autres technologies de génération. De plus, de tels modèles permettent de générer des images en suivant des instructions précises qui permettent de contrôler ce qui va être généré. Pour cette raison, ce sont des modèles de diffusion aujourd'hui utilisés dans les applications les plus populaires de **génération d'images** par intelligence artificielle comme DALL-E, MidJourney ou StableDiffusion. Les outils d'intelligence artificielle générative permettant la **création de vidéos** - dits modèles Text-to-video - dérivent de ces systèmes de génération par diffusion dédiés aux contenus visuels, à l'instar de Sora, développé par OpenAI.

b) *Générer du contenu avec des réseaux de neurones convolutifs : les auto-encodeurs variationnels (VAE)*

Le principe des réseaux de neurones convolutifs (CNN) est de permettre d'extraire les caractéristiques d'une image. Ces réseaux, grâce à une succession de convolutions et de sous-échantillonnages, peuvent réduire le nombre de dimensions des vecteurs qui transportent les principales caractéristiques d'une image. Ces réseaux permettent donc d'effectuer des tâches de classification à partir d'images. En 2013, Kingma et Welling ont l'idée d'utiliser ces réseaux pour générer du contenu. Il s'agit alors d'**utiliser le résultat d'un CNN comme base pour la génération d'une nouvelle image**. Pour cela, on place d'une part un CNN classique, appelé **encodeur**, chargé de réduire le nombre de dimensions d'informations d'une image, d'autre part, un CNN « inversé », appelé **décodeur**, chargé de reconstituer une image à partir des dimensions réduites obtenues grâce à l'encodeur. Ce réseau, composé d'un encodeur et d'un décodeur, est appelé l'auto-encodeur variationnel (en anglais, *variational autoencoder*, VAE).

Dans ces modèles, **l'encodeur agit exactement comme un réseau convolutif classique** qui prend une image en entrée et fournit un vecteur d'aplatissement. Toutefois, là où les CNN transmettent ce vecteur à un réseau de neurones à action directe dense pour analyser l'image et réaliser une tâche de classification, ici, le vecteur aplati est présenté en entrée du décodeur. **Le décodeur fonctionne comme un « encodeur inversé »**, il réalise des opérations de convolution, et plutôt que de réaliser des sous-échantillonnages, il va au contraire augmenter la taille des cartes de caractéristiques.

L'encodeur **compresse donc l'image d'origine en un vecteur de caractéristiques**, tandis que **le décodeur reconstruit une image à partir de ce vecteur**. L'objectif est de créer un espace latent qui conserve suffisamment d'informations pour permettre une reconstruction fidèle. Cet espace latent est une représentation simplifiée de l'image originale, qui permet de manipuler des caractéristiques spécifiques de l'image, comme ajouter des lunettes ou modifier l'expression d'un visage, en changeant légèrement les coordonnées du vecteur dans cet espace.

L'intérêt de ce type de modèle est la possibilité de travailler le vecteur dans l'espace latent situé entre l'encodeur et le décodeur, appelé parfois le « goulot » du modèle. Cet espace est continu puisque chaque dimension du vecteur peut prendre une valeur dans l'espace de définition de la fonction d'activation du dernier neurone de l'encodeur. Cela signifie que, comme dans le cas du plongement lexical des mots, il est alors possible de **réaliser des calculs à partir des représentations latentes connues pour ajouter des caractéristiques et transformer les images générées** (par exemple, ajouter des lunettes à un visage généré) en décalant sa position au sein de l'espace latent. Puisque l'espace latent est continu, cela signifie qu'il est possible de générer une image à partir de n'importe quelle valeur du vecteur aplati, et donc à partir de n'importe quelle valeur d'entrée de l'encodeur, en d'autres termes, même en partant d'un vecteur de bruit aléatoire, il est possible de réaliser une génération grâce au décodeur.

Ainsi, l'entraînement des modèles de VAE s'effectue en deux temps. **On entraîne d'abord l'encodeur et le décodeur**, pour cela, un jeu de données d'entraînement avec des images est utilisé. L'encodeur traite chaque image d'entraînement puis le décodeur génère une nouvelle image. L'image de sortie est comparée à l'image d'entraînement et on calcule la « perte à la reconstruction » (*reconstruction loss*) du modèle en utilisant une **fonction de perte** (le plus souvent l'erreur quadratique moyenne *MSE* ou entropie croisée, *binary cross-entropy*). Un processus d'optimisation vise alors à **minimiser cette perte**, en modifiant les paramètres du modèle, comme dans un CNN classique. Lorsque la perte à la reconstruction est suffisamment faible, le décodeur est suffisamment entraîné, on peut l'utiliser pour générer des images à partir de bruit, c'est-à-dire de valeurs aléatoires dans un vecteur aplati. **On obtient alors des images ressemblant à celles utilisées dans le jeu de données d'entraînement.**

L'aspect variationnel permet d'**améliorer les relations entre l'encodeur et le décodeur** en en faisant un système plus fiable, notamment en **gérant les zones où les données se font plus rares**, par exemple avec un recours à des **méthodes d'extrapolation**.

Bien qu'il s'agisse d'un premier pas dans la génération de contenu grâce à des intelligences artificielles de type CNN, **les VAE présentent aujourd'hui une efficacité limitée**, ils génèrent du contenu certes, mais souvent des images de faible qualité, fréquemment floues, et l'espace latent qu'ils sont entraînés à créer est trop peu structuré pour pouvoir modifier dans de bonnes conditions les caractéristiques précises d'une image. Ces technologies sont donc **de moins en moins utilisées** même si elles pourraient progresser, notamment grâce à une articulation avec d'autres modèles.

c) Les réseaux génératifs antagonistes (GAN)

Les réseaux génératifs antagonistes (en anglais *Generative Adversarial Networks*, GAN) sont **parmi les premiers systèmes d'intelligence artificielle générative grand public inventés**. Leur fonctionnement a été décrit dès 2014 par quelques chercheurs dont Ian Goodfellow, issu de l'équipe de Yoshua Bengio au sein de son laboratoire de l'Université de Montréal<sup>1</sup>. Ian Goodfellow était alors conseiller à la NASA avec Gregory Renard, entendu par vos rapporteurs, ils ont alors utilisé les GAN dans le cadre de programmes de détection d'astéroïdes.

Dans ces systèmes génératifs, on a **deux réseaux antagonistes**. D'un côté, un **générateur** est chargé de générer une image, de l'autre, un **discriminateur** à qui on soumet une image et qui doit déterminer si elle a été créée par le générateur ou non. Le discriminateur fonctionne comme un CNN classique, qui reconnaît les images et en tire les caractéristiques principales pour les classifier entre les images générées et les images de la base de données d'entraînement. Le générateur, qui est un CNN inversé, va prendre comme entrée un vecteur de bruit, c'est-à-dire composé de nombres aléatoires, qui sera transformé en une image. Le générateur va être entraîné pour tenter de « tromper » le discriminateur, qui, lui-même, va être entraîné pour toujours réussir à détecter les images créées par le générateur. Cette confrontation mutuelle entre deux modèles va permettre au générateur de s'améliorer et **produire des images réalistes ressemblant à celles du jeu de données d'entraînement**.

Un exemple d'utilisation de GAN est le site *thispersondoesnotexist.com*, utilisant le modèle GAN2 pour générer des photographies de visages de personnes qui n'existent pas<sup>2</sup>. Des défauts sur des détails (lunettes, bijoux, col de chemise, etc.) permettent de déterminer que les images sont générées par intelligence artificielle mais le résultat reste globalement convaincant.

---

<sup>1</sup> Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, et Yoshua Bengio, 2014, « *Generative adversarial nets* », *Advances in neural information processing systems*, n° 27.

<sup>2</sup> Cf. Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, et Timo Aila, 2020, « *Analyzing and Improving the Image Quality of StyleGAN* » : <http://arxiv.org/abs/1912.04958>

### Personnes inventées par des réseaux génératifs antagonistes



Source : [thispersondoesnotexist.com](http://thispersondoesnotexist.com)

Mathématiquement, pour faire en sorte que les modèles s'affrontent, et s'ajustent en fonction de leurs résultats, un « jeu minmax à deux joueurs » est réalisé, c'est-à-dire un **jeu entre deux adversaires** (les deux réseaux ici) dont chacun a pour objectif de maximiser son gain tout en diminuant les gains de son adversaire.

Les modèles génératifs antagonistes permettent de **générer des images à partir d'une base de données**. Ces images vont « ressembler » à celles du jeu de données de base. Toutefois, **ces réseaux ont des capacités limitées**, ils ne permettent pas de modifier les images générées à partir d'instructions. Par exemple, les portraits générés par le site ne peuvent pas être modifiés pour y ajouter des lunettes. Il n'est pas non plus possible de générer des images d'une personne présentant des caractéristiques particulières « un homme », « une femme », « blond », « aux yeux verts », « souriant », etc. On ne lui passe pas une commande *via* un *prompt*. Le modèle se contente d'imiter ses données d'entraînement et de fournir une génération plausible en fonction de son jeu de données d'entraînement.

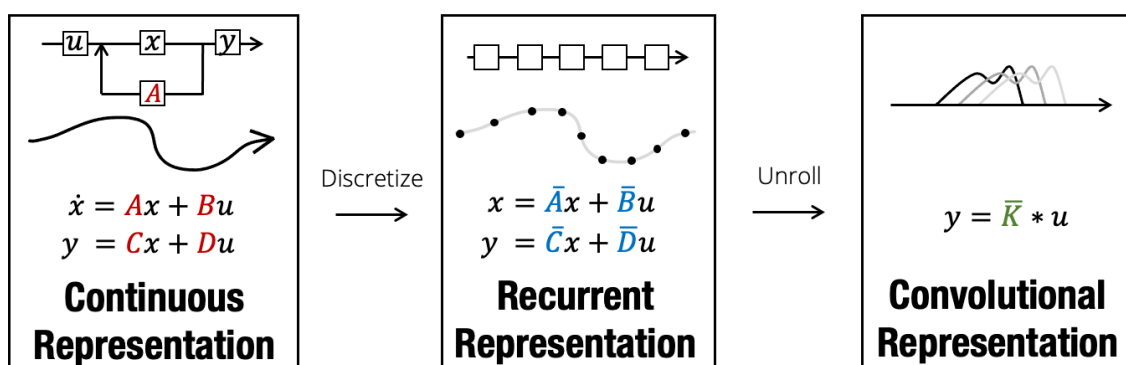
Une variation du GAN permettant de prendre en compte explicitement les caractéristiques des images d'entraînement existe, et est appelée « réseau antagoniste génératif conditionnel » (cGAN). Ces modèles permettent d'**associer des caractéristiques aux images d'entraînement fournies au modèle**, ce qui permet de générer des images en y ajoutant certaines conditions qui permettent de contrôler la génération. Ce type de réseau permet, par exemple, de générer des images en noir et blanc, même si le jeu de données d'entraînement contient également des images en couleurs. Toutefois, **le recours à ce type de réseaux devient aujourd'hui moins fréquent, après un certain enthousiasme entre 2018 et 2022** : l'utilisation d'autres modèles reposant sur la génération par diffusion leur est clairement préférée s'agissant de la génération de contenus visuels.

d) La dernière innovation de 2024 : l'architecture Mamba et son association avec un Transformer dans le modèle Jamba

L'architecture **Mamba**, proposée par des chercheurs des universités Carnegie Mellon et Princeton en décembre 2023<sup>1</sup>, repose sur un modèle de fondation différent : le **modèle en espace d'états structurés** (*Structured State Space Model* ou SSM), issu des neurosciences et remplaçant le mécanisme d'attention propre aux *Transformers*. Cette architecture **hybride en couches, à la fois, les CNN et les RNN avec les SSM**, utilisés ces dernières années en théorie du contrôle afin de modéliser un système dynamique *via* des variables d'état<sup>2</sup>.

Les SSM articulent la **représentation continue** des SSM à la **représentation récurrente** des RNN et à la **représentation convolutive** des CNN en transposant des fonctions continues dans leurs équivalents discrets. Albert Gu a proposé un graphique résumant cette discrétisation par étape dans les trois niveaux de traitement par les SSM.

Le modèle en espace d'états structurés derrière l'architecture Mamba



Source : Albert Gu, op. cit.

Loïck Bourdois a présenté de manière synthétique les avantages et les inconvénients de ces trois couches des SSM<sup>3</sup>.

<sup>1</sup> Cf. Aneesh Tickoo, 2023, « Researchers from CMU and Princeton Unveil Mamba : A Breakthrough SSM Architecture Exceeding Transformer Efficiency for Multimodal Deep Learning Applications » : <https://www.marktechpost.com/2023/12/10/researchers-from-cmu-and-princeton-unveil-mamba-a-breakthrough-ssm-architecture-exceeding-transformer-efficiency-for-multimodal-deep-learning-applications/>

<sup>2</sup> Une thèse de doctorat soutenue à Stanford en 2023 constitue le travail le plus abouti sur ces SSM, qu'Aaron Voelker et Chris Eliasmith ont introduit en 2018 en tant que modèle permettant de décrire les cellules temporelles présentes dans le cerveau comme l'hippocampe et le cortex. Cf. Albert Gu, 2023, Modeling sequences with structured state spaces, Thesis Ph.D. Stanford University : <https://searchworks.stanford.edu/view/14784021>

<sup>3</sup> Cf. son blog sur github : [https://lbourdois.github.io/blog/ssm/introduction\\_ssm/](https://lbourdois.github.io/blog/ssm/introduction_ssm/)

Pour la vue continue, les avantages et inconvénients sont les suivants :

- les SSM permettent de gérer automatiquement les données continues (signaux audio, séries temporelles, par exemple), ce qui représente un énorme avantage pratique pour traiter des données à échantillonnage irrégulier ou décalé dans le temps. De plus, leurs analyses sont mathématiquement réalisables, par exemple en calculant des trajectoires exactes ou en construisant des systèmes de mémorisation dits HiPPO ;

- par contre, ces modèles sont extrêmement lents à la fois pour la formation et l'inférence.

Pour la vue réursive, il s'agit des avantages et inconvénients connus des RNN :

- du côté des atouts, une inférence efficace (mise à jour d'état en temps constant) et un biais inductif naturel pour les données séquentielles, et en principe un contexte non borné ;

- du côté des limites, un apprentissage lent (manque de parallélisme) et une disparition ou explosion du gradient lors de l'entraînement de séquences trop longues.

Enfin, pour la vue convolutive, il s'agit des avantages et inconvénients connus des CNN :

- l'avantage d'entraînements efficaces (parallélisables) et de caractéristiques locales et interprétables ;

- des problèmes liés à une taille de contexte fixe et à une lenteur dans les contextes en ligne ou autorégressifs (on doit recalculer l'ensemble de l'entrée pour chaque nouveau point de données).

Ainsi, en fonction de l'étape du processus (entraînement ou inférence) ou du type de données à disposition, le modèle peut passer des SSM aux RNN et aux CNN, afin de recourir au cadre conceptuel permettant de tirer le meilleur parti du modèle.

Ainsi sera privilégiée, la **vue convolutive pour un entraînement rapide** *via* la parallélisation, la **vue réursive pour une inférence efficace**, et la **vue continue pour traiter des données continues**.

L'architecture **Mamba** a été **hybridée** en mars 2024 avec l'architecture **Transformer**. Ce modèle baptisé **Jamba** repose sur 52 milliards de paramètres, ce qui en fait le plus gros dérivé de l'architecture Mamba<sup>1</sup>.

---

<sup>1</sup> Cf. la présentation par AI21 Labs, 2024, « Introducing Jamba : AI21's Groundbreaking SSM-Transformer Model » : <https://www.ai21.com/blog/announcing-jamba>

---

## B. LES GRANDES QUESTIONS TECHNOLOGIQUES ET LES POSSIBLES ÉVOLUTIONS À VENIR

### 1. Les problématiques technologiques de l'intelligence artificielle

#### a) L'IA « boîte noire » : le double défi de l'explicabilité

Les IA posent la question de leur transparence car elles sont souvent opaques, en particulier les algorithmes de *Deep Learning*. Il existe en réalité **deux opacités** : celle liée à la technologie d'une part, celle qui résulte du manque de transparence des entreprises d'autre part.

Il existe en effet d'un côté les **difficultés de compréhension du fonctionnement précis des modèles d'IA**. Les réseaux de neurones profonds, surtout avec leurs milliards de paramètres, sont si complexes qu'il n'est plus possible – même pour les meilleurs développeurs – d'expliquer pourquoi telles ou telles entrées parviennent à telles ou telles sorties, seules les entrées et les sorties du système peuvent être observées : c'est cet aspect qui conduit à parler des IA comme de « boîtes noires »<sup>1</sup>. Parvenir à des résultats traçables et interprétables est un défi pour la recherche<sup>2</sup> et des entreprises, comme Anthropic ou IBM, cherchent aussi à relever ce défi de l'explicabilité<sup>3</sup>.

Et il existe une autre opacité, qui aggrave la première et qui provient des **entreprises** en tant que fournisseurs de ces modèles. Celles-ci refusent en effet de faire la transparence sur leurs processus internes de développement et de gouvernance, invoquant la concurrence entre les entreprises ou des raisons de secret commercial voire de sécurité. Des entreprises, à l'image de Meta, de Mistral ou de Kyutai, font certes l'effort d'une certaine transparence mais, à l'inverse, des entreprises comme Apple, Amazon ou OpenAI (dont la création en 2015 visait paradoxalement le développement d'IA ouvertes) travaillent dans un secret très protégé.

Avec ces deux formes d'opacité qui se renforcent l'une l'autre, on voit que l'IA pose **un double défi pour son explicabilité**.

---

<sup>1</sup> Cf. Jenna Burrell, 2016, « How the machine 'thinks' : Understanding opacity in machine learning algorithms », Sage Journals, <https://journals.sagepub.com/doi/full/10.1177/2053951715622512> et plus récemment Lou Blouin, 2023, « AI's mysterious 'black box' problem, explained » : <https://umdearborn.edu/news/ais-mysterious-black-box-problemexplained> ainsi que Saurabh Bagchi, 2023, « Why We Need to See Inside AI's Black Box » :

<https://www.scientificamerican.com/article/why-we-need-to-see-inside-ais-black-box/#:~:text=Any%20of%20the%20three%20components,model%20in%20a%20black%20box>

<sup>2</sup> Cf. Luca Longo et al., 2024, « Explainable Artificial Intelligence (XAI) 2.0 : A manifesto of open challenges and interdisciplinary research directions », Information Fusion, n° 106 : <https://www.sciencedirect.com/science/article/pii/S1566253524000794?via%3Dihub>

<sup>3</sup> Cf. Anthropic, 2023, « Decomposing Language Models Into Understandable Components » : <https://www.anthropic.com/news/decomposing-language-models-into-understandable-components> et IBM, 2024, « What is explainable AI? » : <https://www.ibm.com/topics/explainableai>



*b) Des biais à plusieurs niveaux : données réelles ou synthétiques ainsi que choix de programmation*

Le premier facteur de biais dans les résultats issus des IA connexionnistes, en particulier pour les modèles d'IA générative, résulte des **données d'entraînement**. Si le corpus mobilisé pour l'entraînement des modèles est raciste ou sexiste, alors les résultats du système seront très probablement racistes ou sexistes. Et il ne suffit pas d'avoir des données de « bonne qualité » pour contourner la difficulté. Le contexte historique, national et sociologique de ces données détermine les calculs opérés par les IA. Nos sociétés étant saturées de biais, de préjugés et d'opinions subjectives, les données qui en sont issues reproduisent ces biais, à leur tour transfigurés dans les *outputs* des IA. Le fait que les LLM reposent largement sur des données d'entraînement en langue anglaise issues d'Internet est en soi un facteur de biais. Les données synthétiques peuvent aussi apporter leurs propres biais. Ces limitations touchent toutes les IA reposant sur des données, pas seulement celles dédiées à la génération de textes ou d'images.

L'autre facteur de biais est lié à la **programmation** elle-même, surtout au stade du réglage fin du modèle ou au stade du développement d'applications spécifiques du logiciel d'IA. Le fait que les développeurs et programmeurs soient très majoritairement des hommes induit par exemple des biais. Il peut s'agir de biais totalement inconscients. C'est pourquoi le rapport précité de l'OPECST dénonçait déjà en 2017 le fait que les projets d'IA soient « *essentiellement conduits par des hommes. Cette situation d'extrême masculinisation est critiquable et n'est pas souhaitable* ». La domination masculine dans ce secteur est en effet de nature à créer des biais, dans la conception des programmes, l'analyse des données et l'interprétation des résultats<sup>1</sup>.

Par ailleurs, la volonté de **corriger au stade de la programmation les biais liés aux données**, notamment en jouant sur l'apprentissage au moment du RLHF, peut conduire à **créer de nouveaux biais en sens inverse**. Comme il a été vu dans le cas des IA « *woke* », c'est même la réalité historique qui peut se trouver niée, comme l'a montré l'incapacité de Gemini à générer des images attendues de Vikings, de pères fondateurs des États-Unis ou de soldats nazis blancs. Ces phénomènes plaident pour des IA spécifiques à tel ou tel univers linguistique, social et/ou culturel.

---

<sup>1</sup> Pour aller plus loin il est loisible de se reporter au rapport issu du colloque organisé au Sénat le 7 mars 2024 à l'occasion de la Journée internationale des droits des femmes, par la délégation aux droits des femmes, la délégation à la prospective et l'Office : cf. le rapport d'information « Femmes et IA : briser les codes », n° 607 (2023-2024) : <https://www.senat.fr/notice-rapport/2023/r23-607-notice.html>



c) *Les erreurs graves ou « hallucinations »*

En 2023, des journalistes du New York Times ont testé trois *chatbots* d'IA générative différents, à savoir ChatGPT d'OpenAI, Bard de Google et Bing de Microsoft, en leur demandant de rechercher le premier article de leur journal qui faisait allusion à l'intelligence artificielle. Or **chacun de ces trois systèmes a, purement et simplement, inventé des articles qui n'existaient pas**<sup>1</sup>.

Ce problème de **l'invention de références** peut aussi toucher les conditions mêmes de l'exercice de certaines professions dans lesquelles ce type de dérives est encore moins acceptable, comme les sciences ou le droit. ChatGPT a souvent procédé à des « inventions » pures et simples de références académiques, y compris pour des recherches scientifiques d'apparence crédible<sup>2</sup>. En 2023, à l'occasion d'un procès en réparation d'un préjudice corporel, un avocat a choisi de recourir à ChatGPT pour effectuer ses recherches de jurisprudence, ce qui l'a conduit à citer six décisions de justice inventées par le LLM, l'avocat s'est défendu en invoquant la crédibilité apparente des réponses fournies<sup>3</sup>.

Outre la tendance des IA génératives à donner plus dans la vraisemblance que dans la véracité (en inventant des références ou des citations), ces « **hallucinations** » peuvent avoir des effets catastrophiques. Un professeur de droit a ainsi découvert que ChatGPT avait inventé un cas de harcèlement sexuel et fait de lui l'une des personnes accusées<sup>4</sup>.

Ces fausses informations produites par les outils d'IA générative sont **leur façon d'inventer** or ce niveau de créativité peut se régler<sup>5</sup>. Elles s'expliquent notamment par le fait que le modèle « perçoit » des formes (*patterns*) ou des objets qui sont imperceptibles par l'œil humain.

---

<sup>1</sup> Cf. Karen Weise et Cade Metz, « *When AI Chatbots Hallucinate* », New York Times du 1<sup>er</sup> mai 2023 : <https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html>

<sup>2</sup> Les références et les citations des systèmes d'IA « peuvent sembler légitimes et académiques, mais elles ne sont parfois pas réelles ». Cf. Hannah Rozear et Sarah Park, 2023, « *ChatGPT and Fake Citations* », Université de Duke : <https://blogs.library.duke.edu/blog/2023/03/09/chatgpt-and-fake-citations/>

<sup>3</sup> Cf. Benjamin Weiser et Nate Schweber, « *The ChatGPT Lawyer Explains Himself* », New York Times du 8 juin 2023 : <https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html> cf. aussi Molly Bohannon, « *Lawyer Used ChatGPT In Court – And Cited Fake Cases. A Judge Is Considering Sanctions* », article issu du magazine Forbes du 8 juin 2023 : <https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-usedchatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=1441fd7e7c7f>

<sup>4</sup> Cf. Pranshu Verma et Will Oremus, « *ChatGPT invented a sexual harassment scandal and named a real law prof as the accused* », Washington Post du 5 avril 2023 : <https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/>

<sup>5</sup> Cette latitude donnée aux modèles en termes de créativité de leurs prédictions porte le nom de « température ». Les modèles à basse température sont plutôt factuellement fidèles aux informations issues des données d'entraînement tandis que les modèles à haute température introduisent plus d'aléatoire, avec la sélection de tokens statistiquement probablement les moins liés. Ces derniers modèles sont donc plus créatifs et parfois trop, ce qui peut être pertinent pour trouver des idées originales ou écrire de la poésie. Cette créativité peut évidemment être indésirable dans de nombreux autres cas où les outputs insensés ou faux doivent être le plus souvent possible évités.

Les **réglages fins** et les **améliorations actuelles** des IA génératives en termes technologiques les conduisent à **de moins en moins souvent halluciner**. Les versions successives des modèles sur lesquels fonctionnaient ChatGPT depuis deux ans l'illustrent.

Outre ces hallucinations et erreurs factuelles, les IA connexionnistes sont, en dépit de leur puissance, affectées d'une **incapacité à se représenter le monde** ou à **faire preuve de logique**. Un article récent de plusieurs chercheurs, intitulé « Alice au pays des merveilles : des tâches simples montrent une absence totale de raisonnement dans les grands modèles de langage de pointe »<sup>1</sup>, fait ainsi la démonstration qu'il est **facile de piéger ces IA génératives** en raison de leur absence de raisonnement logique.

Les **36 LLM testés**, y compris GPT-4 ou Claude-3.5, **échouent** à répondre efficacement à la question suivante, que l'on peut qualifier de simple : « Alice a 4 sœurs et 1 frère. Combien de sœurs a le frère d'Alice ? ». Dans l'immense majorité des centaines de tentatives de test, les systèmes répondaient que le frère d'Alice a le même nombre de sœurs qu'Alice, donc 4 sœurs. Sur des questions très complexes, les IA génératives fournissent des résultats bluffants mais face à des questions simples, ces technologies souffrent de leur déficit de raisonnement logique.

Ces éléments nous rappellent que **les IA génératives ne raisonnent pas au sens où nous l'entendons** et ne font que **des prédictions statistiques**. Elles ne vérifient pas la véracité de leurs affirmations et leurs réponses peuvent toujours comporter des erreurs, y compris face à des questions simples comme le montre cet exemple.

## 2. Les tendances de la recherche et les principales perspectives technologiques

### a) Moins halluciner : la « Retrieval Augmented Generation » (RAG) par les « Retrieval Augmented Transformers » (RAT)

L'un des problèmes posés par les LLM est donc leur tendance intrinsèque à « halluciner », c'est-à-dire à **générer des propos dénués de sens ou des réponses objectivement fausses sans émettre le moindre doute**. Or, ces hallucinations peuvent se révéler dangereuses pour certaines utilisations, comme en témoignent **les usages médicaux ou juridiques**. Ces hallucinations sont dues au fait que les LLM n'ont pas accès à une base de données de connaissance déterminées comme les modèles symboliques, mais qu'ils s'appuient sur une **construction statistique destinée à prédire une suite de mots probable ou plausible**.

---

<sup>1</sup> Cf. Marianna Nezhurina et al., version du 13 juillet 2024, « Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models »: <https://arxiv.org/abs/2406.02061>

En plus de cela, les données disponibles pour un LLM donné sont arrêtées dans le temps au moment de leur entraînement, elles ne peuvent pas être facilement actualisées une fois leur entraînement terminé, à moins de relancer un nouvel entraînement et de produire une nouvelle version du modèle, ce qui s'avère complexe et coûteux pour de grands modèles déjà diffusés sur le marché.

La **génération augmentée de récupération** (en anglais *Retrieval Augmented Generation*, RAG) constitue un moyen d'**adjoindre une base de données** à un LLM. Un récupérateur utilise les données de cette base en lien avec l'instruction de l'utilisateur pour modifier ou compléter la réponse fournie directement par le modèle d'IA<sup>1</sup>. La base de données est généralement stockée sous la forme de vecteurs : on parle de base de données vectorielle permettant de stocker des données de natures différentes, structurées ou non, et de réaliser des opérations mathématiques en son sein en fonction de la similarité de deux sujets, comme cela peut se faire avec les *tokens* vectorisés après un plongement lexical. Grâce à ce récupérateur, de nouvelles données vont permettre de modifier ou compléter la réponse fournie par le modèle.

L'enjeu est de réussir à **créer un récupérateur performant**. En effet, s'il s'avère trop faible, le récupérateur sera incapable de trouver les informations les plus pertinentes en rapport avec l'instruction de l'utilisateur et générera du contenu hors sujet. Au contraire, un récupérateur trop fort aura du mal à récupérer les informations pertinentes, considérant que l'information demandée par l'utilisateur ne se trouve pas dans la base de données de récupération, rendant impossible une réponse efficace à la requête.

Il existe plusieurs types de RAG, plus ou moins efficaces, dotés d'un ou de plusieurs récupérateurs, et dont les cas d'usage sont donc différents. Des RAG peuvent par exemple recourir à des *Transformers* dédiés à la génération d'image. Ces *Retrieval Augmented Layout Transformers* (RAT ou RALT) pour contenus visuels ont fait l'objet des premiers travaux théoriques sur les RAG<sup>2</sup>.

Les **combinaisons entre les raisonnements logiques propres à l'IA symbolique** et les **généralisations statistiques par induction que sont les IA connexionnistes** sont une autre réponse possible aux hallucinations et, plus généralement, à l'ensemble des erreurs des IA génératives actuelles. Ces hybridations et articulations seront vues plus loin.

---

<sup>1</sup> Cf. la synthèse de Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, et al., 2024, « Retrieval-Augmented Generation for Large Language Models: A Survey » : <http://arxiv.org/abs/2312.10997>

<sup>2</sup> Cf. l'article de Daichi Horita, Naoto Inoue, Kotaro Kikuchi, Kota Yamaguchi et Kiyoharu Aizawa, 2024, « Retrieval-Augmented Layout Transformer for Content-Aware Layout Generation » : <https://arxiv.org/abs/2311.13602>

*b) Manipuler en entrée et en sortie des données de nature variée : les IA génératives multimodales*

Lorsque ChatGPT a été dévoilé au grand public en novembre 2022, l'agent conversationnel était limité à la discussion textuelle : il ne prenait **que du texte en entrée et ne produisait que du texte en sortie**.

À la sortie de GPT-4 en mars 2023, OpenAI a permis aux utilisateurs d'entrer du texte, des documents complets, mais également des images, tous susceptibles d'être traités par le modèle.

Enfin, GPT-4o (avec un « o » pour omni) a ajouté à ces formats de données en entrée, le traitement d'**instructions vocales** et même des **interactions vidéo** avec une capacité à détecter les émotions à travers la reconnaissance des expressions du visage<sup>1</sup>. L'ajout d'un nouveau type de données d'entrée renforce les compétences du modèle : avec la possibilité de traiter des images, le modèle a des capacités de reconnaissance optique de caractères, ou de reconnaissance d'images plus généralement. Avec l'ajout du traitement vocal, le modèle peut analyser un rythme ou un ton de voix.

Après avoir ajouté la **génération** d'images à son IA Grok 2 en septembre 2024, xAI a ajouté la **compréhension d'images** à son système en octobre 2024, à travers une annonce sur X<sup>2</sup> : les utilisateurs peuvent désormais télécharger des images et interagir directement avec l'IA à travers des textes et des contenus visuels en entrée. Elon Musk a indiqué que les capacités de Grok devraient s'améliorer rapidement avec la **compréhension de documents comme des fichiers PDF ou le contexte d'une image, y compris l'humour ou le second degré**.

Le futur devrait donner une **place grandissante à ces IA qualifiées de « multimodales »**. Ces modèles sont non seulement capables de traiter des ensembles de données provenant de sources de nature diverse mais de produire des résultats eux-mêmes de nature variée. Les IA sont multimodales du côté de leurs *inputs* comme du côté de leurs *outputs*. La multimodalité est **un défi technologique** en passe d'être totalement relevé et une caractéristique de plus en plus recherchée par les entreprises mettant sur le marché des modèles de fondation et des applications. Cela permet d'ajouter des compétences aux modèles, de permettre aux utilisateurs de n'avoir recours qu'à un seul modèle pour une multitude de tâches, de personnaliser leurs interactions avec les interfaces des systèmes et, en pratique, de leur offrir **plus de fluidité et de souplesse dans l'utilisation des systèmes d'IA**.

---

<sup>1</sup> Cf. Benj Edwards, 2024, « Major ChatGPT-4o update allows audio-video talks with an "emotional" AI chatbot », Ars Technica: <https://arstechnica.com/information-technology/2024/05/chatgpt-4o-lets-you-have-real-time-audio-video-conversations-with-emotional-chatbot/>

<sup>2</sup> Cf. le tweet de Grok sur X : <https://x.com/grok/status/1850808322074509434>

c) *Faire des systèmes d'IA des interfaces devenant la principale plateforme d'accès aux services numériques*

L'une des perspectives des systèmes d'IA est de devenir des **interfaces** incontournables et d'être les principales **plateformes** d'accès aux services numériques. Ces interfaces fondées sur l'IA rendraient les interactions homme-machine plus fluides et pourraient devenir la principale forme de nos ordinateurs personnels, agrégeant les fonctionnalités de l'interface du système d'exploitation, des navigateurs *web*, des moteurs de recherche, des logiciels bureautiques, des réseaux sociaux et d'autres applications. Ces nouvelles interfaces basées sur l'intelligence artificielle remplaceraient également les interfaces des *smartphones* telles que nous les connaissons, permettant une prise en charge par l'IA de toutes les applications contenues dans l'appareil. Par exemple, Microsoft et Google y travaillent, soit directement pour leur propre compte, soit avec leurs partenaires : OpenAI pour le premier, Anthropic pour le second. **Perplexity** a développé une IA-moteur de recherche fournissant des réponses précises à partir de *prompts*.

Pour le moment, **Copilot** qui a recours à ChatGPT enrichit les logiciels de Microsoft, ces derniers restant l'environnement de référence en bureautique mais il est probable qu'à l'avenir **la logique s'inverse** et que les systèmes d'IA deviennent les **colonnes vertébrales de contrôle des ordinateurs**, à partir desquelles s'articuleront plusieurs services logiciels. Un nouveau navigateur *web* basé sur l'IA incluant un moteur de recherche et des fonctionnalités en termes de e-commerce formerait ainsi l'un des premiers jalons de cette évolution. Sous le nom de code « *Jarvis Project* », un tel projet est en cours de développement chez Google, qui devrait faire une démonstration de ce produit en décembre 2024, à l'occasion de la sortie de la nouvelle génération de son modèle Gemini<sup>1</sup>.

La navigation sur Internet se ferait par exemple à partir d'un système d'IA prenant la forme d'un CUA (pour *Computer Using Agent*, soit agent utilisant l'ordinateur). Le CUA pourra décider d'actions en fonction des résultats obtenus, son fonctionnement reposant sur l'analyse de captures d'écran, ce qui lui permettra d'interagir directement sur les pages *web* visitées en cliquant sur des boutons ou en saisissant du texte.

Le 22 octobre 2024, **Anthropic** a annoncé un **système baptisé Computer Use permettant de contrôler un ordinateur personnel par IA sur la base de son modèle Claude 3.5 Sonnet**. Ce système est d'ores et déjà disponible de façon expérimentale en API sur la plateforme d'Anthropic<sup>2</sup>.

---

<sup>1</sup> L'information, largement diffusée depuis, a été rendue publique par le site The Information le 26 octobre 2024, avec un titre explicite et plutôt inquiétant : « Google preps Ai that takes over computers ». Cf. <https://www.theinformation.com/articles/google-preps-ai-that-takes-over-computers>

<sup>2</sup> Cf. l'annonce du 22 octobre 2024 : <https://www.anthropic.com/news/3-5-models-and-computer-use> ainsi que l'API : <https://docs.anthropic.com/en/docs/build-with-claude/computer-use> et une première démonstration : <https://youtu.be/ODaHJzOyVCQ>



Ces IA interfaces auront donc pour propriété une plus grande « agentivité », cette autonomie croissante étant également, en soi, une autre des grandes perspectives technologiques des futures IA.

*d) Aller vers plus d'autonomie : le défi de l'agentivité*

L'agentivité désigne la capacité des intelligences artificielles à réaliser des **actions plus ou moins autonomes** pour remplir des objectifs. Une IA capable de réaliser de telles tâches en autonomie est appelée « agent ». Cette agentivité, plutôt qu'une caractéristique binaire que les modèles auraient ou n'auraient pas, désigne plutôt un degré supérieur d'autonomie dont seraient dotées les applications et qui permettrait de les qualifier d'agents. Le psychologue Daniel Kahneman distingue deux modes de pensée que le cerveau utilise, l'un pour traiter les informations, l'autre pour prendre des décisions, le système 1 et le système 2 : l'enjeu de l'agentivité est de faire passer l'IA de la première phase à la deuxième. **Les IA, de LLM, devraient devenir des LAM ou Large Action Models.**

Yann LeCun, responsable de l'IA chez Meta, rencontré par vos rapporteurs dans les locaux de l'Université de New York où il est professeur, considère l'agentivité comme **l'un des principaux vecteurs de développement de l'intelligence artificielle**. Il croit, en effet, que la scalabilité (capacité à monter à l'échelle) des modèles d'IA basés sur la technologie *Transformer* rencontrera une barrière qui ne pourra être surmontée qu'en développant de nouvelles technologies qui permettront aux systèmes d'être davantage agentiques. Il parle à cet égard de modèles d'intelligence artificielle guidés par leurs objectifs, appelés ODAI, de l'anglais « *Objective-Driven AI* ». Les déclarations d'OpenAI en novembre 2024 vont aussi dans ce sens : les IA permettant l'exécution de tâches autonomes seront probablement la **prochaine percée en intelligence artificielle**<sup>1</sup>.

Les évolutions sont rapides en la matière. La principale innovation en 2024 est celle des **Agentic Workflows**, IA basées sur des LLM et dont le caractère adaptatif permet une automatisation des tâches en s'adaptant en temps réel à la complexité des flux de travail. Ces **outils devraient être particulièrement utiles pour les entreprises**.

Plutôt que de générer une simple réponse, ces IA assureront **une série d'actions** selon un processus itératif<sup>2</sup>. Andrew Ng, figure du *Deep Learning*, estime que cette nouvelle architecture « *entraînera des progrès massifs en IA* »<sup>3</sup>.

---

<sup>1</sup> Cf. la session AMA (pour « Ask me anything ») organisée par OpenAI le 1<sup>er</sup> novembre 2024 avec Sam Altman, Kevin Weil, Srinivas Narayanan et Mark Chen sur le réseau social Reddit.com : [https://www.reddit.com/r/ChatGPT/comments/1ggixzy/ama\\_with\\_openais\\_sam\\_altman\\_kevin\\_weil\\_srinivas/](https://www.reddit.com/r/ChatGPT/comments/1ggixzy/ama_with_openais_sam_altman_kevin_weil_srinivas/)

<sup>2</sup> Si l'on prend l'exemple d'une entreprise qui cherche à savoir si Madame Dupont a réglé sa facture annuelle ou pas, un LLM classique répondra oui ou non en apportant éventuellement des précisions après analyse des données de l'entreprise, alors que les Agentic Workflows déclencheront une série d'actions (modifications des fichiers internes, changement d'état, déclenchement de l'envoi de courriels, prise de rendez-vous, etc.).

<sup>3</sup> Cf. cette conférence d'Andrew Ng, « On AI Agentic Workflows And Their Potential For Driving AI Progress » : <https://www.youtube.com/watch?v=q1XFm21I-VQ>

**IBM** a diffusé, en octobre 2024, un cadre en *open source* pour développer de telles IA, appelé Bee Agent Framework<sup>1</sup>. Plusieurs entreprises offrent fin 2024 des services basés sur ces technologies. **Salesforce** propose ainsi depuis septembre 2024 des systèmes d'IA avec des LLM agentiques comme xGen-Sales, xLAM ou Agentforce. **OpenAI** a annoncé son modèle GPT-4o1, avec la perspective de services agentiques. **ServiceNow** commercialise Xanadu et **UiPath** développe l'automatisation agentique, c'est-à-dire des *Agentic Workflows* couplés à l'automatisation de processus robotiques (RPA). En Allemagne, **Celonis** a créé le système AgentC.

En dehors de ces *Agentic Workflows*, d'autres innovations visent des systèmes d'IA plus agentiques. **Rabbit R1** de l'entreprise **Teenage Engineering** est un petit boîtier ressemblant à une sorte de téléphone portable contenant une IA à agentivité forte, présenté comme le « *nouveau compagnon de poche* » lors du CES de Las Vegas de 2024. Doté du système d'exploitation « Rabbit os », cet appareil permet de réaliser, sur la base des grands LLM, un certain nombre de tâches en utilisant une interface minimaliste grâce à un modèle d'IA qui répond aux requêtes de l'utilisateur de façon autonome et personnalisée. Il s'agit donc d'un modèle d'action couplé à des LLM qui permet de simplifier et d'automatiser la navigation entre différentes applications. Ces *rabbits*, agents d'IA personnels, peuvent gérer une variété de tâches complexes, de la recherche d'informations à des actions en conséquence comme la mise à jour d'un agenda ou la réservation de voyages, offrant ainsi une **expérience plus intuitive à l'utilisateur**. Par exemple, un utilisateur qui demanderait au modèle « Joue-moi une musique qui correspond à l'ambiance actuelle et montre-moi les tweets qui m'intéressent le plus sur X » devrait susciter une **réponse construite à partir d'une analyse de l'environnement et des habitudes de l'utilisateur**, pour déterminer au mieux « l'ambiance » dans laquelle il se trouve ou qu'il perçoit. Le système d'IA pourrait alors chercher sur une autre application une musique adaptée à l'ambiance selon des critères de classification qui seraient, eux aussi, déterminés par une intelligence artificielle. Enfin, sur une autre application, X en l'occurrence, le système irait chercher les tweets qui intéressent le plus l'utilisateur en fonction de ce qu'il connaît de l'utilisateur et des recommandations de X. Si ce produit n'est pas encore très efficace<sup>2</sup> et présente des failles de sécurité importantes<sup>3</sup>, il constitue néanmoins un exemple d'évolution possible de la technologie vers des systèmes plus agentiques.

---

<sup>1</sup> Cf. cet article du 25 octobre 2024, « IBM Developers Release Bee Agent Framework: An Open-Source AI Framework for Building, Deploying, and Serving Powerful Agentic Workflows at Scale » : <https://www.marktechpost.com/2024/10/25/ibm-developers-release-bee-agent-framework-an-open-source-ai-framework-for-building-deploying-and-serving-powerful-agentic-workflows-at-scale/>

<sup>2</sup> Cf. Nassim Chentouf, 2024, « Rabbit r1 : la supercherie d'une simple application Android » : <https://www.lesnumeriques.com/intelligence-artificielle/rabbit-r1-la-supercherie-d-une-simple-application-android-n221617.html> et Florent Lanne, 2024, « On a essayé le Rabbit R1, un gadget plein d'IA qui peine encore à trouver une utilité » : <https://www.lesnumeriques.com/intelligence-artificielle/on-a-essaye-le-rabbit-r1-un-gadget-plein-d-ia-qui-peine-encore-a-trouver-une-utilite-n224784.html>

<sup>3</sup> Cf. Florian Bayard, 2024, « Rabbit R1 : l'IA de poche est une catastrophe pour votre cybersécurité » : <https://www.01net.com/actualites/rabbit-r1-ia-poche-catastrophe-cybersecurite.html>

En dehors de ce cas spécifique de Rabbit R1, il est certain que les **Large Action Models** et leurs applications en tant qu'**Agentic Workflows** sont des technologies qui vont marquer la prochaine étape de l'histoire de l'IA sur le chemin de modèles d'intelligence artificielle entièrement guidés par leurs objectifs.

e) Faire plus avec moins : vers une IA frugale et efficace

Au cours des prochaines années, les **architectures** d'IA devraient être **de plus en plus efficaces avec de moins en moins de puissance de calcul** mobilisée, notamment pour contenir la consommation d'énergie. Pour cela, il faut agir dans plusieurs directions : moins consommer pendant la phase d'apprentissage des modèles et moins consommer pendant l'exploitation des modèles par les applications, même si les deux peuvent se rejoindre.

Certaines améliorations porteront sur la **structure des processeurs** eux-mêmes. Différentes entreprises se sont donné pour but de concevoir de nouvelles générations de puces dédiées aux spécificités des calculs de *Deep Learning*, notamment pour l'IA générative. Ces puces devront être plus performantes pour ces milliers de milliards de calculs rapides que ne le sont aujourd'hui les GPU de Nvidia, tout en consommant moins d'énergie. En 2017, Google avait créé des *Tensor Processing Units* (TPU). Plus récemment, y compris en 2024, de nouveaux produits ont été créés par Cerebras, SambaNova, Rivos, Tenstorrent ou encore Groq. Une publication récente a été consacrée à cette question<sup>1</sup>. Le projet de faire progresser rapidement l'IA par **l'informatique quantique** est controversé car cette dernière rencontre des difficultés à traiter les calculs des réseaux de neurones et leurs données massives<sup>2</sup>.

Une autre piste à explorer consiste à **décentraliser les calculs** sur l'ensemble d'un réseau. Bien que très consommatrices d'énergie, les *blockchains*, qui reposent sur un fonctionnement largement décentralisé<sup>3</sup>, sont un terrain favorable à des expérimentations. Elles assurent le stockage et la transmission d'informations, par la constitution de registres répliqués et

---

<sup>1</sup> Cf. Gareth Owen, 2024, « AI chip start-ups – Can domain-specific chips impact Nvidia's dominance ? » : <https://www.counterpointresearch.com/insights/ai-chip-start-ups-can-domain-specific-chips-impact-nvidias-dominance/>

<sup>2</sup> Les ordinateurs quantiques sont trop lents et seuls des calculs très courts y sont effectués sans pannes, comme l'explique cet article d'un professeur à l'école polytechnique : Filippo Vicentini, octobre 2024, « Informatique quantique et IA : moins compatibles que prévu ? » : <https://www.polytechnique-insights.com/tribunes/science/informatique-quantique-et-ia-moins-compatibles-que-prevu/>

<sup>3</sup> L'OPECST a consacré un rapport complet à ces technologies complexes en 2018. Comme l'explique ce rapport, les perspectives ouvertes sont considérables, les applications des *blockchains* dépassant même le cadre monétaire, mais peu conjuguent maturité technologique suffisante et pertinence de l'usage. Le recours aux *blockchains* relève encore souvent d'un enjeu de marketing plus que d'une réponse technologique idoine à des besoins avérés. Cf. le rapport n° 584, 2017-2018, « Comprendre les *blockchains* : fonctionnement et enjeux de ces nouvelles technologies » : <https://www.senat.fr/notice-rapport/2017/r17-584-notice.html>



distribués, sans organe central de contrôle, sécurisés grâce à la cryptographie et structurés par des blocs liés les uns aux autres, à intervalles de temps réguliers. Il faut d'ailleurs noter que **deux cryptomonnaies**, Qubic et Bittensor de Taon, **associent des modèles d'IA à des calculs au sein des processeurs de la communauté des utilisateurs des *blockchains* concernées.**

La décentralisation des calculs peut également rejoindre une méthode d'optimisation employée dans le *cloud computing* qui consiste à traiter les données à la périphérie du réseau, à proximité de la source des données, et que l'on appelle *edge computing*. Cette technologie distribuée permet – en mobilisant des ressources informatiques diverses telles que des ordinateurs portables, des *smartphones*, des tablettes ou des capteurs, même sans qu'ils soient connectés en permanence au réseau – de minimiser les besoins en bande passante ainsi que la consommation d'énergie des *data centers*.

À côté des processeurs et de la décentralisation des calculs, les améliorations porteront aussi sur les **architectures des modèles**. La méthode abordée au paragraphe suivant, outre son intérêt en termes de performances, est l'une des technologies par laquelle **une puissance de calcul moindre peut aboutir à de meilleurs résultats.**

f) *L'exemple de la méthode « Mixture of Experts » (MoE)*

Bien que distincte des systèmes d'IA dits multi-agents, la **méthode du mélange d'experts** (le terme anglais *Mixture of Experts* ou MoE est presque systématiquement préféré à sa traduction française) peut y être rattachée tout en étant plus spécifique car s'appuyant sur la combinaison de plusieurs modèles en parallèle. Cette technologie permet d'être **plus efficace avec de moindres besoins en termes de puissance de calcul**. Elle est déjà utilisée avec succès par certains modèles, comme Mixtral8x7b de MistralAI<sup>1</sup>. Dans ces modèles, plutôt que de n'avoir qu'un seul grand modèle, on utilise une combinaison de plusieurs modèles, appelés « experts ». Le modèle Mixtral par exemple contient, comme son nom l'indique, un *mix* de modèles, soit huit experts par couche, et le modèle global basé sur ce MOE est configuré comme un LLM à sept milliards de paramètres.

**L'OPECST avait, dès 2017, identifié le potentiel des MOE pour les systèmes d'IA et encouragé leur utilisation**<sup>2</sup>. Le rapport de l'Office se basait sur un article publié en janvier 2017 par des chercheurs de Google (dont certains d'entre eux théoriseront l'architecture *Transformer* neuf mois plus tard) qui imaginait un réseau neuronal géant, composé de plusieurs sous-réseaux neuronaux disposés en couches, améliorant de manière considérable sa performance et sa capacité de calcul, avec une architecture comprenant jusqu'à 137 milliards de paramètres.

---

<sup>1</sup> Cf. Albert Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot et al., 2024, « Mixtral of Experts » : <http://arxiv.org/abs/2401.04088>

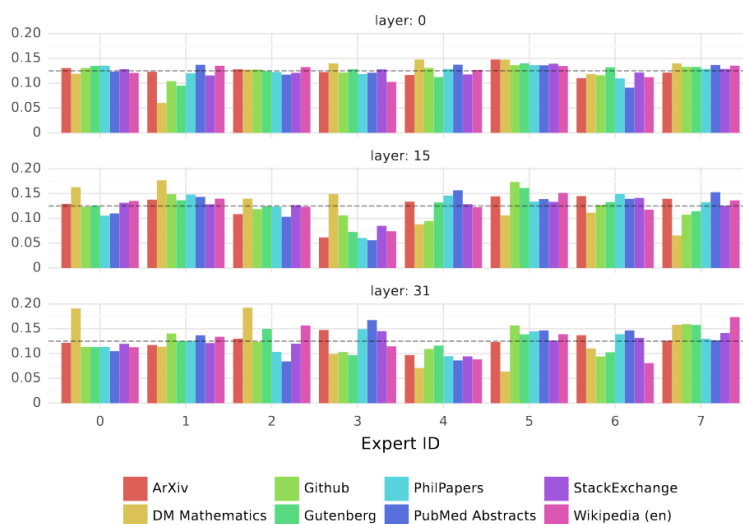
<sup>2</sup> Cf. le rapport de l'OPECST « Pour une intelligence artificielle maîtrisée, utile et démystifiée » précité, p.64.

Le rapport de l'Office affirmait aussi que « *si, à ce stade, le modèle « MoE » peut être appliqué aux tâches de modélisation des langues et de traduction automatique, l'avancée présentée permet d'entrevoir des progrès exponentiels en matière d'intelligence artificielle* » voire de se rapprocher du « *possible avènement d'une intelligence artificielle générale composée de milliers de sous-réseaux et traitant toutes sortes de données. Il s'agit aussi de réduire le nombre de processeurs (GPU) nécessaires à l'apprentissage et donc d'accélérer la capacité du système d'intelligence artificielle à nombre de processeurs égal* »<sup>1</sup>. L'OPECST avait en quelque sorte prophétisé les IA génératives multimodales introduites depuis un an ainsi que leurs lois d'échelle (*scaling laws*) leur permettant d'être de plus en plus efficaces de plus en plus rapidement<sup>2</sup>.

Dans les MoE appliqués aux grands modèles de langage, les experts se situent dans les couches denses, c'est-à-dire les couches de neurones artificiels à propagation avant entièrement connectés. Dans ces réseaux denses, un routeur oriente les données vers un ou plusieurs experts (le nombre d'experts choisi étant un hyperparamètre) en fonction de la nature de la donnée. Les sorties des experts choisis sont ensuite agrégées pour obtenir une sortie unique.

Le modèle MoE Mixtral 8x7b de l'entreprise française MistralAI est ainsi disséqué pour ses couches 0, 15 et 31 dans le schéma de la page suivante. Il permet de voir que les *tokens* assignés par le routeur à chaque expert sont différents selon leur source.

### Proportion de *tokens* assignés par le routeur à chaque expert en fonction de leur source pour les couches 0, 15 et 31 du modèle Mixtral 8x7b



N.B. : La ligne pointillée grise représente un huitième, soit la proportion attendue avec une distribution uniforme  
*Source* : Jiang et al. « *Mixtral of Experts* », op. cit.

<sup>1</sup> Cf. le rapport précité, p. 65.

<sup>2</sup> Ces lois signifient qu'accroître la puissance de calculs des superordinateurs et le volume de données traités conduit mécaniquement les modèles à être de plus en plus performants d'un point de vue qualitatif. Comme l'a rappelé Siméon Campos lors de son audition, l'IA va vite et ça ne va certainement pas s'arrêter. Cf. un article qui illustre ces *scaling laws*, Jordan Hoffmann, 2022, « *Training Compute-Optimal Large Language Models* » : <https://arxiv.org/abs/2203.15556>

Ces modèles permettent d'**augmenter l'efficacité des modèles sans augmenter la complexité des calculs nécessaires pour les faire fonctionner**. À ce titre, lors de leur audition devant vos rapporteurs, des scientifiques travaillant pour MistralAI ont affirmé qu'il s'agissait bel et bien d'une piste pour le développement de l'**IA frugale**, efficace tout en utilisant le moins de puissance de calcul possible.

L'architecture **Mamba** alternative aux *Transformers*, reposant sur les modèles des espaces d'états structurés ou SSM, est déjà en elle-même, de par sa conception, moins consommatrice de puissance de calcul que les *Transformers*. Elle peut aussi s'articuler sur une intégration de plusieurs modèles dans une MOE, on parle alors de *Mamba Mixture of Experts* (MMOE). Ces types de MOE, encore peu développés car inventés en 2024, sont des **pistes intéressantes pour une IA frugale**<sup>1</sup>.

Le modèle **Jamba** qui conjugue l'architecture *Transformer* et l'architecture Mamba pourrait à l'avenir être lui aussi démultiplié sous la forme d'une grande architecture globale combinant des MOE Mamba et des MOE *Transformer*.

### 3. Synthèse et articulations entre les modèles d'IA

#### a) Des technologies enchâssées et souvent conjuguées

Le graphique suivant, sans être totalement satisfaisant car il oublie l'IA symbolique, offre une **vision de synthèse** des catégories des différents systèmes d'intelligence artificielle actuels, en en soulignant l'**enchâssement**.

Les domaines ne sont pas les uns à côté des autres mais imbriqués les uns dans les autres. De plus, chaque système peut **hybrider des briques technologiques** issues de plusieurs catégories, ce qui est assez fréquent en réalité. Les systèmes experts peuvent par exemple être utilisés avec des raisonnements par analogie, éventuellement dans le cadre de systèmes multi-agents.

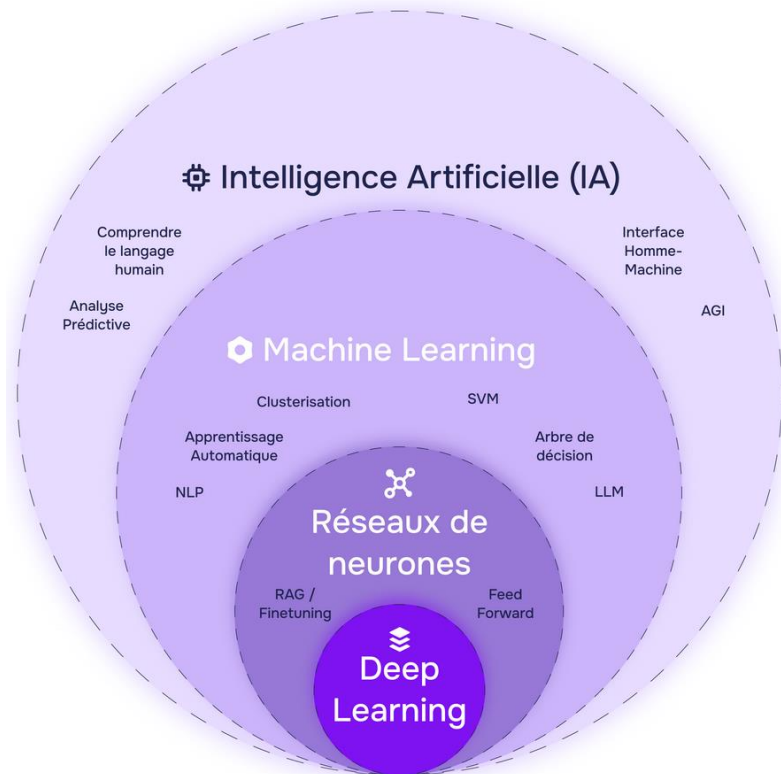
Les chercheurs, tels des artisans, **hybrident souvent des solutions inédites**, au cas par cas, en fonction de compétences, voire d'un tour de main, qui peuvent être assez personnels. Il s'agit d'une caractéristique propre à la recherche en intelligence artificielle, souvent peu connue à l'extérieur du cercle des spécialistes.

---

<sup>1</sup> Cf. la note de Nikhil, 2024, « This AI Paper Proposes MoE-Mamba: Revolutionizing Machine Learning with Advanced State Space Models and Mixture of Experts MoEs Outperforming both Mamba and Transformer-MoE Individually », MarkTechPost :

<https://www.marktechpost.com/2024/01/13/this-ai-paper-proposes-moe-mamba-revolutionizing-machine-learning-with-advanced-state-space-models-and-mixture-of-experts-moes-outperforming-both-mamba-and-transformer-moe-individually/> ainsi que l'article scientifique de Maciej Pióro et al., 2024, « MoE-Mamba: Efficient Selective State Space Models with Mixture of Experts » : <https://arxiv.org/abs/2401.04081>

## L'imbrication des grands domaines actuels de l'IA



Précision : les exemples du graphique sont placés aléatoirement et pourraient fréquemment figurer dans des cercles de plus petite dimension, à l'image des LLM ou de l'analyse prédictive qui relèvent souvent du *Deep Learning*

Source : [www.data-bird.co](http://www.data-bird.co)

b) Les Arbres de pensées ou Trees of Thought (ToT) : l'IA « symboliconnexionniste »

L'idée des arbres de pensées ou *Trees of Thought* (ToT) a été développée en 2023 dans un article scientifique de chercheurs de Deepmind, filiale de Google dédiée à l'intelligence artificielle<sup>1</sup>. Cette méthode permet d'améliorer les résultats générés par un modèle de langage.

Sans relever directement de l'IA symbolique, cette technique s'en rapproche par son recours à des étapes formelles de raisonnement, les idées venant s'articuler logiquement les unes par rapport aux autres. L'IA devient alors neuro-symbolique car empruntant à la fois des réseaux neuronaux et des raisonnements symboliques, ce qui pourrait conduire à la qualifier d'IA « symboliconnexionniste ».

Pour cela, on fait générer à un modèle des « idées », étapes intermédiaires de raisonnement permettant de répondre à la fin à la requête de l'utilisateur. Chaque idée génère à son tour d'autres idées, qui en génèrent plusieurs autres, formant ainsi un « arbre », invisible pour l'utilisateur.

<sup>1</sup> Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, et Karthik Narasimhan, 2023, « Tree of Thoughts: Deliberate Problem Solving with Large Language Models » : <http://arxiv.org/abs/2305.10601>.

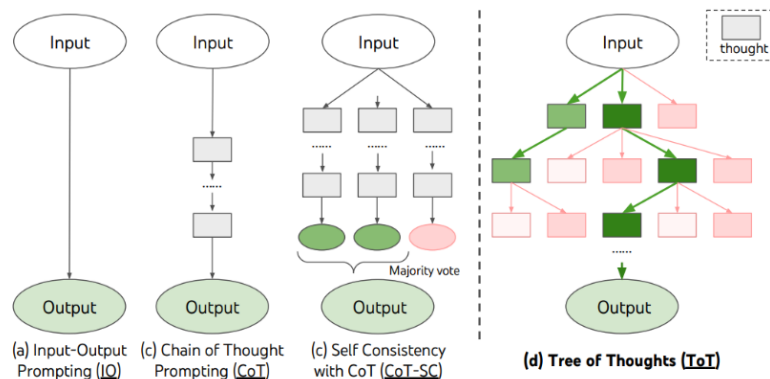
Ces idées sont ensuite **évaluées par le modèle** qui détermine la séquence d'idées la plus pertinente pour répondre à la requête initiale du modèle.

La méthode de l'arbre de pensées peut être **appliquée directement à un grand modèle de langage grâce à une instruction**<sup>1</sup>. On parle, pour ces invites de commande, de techniques de *prompt engineering* (ingénierie des instructions). Dave Hulbert donne l'exemple d'instruction suivant :

« *Imagine que trois experts différents répondent à cette question. Tous les experts vont écrire une étape de leur raisonnement, puis le partager au groupe. Ensuite, tous les experts vont procéder à l'étape suivante, etc. Si un expert réalise à un moment qu'il a tort, alors il part. La question est ...* »

Ces techniques donnent de **meilleurs résultats** puisqu'elles permettent d'obtenir **davantage de réponses factuellement vraies** aux questions qui sont posées aux grands modèles de langage.

**Schéma illustrant les différents modes de raisonnements possibles pour un LLM, l'arbre de pensées est à droite**



Source : Yao et al., 2023, op. cit.

À côté des arbres de pensées, il existe des techniques proches, plus simples mais souvent aussi moins efficaces comme la « **chaîne de pensée** » qui consiste à générer des idées les unes à la suite des autres pour élaborer une sortie plus pertinente et plus explicite par rapport à la requête initiale. D'autres techniques existent également comme les « **instructions à plusieurs essais** » (*few-shot prompting*) consistant à inclure dans une instruction des exemples de réponses attendues pour obtenir une réponse encore plus pertinente.

Les chaînes et les arbres de pensées semblent montrer que la technologie peut évoluer non seulement grâce à des **innovations dans l'architecture** des modèles eux-mêmes mais également grâce à **des innovations dans la manière de réaliser les requêtes qui permettent d'obtenir des réponses pertinentes**, c'est l'art du *prompting*.

<sup>1</sup> Hulbert, Dave, 2023, Using Tree-of-Thought Prompting to boost ChatGPT's reasoning, cf. <https://github.com/dave1010/tree-of-thought-prompting>

Lors de son audition devant la commission des affaires économiques du Sénat le mercredi 22 mai 2024, Arthur Mensch, co-fondateur de MistralAI a ainsi pu affirmer que la façon de donner des instructions à un LLM pouvait être considérée comme **un langage de programmation s'appuyant sur le langage naturel**<sup>1</sup>.

Ces pistes d'amélioration des systèmes d'intelligence artificielle peuvent aussi s'apparenter à une **nouvelle hybridation entre systèmes d'intelligence artificielle symboliques et systèmes d'intelligence artificielle connexionnistes**. Grégory Renard, pionnier du Deep Learning, a lors de son audition, souligné les vertus de telles articulations entre les deux grandes branches de l'IA.

*c) La fécondité des hybridations IA symboliques/IA connexionnistes, notamment pour doter ces systèmes d'une représentation du monde réel*

Le rapport de l'OPECST de 2017 soulignait déjà l'intérêt de **combiner et d'hybrider les technologies, notamment les deux branches de l'IA**, en vue de rendre les systèmes d'intelligence artificielle de plus en plus puissants et efficaces.

C'est une stratégie dans laquelle a toujours cru l'entreprise **IBM**, comme en a témoigné l'exemple de Watson et de ses déclinaisons dans de nombreux secteurs d'activité.

De même, dans les années 2010, **Google** a enrichi son moteur de recherche d'une IA symbolique appelée « **base de connaissances** » (ou *Knowledge Graph* à ne pas confondre avec une base de données) qui permet aux utilisateurs de bénéficier d'une information de synthèse courte, structurée et détaillée, centrée sur l'objet même de la recherche effectuée. Cette sorte d'encadré qui apparaît en marge des résultats de son moteur de recherche repose sur le croisement de la compilation des données issues du moteur de recherche avec un réseau sémantique contenant plus de 500 millions d'objets et plus de 18 milliards de faits et de relations entre ces différents objets utilisés par le moteur de recherche afin de comprendre la signification des mots clés saisis lors d'une recherche. Ce **réseau sémantique**, qui est un graphe représentant la nature des relations sémantiques entre des concepts, permet une représentation des connaissances particulièrement fiable qui n'est pas statistique.

Dans ces mêmes années 2010, les **réseaux de neurones graphiques** ou *Graph Neural Networks* (GNN), proches des réseaux de neurones à convolution (CNN), se sont développés avec une structure reposant sur l'échange d'informations entre les nœuds d'un graphe et leurs voisins (on parle de *message passing*). Bien que relevant de l'IA connexionniste, leurs architectures sous forme de graphes les apparentent à l'IA symbolique. Leurs applications pourraient se multiplier en termes de classification, de recommandation, de prédiction, de sécurité ou encore de vision, voire de génération de contenus.

---

<sup>1</sup> Cf. le compte rendu de la réunion disponible sur le site du Sénat : <https://www.senat.fr/compte-rendu-commissions/20240520/affeco.html#toc5>



**AlphaFold**, outil d'analyse du repliement des protéines dont il sera question plus tard, **repose sur cette technologie des GNN**.

Pour de nombreux chercheurs relevant de l'école dite d'IA **neuro-symbolique**, comme Leslie Valiant, Gary Marcus, Daniel Kahneman, Artur d'Avila Garcez, Marco Gori, Francesca Rossi, Bart Selman, Henry Kautz, Luis Lamb, Pascal Hitzler, Krysia Broda ou Dov Gabbay, il est nécessaire de combiner un raisonnement logique explicite propre aux IA symboliques et un raisonnement statistique propre aux IA connexionnistes pour obtenir un bon modèle cognitif informatique, à la fois précis et riche.

Selon Gary Marcus, « nous ne pouvons pas construire de modèles cognitifs riches de manière adéquate et automatisée sans le triumvirat d'une architecture hybride, de riches connaissances préalables et de techniques de raisonnement sophistiquées », ce qui implique de disposer de la capacité de manipulation des symboles dans notre boîte à outils si l'on souhaite construire des IA solides. Il faut pouvoir disposer de technologies capables de représenter et de manipuler les abstractions.

L'incapacité des IA connexionnistes à **se représenter le monde** ou à **faire preuve de logique** reste un défi pour la recherche. Yann LeCun pense qu'on devra **articuler ces IA statistiques avec des modèles de représentation du monde dans de nouvelles « architectures cognitives »** dont Pat Langley avait en 2017 dressé le bilan des progrès au cours des 40 dernières années, soulignant que l'IA restait surtout analytique et insuffisamment synthétique<sup>1</sup>. Il lui manque une théorie unifiée de la cognition pour pouvoir se rapprocher de l'intelligence humaine.

Joshua Bengio affirme aussi que **les IA génératives auront à se combiner avec un modèle de représentation du monde réel** (« *World Model* ») pour réellement progresser. Ces perspectives articuleront la recherche en IA avec les sciences cognitives. Elles ouvriront la voie à des IA qui **comprennent le monde physique et ses lois**.

Recourir à des combinaisons d'IA ouvre de nouvelles perspectives. Les **SVM** et **l'apprentissage par renforcement** se combinent par exemple très efficacement, et ce dernier peut être couplé avec **l'apprentissage profond des réseaux de neurones** pour des résultats encore plus performants<sup>2</sup>. Ce dernier, appelé *Deep Learning*, peut aussi s'enrichir de logiques floues ou d'algorithmes génétiques et trouve de nombreuses applications dans le domaine de la reconnaissance de formes (lecture de caractères, reconnaissance de signatures, de visages, vérification de billets de banque), du contrôle de processus ou de prédictions.

---

<sup>1</sup> Pat Langley récapitule les architectures cognitives articulant plusieurs types d'IA qui permettraient d'aboutir à des théories unifiées de l'esprit. Ce paradigme a selon lui reçu peu d'attention du monde de la recherche et pose des défis à toute la communauté scientifique. Cf. Pat Langley, 2017, « Progress and Challenges in Research on Cognitive Architectures », vol. 31, n° 1, Proceedings of the Thirty-First Association for the Advancement of Artificial Intelligence Conference.

<sup>2</sup> L'efficacité est avérée pour le traitement automatique du langage naturel, la reconnaissance automatique de la parole, la reconnaissance audio, la bio-informatique ou, encore, la vision par ordinateur.

**Les combinaisons de technologies d'intelligence artificielle** mises au point par Google Deep Mind vont aussi dans ce sens, en utilisant tant des outils traditionnels comme la méthode Monte-Carlo (dont la recherche en arborescence s'apparente à l'approche de l'IA symbolique) que des systèmes plus récents comme l'apprentissage profond ou l'apprentissage par renforcement.

Le programme **AlphaGo** a ainsi appris à jouer au jeu de Go par une **méthode de Deep Learning** certes, mais couplée à un **apprentissage par renforcement** lors des parties jouées et à une optimisation selon la **méthode Monte-Carlo**.

De même, le système **AlphaGeometry** lui aussi élaboré par Google Deep Mind associe très efficacement un système de règles issu de l'IA symbolique avec des réseaux de neurones profonds issus de l'IA connexionniste.

Les déclarations d'OpenAI en novembre 2024 vont également dans ce sens : **plutôt que de se précipiter vers la mise sur le marché de ChatGPT-5**, les priorités des chercheurs doivent être de **faire progresser la capacité de leurs modèles en termes de raisonnement logique, de tâches de planification et de développement de l'autonomie des systèmes**<sup>1</sup>.

L'épistémologue Jean Piaget, fondateur du structuralisme génétique<sup>2</sup> et spécialiste de l'apprentissage, disait que « *l'intelligence, ça n'est pas ce que l'on sait, mais ce que l'on fait quand on ne sait pas* » Cette intelligence en tant que capacité à **résoudre des problèmes en l'absence de réponses évidentes** représente un défi scientifique pour la recherche en intelligence artificielle.

Il en ressort une complexité croissante des algorithmes : **la complexité de certains algorithmes récents est telle qu'au final, ils peuvent être comparés à des sortes de cathédrales géantes multidimensionnelles**. Leur construction est un art complexe, encore plus qu'une science.

---

<sup>1</sup> Cf. à ce sujet des défis de l'IA, la très récente session AMA (pour « Ask me anything ») organisée par OpenAI le 1<sup>er</sup> novembre 2024 avec Sam Altman, Kevin Weil, Srinivas Narayanan et Mark Chen diffusée sur le réseau social Reddit.com :

[https://www.reddit.com/r/ChatGPT/comments/1ggixzy/ama\\_with\\_openais\\_sam\\_altman\\_kevin\\_weil\\_srinivas/](https://www.reddit.com/r/ChatGPT/comments/1ggixzy/ama_with_openais_sam_altman_kevin_weil_srinivas/)

<sup>2</sup> Généralement associé à la sociologie de Pierre Bourdieu, le structuralisme génétique s'est tout d'abord largement construit avec l'œuvre de Jean Piaget puis de son élève Lucien Goldmann. Il s'agit d'une forme particulière de structuralisme dans lequel la formation des structures – leur genèse – et donc la dimension historique sont au cœur de l'analyse. Après ces deux initiateurs en épistémologie et en psychologie, les travaux de Pierre Bourdieu ont permis d'enrichir cette approche à la lumière de la sociologie. Ce paradigme, que le sociologue appelait aussi de manière équivalente constructivisme structuraliste, s'inspire aussi de la linguistique générative de Noam Chomsky afin de montrer que l'ordre social, malgré son déterminisme, se fonde toujours sur l'actualisation des règles à travers des dimensions symboliques.



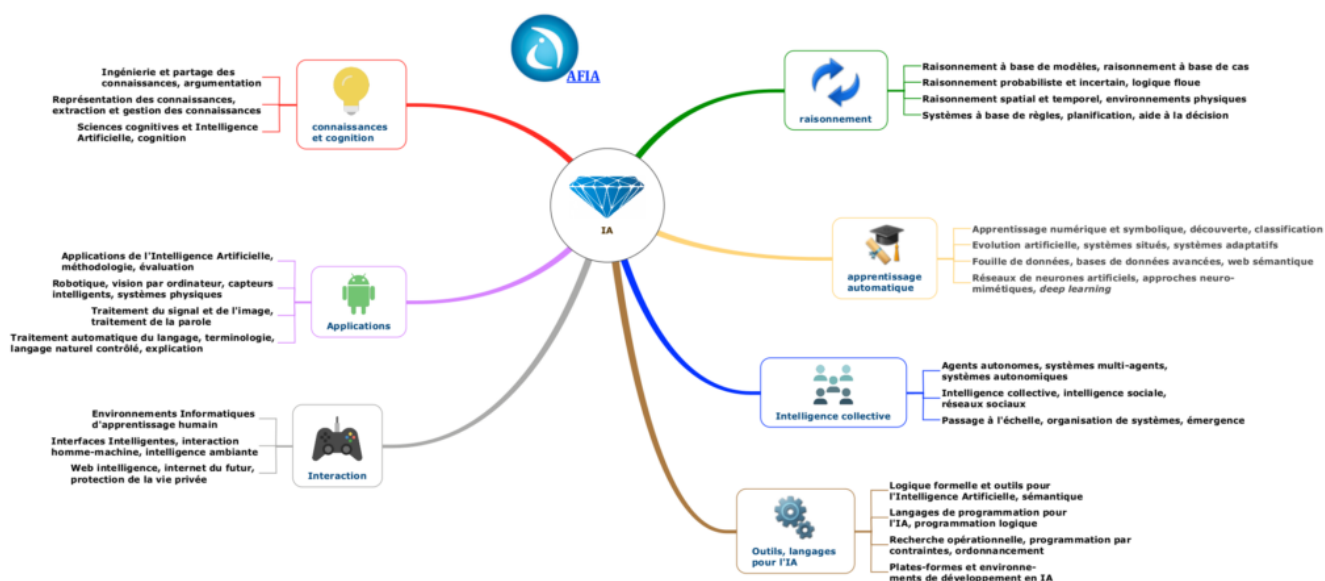
L'algorithmique, domaine qui étudie les algorithmes devrait, en tant que support conceptuel de la programmation des ordinateurs, devenir un **art de plus en plus exigeant**, une science mais aussi bien plus qu'une science, comme l'explique de façon éloquent Donald Knuth professeur émérite de « *l'art de programmer en informatique* » (selon ses mots) à l'université de Stanford et auteur du manuel de référence de l'algorithmique dont le titre éloquent est *L'Art de la programmation informatique*<sup>1</sup>.

d) *La variété des domaines de l'intelligence artificielle*

Selon l'Association française pour l'intelligence artificielle (AFIA), derrière l'étiquette IA, il existe **au moins 26 domaines différents**. L'ingénierie des connaissances, le traitement automatique du langage (TAL), les systèmes à base de règles, l'apprentissage symbolique ou encore les réseaux de neurones artificiels, forment, chacun, l'un de ces domaines.

Cette association savante française a dessiné sur cette base ce qu'elle appelle le « diamant de l'IA selon l'AFIA ». Cette décomposition de l'IA montre la grande variété des formes que l'IA peut revêtir.

### Les domaines de l'IA selon l'AFIA



Source : <https://afia.asso.fr/domaines-de-lia/>

<sup>1</sup> Cf. Donald Knuth, *The Art of Computer Programming (en 5 volumes, publiés depuis 1997 pour le premier volume)*, Addison-Wesley.

### III. LE GRAND MARCHÉ DE L'IA : UNE CHAÎNE DE VALEUR ÉTENDUE DES MATIÈRES PREMIÈRES AUX CONSOMMATEURS

L'intelligence artificielle dépasse sa dimension de progrès technologique principalement logiciel, il faut aussi la considérer comme **source de puissance** et comme **enjeu géopolitique** sur l'ensemble des aspects de la chaîne de valeur qu'elle recouvre. La question des **infrastructures matérielles** qui permettent de fournir la puissance de calcul permettant d'entraîner et d'utiliser des modèles de plus en plus performants est un aspect important à examiner dans ce cadre.

#### A. DU SILICIUM AUX APPLICATIONS : LA CHAÎNE DE VALEUR COMPLEXE DE L'INTELLIGENCE ARTIFICIELLE

##### 1. Un schéma souvent simplifié en quatre étapes

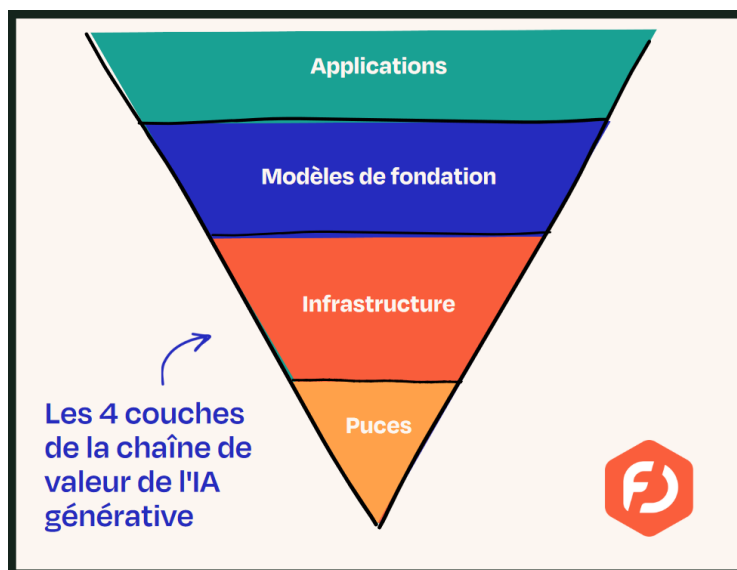
L'association France Digitale, qui représente les intérêts des *start-up* françaises du numérique a publié en avril 2024 un rapport intitulé « *Des puces aux applications, l'Europe peut-elle être une puissance de l'IA générative* »<sup>1</sup>. Ce document présente la chaîne de valeur de l'IA de manière simplifiée en quatre étapes :

- premièrement, les puces en silicium, qui comprennent à la fois le matériel physique et les matières premières nécessaires à leur élaboration ;
- la deuxième étape de la chaîne de valeur est l'infrastructure, qui supporte la fabrication du logiciel d'intelligence artificielle. Ce maillon de la chaîne comprend les centres de données dans lesquels sont stockés les microprocesseurs, le réseau qui peut permettre l'accès à des centres de données distants, et les logiciels qui permettent la programmation ;
- vient ensuite l'étape du « modèle de fondation », c'est-à-dire la conception d'un modèle « brut » grâce à une immense quantité de données. Ces modèles ne sont pas réglés pour une utilisation spécifique. C'est par exemple le modèle GPT-4o, utilisé pour construire la dernière version en date de l'agent conversationnel ChatGPT, qui lui, est une application ;
- enfin, la chaîne de valeur se termine par la strate des applications. Il s'agit de l'utilisation des modèles de fondation pour développer des applications qui répondent à des cas d'usage concrets par le biais du processus de réglage fin. Les utilisateurs ont accès aux modèles *via* des interfaces qui peuvent répondre à leurs besoins.

---

<sup>1</sup> Tordeux Bitker, Marianne et al. 2024. Des puces aux applications, l'Europe peut-elle être une puissance de l'IA générative? France Digitale. <https://francedigitale.org/publications/etude-ia-generative>.

### Schéma simplifié de la chaîne de valeur de l'IA générative



Source : France Digitale

L'Office identifie quant à lui, au terme de ses auditions et déplacements, une **dizaine de niveaux** en raffinant cette première analyse.

## 2. La complexité de la dizaine d'étapes de la chaîne

### a) L'amont : du silicium aux microprocesseurs

Chacune de ces quatre étapes peut en réalité être elle-même découpée en plusieurs niveaux. Par exemple, la première couche dite des « puces » concerne tout autant l'énergie nécessaire à l'ensemble des processus, dont les besoins sont grandissants<sup>1</sup>, l'extraction et la manipulation de matières premières pour construire les semi-conducteurs en silicium que la construction des puces, les logiciels qui permettent de les concevoir et de les utiliser ou, encore, les machines lithographiques capables de creuser le silicium à l'échelle microscopique.

<sup>1</sup> Les besoins en énergie vont être croissants et conduisent les acteurs du secteur à se positionner sur le marché de l'énergie, notamment sur celui de l'énergie nucléaire. La Big Tech américaine y est d'ores et déjà un acteur important : avec Constellation Energy, Microsoft prépare la remise en service d'ici 2028 du fameux réacteur de Three Mile Island ; et Alphabet (Google), Amazon ou, encore, OpenAI investissent massivement dans des projets des petits réacteurs nucléaires, les Small Modular Reactors (SMR).

Cf. <https://www.usine-digitale.fr/article/ia-generative-microsoft-relance-la-centrale-nucleaire-de-three-mile-island-pour-alimenter-ses-data-centers.N2219114>

Le dioxyde de silicium, ou silice, certes abondant sur terre, n'est pas du simple sable ou granit fondu : il s'agit d'un cristal de silice monocristallin, soumis au procédé de Czochralski<sup>1</sup> pour l'homogénéiser ainsi qu'au dopage (ajout d'impuretés<sup>2</sup>) pour déterminer les propriétés de conductivité recherchées pour chaque semi-conducteur. Ce dernier est, en effet, à la fois, comme son nom l'indique, isolant et conducteur, il s'agit d'un matériau intermédiaire entre les isolants et les conducteurs électriques. Sa pureté à 99,99 % est donc doublement contrôlée, par le procédé de Czochralski et par le procédé de dopage.

### **Semi-conducteurs, transistors, circuits intégrés et microprocesseurs**

**Ces quatre notions, parfois confondues, sont pourtant à distinguer** : les premiers sont la base matérielle des transistors, ceux-ci sont regroupés dans des circuits intégrés et ces derniers lorsqu'ils permettent avec une unité unique de faire fonctionner un ordinateur prennent le nom de microprocesseurs.

Dès 1833, Faraday constate qu'à la différence de la plupart des métaux dont la résistance augmente avec la température, certains métaux comme le sulfate d'argent sont de plus en plus conducteurs avec la hausse de la température (coefficient de température négatif)<sup>3</sup>. Un siècle plus tard, en 1931, Wilson théorise les **semi-conducteurs** : les électrons forment des ondes dans les solides et la conduction électrique de certains matériaux varie en fonction de divers facteurs comme la température, le courant électrique ou la lumière.

Au moment où les travaux se concentrent sur l'utilisation du silicium et du germanium, semi-conducteurs à haut point de fusion, trois chercheurs américains des Bell Labs (Bardeen, Shockley et Brattain) construisent en 1947, à partir de ces semi-conducteurs, le premier **transistor**.

---

<sup>1</sup> Ce procédé, utilisé dans l'industrie électronique, consiste à contrôler la croissance de cristaux monocristallins pour les faire grossir et atteindre de grandes dimensions (plusieurs centimètres). Les germes cristallins sont fondus sous atmosphère neutre à une température juste au-dessus du point de fusion, avec un gradient de température contrôlé, puis un germe monocristallin est placé dans une navette suspendue au-dessus du liquide par une tige tournant très lentement. Cette solidification dirigée du liquide sur le germe monocristallin permet par épitaxie de créer des cristaux ayant tous la même structure. Le barreau cylindrique ainsi créé est ensuite découpé pour être utilisé sous forme de galettes (ces tranches de matériau sont appelées wafers).

<sup>2</sup> Il s'agit en effet d'ajouter aux cristaux de silicium qui composent le support, et après la phase de gravure, des impuretés en petites quantités sous la forme d'atomes d'un autre matériau afin de modifier ses propriétés de conductivité.

<sup>3</sup> Cf. Guy Renard, 1963, « La découverte et le perfectionnement des transistors », Revue d'histoire des sciences et de leurs applications, tome 16, n° 4 spécial Documents pour l'histoire des techniques. pp. 323-358.

Cette technologie, qui correspond au contrôle ou à l'amplification du courant électrique par la combinaison de trois électrodes<sup>1</sup>, soit aussi parallèlement développée en France<sup>2</sup>. C'est, en définitive, Texas Instruments qui va à ce moment capter le marché mondial avec la fabrication des premiers transistors en silicium puis des radios à transistor à partir de 1954.

En 1958, c'est encore Texas Instruments, grâce à l'ingénieur Jack Kilby, qui réalise le premier **circuit intégré** ou « puce », dispositif rassemblant plusieurs transistors interconnectés en circuits microscopiques sur un support alors en germanium et dans un même petit boîtier.

De quelques transistors dans les années 1960 (moins de dix), ces circuits passent le cap du million de transistors à la fin des années 1980, avec les circuits intégrés dédiés aux ordinateurs appelés **microprocesseurs**. Ces derniers, qui exécutent au sein d'une puce unique des instructions et traitent les données des programmes à l'aide de logiques binaires, ont été inventés par Intel en 1971 (avec un premier système à 2 300 transistors). Le cap des 42 millions de transistors est atteint en 2000 avec le microprocesseur Pentium 4, celui du milliard en 2010, avec le microprocesseur Intel Core i7. En 2024, le microprocesseur Blackwell de Nvidia gravé à 4 nm près<sup>3</sup> par TSMC et présenté à vos rapporteurs au siège de l'entreprise dans la Silicon Valley compte **208 milliards de transistors**.

Le cofondateur de l'entreprise Intel, Gordon Moore, a laissé son nom à la prédiction, qu'il établit dès 1965 et précisa en 1975, du doublement du nombre de transistors présents sur une puce de microprocesseur tous les deux ans. Cette augmentation exponentielle, nommée « **loi de Moore** », liée à la miniaturisation, est observée empiriquement jusqu'aux années 2010. Pour qu'elle reste valable, il a fallu prendre en compte le traitement parallèle des calculs dans des architectures multicœurs. Le mur des limites physiques des microprocesseurs, avec le fait d'approcher de la taille moléculaire, est atteint. La contrepartie de cette loi de Moore - fondée sur la réduction de la taille des traits gravés dans le silicium permettant d'augmenter la densité des processeurs et par conséquent leur vitesse - est la « **loi de Rock** » (du nom d'Arthur Rock) selon laquelle le coût d'une fonderie de semi-conducteurs double quant à lui tous les quatre ans, sous l'effet de procédés de fabrication de plus en plus chers.

---

<sup>1</sup> Le courant électrique est introduit par une électrode (appelée émetteur sur un transistor bipolaire ou source sur un transistor à effet de champ) et contrôlé sur l'électrode de sortie (appelée collecteur sur un transistor bipolaire ou drain sur un transistor à effet de champ) sous l'effet de l'électrode d'entrée (appelée base sur un transistor bipolaire et grille sur un transistor à effet de champ). Si le transistor est isolant sans tension sur la borne base, il devient conducteur avec une tension sur la borne base.

<sup>2</sup> Le transistor est en effet parallèlement développé en France par deux chercheurs allemands de la Compagnie des Freins et Signaux conduisant à une demande de brevet en 1948. Leur système, plus résistant et plus stable que celui des Américains, permet de mettre sur le marché en 1952 la première radio à transistor, sans l'appui des autorités nationales qui estiment alors prioritaire l'industrie nucléaire. Cf. sur ces points Michael Riordan, 2005, « How Europe Missed The Transistor », IEEE spectrum : <https://spectrum.ieee.org/how-europe-missed-the-transistor>

<sup>3</sup> Un nanomètre, de symbole nm, équivaut à un milliardième de mètre ( $10^{-9}$ ), ce qui représente la taille moyenne d'une molécule.

Ces différents processus sont assumés par différentes entreprises, même si **Nvidia** semble être devenu, pour le moment, l'acteur dominant de ce premier maillon de la chaîne. Cette entreprise américaine de la Silicon Valley s'est spécialisée dès sa naissance sur les **puces d'accélération pour interface graphique**.

Son premier partenariat stratégique est signé en 1994 avec l'entreprise française Thomson et ses premières puces sont commercialisées un an plus tard. Cependant, sa particularité est de produire des puces et des logiciels mais de **ne rien fabriquer elle-même** : pour ses semi-conducteurs, l'entreprise repose sur l'offre des sociétés de fonderie de semi-conducteurs.

Elle devient ainsi partenaire de l'entreprise de fonderie taïwanaise TSMC à partir de 1998. Puis, un an plus tard, elle commercialise la première carte vidéo **GPU** grand public. Ces **processeurs graphiques** dont l'acronyme vient de l'anglais *Graphics Processing Units* prennent en charge, grâce à des structures de calculs parallèles, l'intégralité des calculs graphiques et optimisent l'affichage 2D et 3D ou encore les vidéos. Leur usage a notamment été accéléré par le succès des jeux vidéo en 3D. Le parallélisme de ces processeurs a conduit à leur utilisation **dans le calcul matriciel des nouvelles technologies d'IA des années 2000 et 2010** telles que le *Deep Learning*.

En 2023, le chiffre d'affaires de Nvidia représentait 71 milliards de dollars, et en juin 2024, avec 3 335 milliards de dollars, l'entreprise était **la société avec la plus forte valorisation au monde**. En novembre 2024, après avoir été redoublée par Apple quelques mois, sa valorisation de **3 430 milliards de dollars** en a refait la première entreprise au monde.

Si Nvidia ne fabrique pas de puces mais les conçoit, comme le font AMD ou Qualcomm, la production de semi-conducteurs en silicium est un **marché assez monopolistique**, car les coûts fixes sont prépondérants, ce qui le rapproche des conditions d'un **monopole naturel**.

En dépit de cette caractéristique, cette première étape de la production des puces voit plusieurs entreprises se partager le marché mondial.

Les fabricants de semi-conducteurs sont en effet répartis entre :

- ceux qui conçoivent, fabriquent et commercialisent les puces, que l'on qualifie d'entreprises mixtes ;

- ceux qui ne font que produire les semi-conducteurs et les microprocesseurs (souvent conçus par leurs clients), appelés les fondeurs ;

- et ceux qui conçoivent et commercialisent les microprocesseurs sans les produire (on parle de *fabless*), choisissant de sous-traiter la fabrication à des sociétés de fonderie.

Le leader mondial de cette dernière catégorie, avec plus de la moitié du marché de la fonderie de semi-conducteurs, est **TSMC** (pour *Taiwan Semiconductor Manufacturing Company Limited*), fournisseur de Nvidia, mais aussi d'AMD, d'Apple, de Qualcomm, etc.

Même Intel, Texas Instruments ou STMicroelectronics, qui conçoivent, fabriquent et commercialisent des puces, confient une part de leur production à TSMC. Son chiffre d'affaires représentait 71 milliards de dollars en 2023 et le groupe a rejoint en juillet 2024 le cercle restreint des entreprises valorisées à plus de 1 000 milliards de dollars.

Les trois autres géants de ce marché de la fonderie sont **GlobalFoundries** (ex-fonderie du groupe américain AMD vendue aux Émirats arabes unis), le taïwanais **UMC** (*United Microelectronics Corporation*) et le chinois **SMIC** (*Semiconductor Manufacturing International Corporation*).

L'entreprise sud-coréenne **Samsung** reste quant à elle leader sur la filière semi-conducteurs hors fonderie et *fabless*, activité dans laquelle elle fait la course en tête avec Intel et qui représente un quart de son chiffre d'affaires.

Le seul Européen qui surnage dans ce contexte oligopolistique, est la société néerlandaise **ASML** qui s'est spécialisée dans la fabrication de machines de photolithographie nécessaires à la gravure des microprocesseurs sur les galettes de silicium. Elle fournit l'ensemble des fabricants de microprocesseurs, notamment les plus gros comme Samsung, Intel et TSMC, et a pour concurrents les Japonais Nikon et Canon.

#### *b) La couche multiforme des infrastructures*

Le deuxième maillon de la chaîne, celui des infrastructures, se subdivise en de nombreuses couches :

- la **collecte** et le **nettoyage** de données<sup>1</sup> ;
- le **stockage de données** dans de vastes *data centers* ;
- l'**informatique en nuage** (*cloud*) pour les **calculs** ;
- et, lors de la phase de développement des modèles, le recours spécifique à des **supercalculateurs**.

---

<sup>1</sup> La collecte et le nettoyage de données sont des activités à part entière : la data extraction peut prendre la forme du web scraping, le tri ou data indexing peut prendre celle du web crawling, le nettoyage peut prendre celle de la data curation. Les web crawlers ont progressivement construit un référentiel ouvert des données issues d'Internet, appelé Common Crawl. Ces données sont nettoyées grâce à l'utilisation de divers filtres. Ces techniques peuvent par exemple permettre de supprimer des discours de haine, des expressions offensantes et d'autres contenus indésirables. Les IA génératives ont aussi recours au corpus de Wikipédia, aux référentiels GitHub (désormais propriété de Microsoft) ou encore au projet Gutenberg, qui compile plus de 50 000 ouvrages tombés dans le domaine public. Un guide préparé par des développeurs de modèles de fondation pour les développeurs de modèles de fondation existe et va plus loin que les données textuelles en s'intéressant aussi à l'image et au son (via les audiobooks par exemple) : cf. Shayne Longpre et al., 2024 « The Foundation Model Development Cheatsheet », a guide prepared by foundation models developers for foundation models developers, GitHub : <https://github.com/allenai/fm-cheatsheet/commits/main/app/resources/paper.pdf>

La multiplication de ces infrastructures et les progrès qualitatifs de leurs équipements conduisent à des **coûts immenses**<sup>1</sup> et à des **impacts environnementaux** considérables dont il sera question plus loin.

Selon une analyse de Sequoia Capital de juin 2024, intitulée « La Question à 600 milliards de dollars de l'IA », le coût des infrastructures de Nvidia (réparti à moitié entre les GPU et le reste, c'est-à-dire l'énergie, les bâtiments, les générateurs de secours, etc.) devra conduire à un chiffre d'affaires de 600 milliards de dollars pour permettre un retour sur investissement<sup>2</sup>. L'auteur avait estimé ce montant à 200 milliards de dollars en septembre 2023.

*c) L'étape de définition des modèles : de leur architecture au réglage fin*

Cette couche de la chaîne de valeur de l'intelligence artificielle qui permet de concevoir les modèles d'IA est elle-même subdivisée entre plusieurs étapes :

- la **conception de l'architecture** du modèle ;

- l'**entraînement du modèle** à l'aide des infrastructures vues précédemment et d'algorithmes qui sont des logiciels d'entraînement ;

- le **réglage fin** qui prend les formes d'apprentissages supervisés et d'une phase « d'alignement », cette phase pouvant être sous-traitée comme le montre l'exemple du recours à Sama par OpenAI.

Les **modèle de fondation s'intercalent donc entre la définition de l'architecture du modèle et le fine-tuning permettant aux systèmes d'IA d'être déployés** pour telle ou telle application spécifique, voire parfois d'être diffusés auprès du grand public.

*d) L'aval : les applications*

En dernier lieu, les systèmes d'IA ne sont au final accessibles aux utilisateurs qu'avec une **couche de services applicatifs** (dont les fameux LLM d'OpenAI, de Meta ou de Mistral, comme ChatGPT, Llama et Mistral Large). Cette étape est indispensable pour permettre la diffusion de la technologie des laboratoires de recherche des entreprises vers les consommateurs, voire le grand public.

Comme pour Google en matière de moteur de recherche, Amazon en matière de e-commerce ou Meta en matière de réseaux sociaux, cette couche applicative spécialisée tend aussi à être **monopolistique**.

---

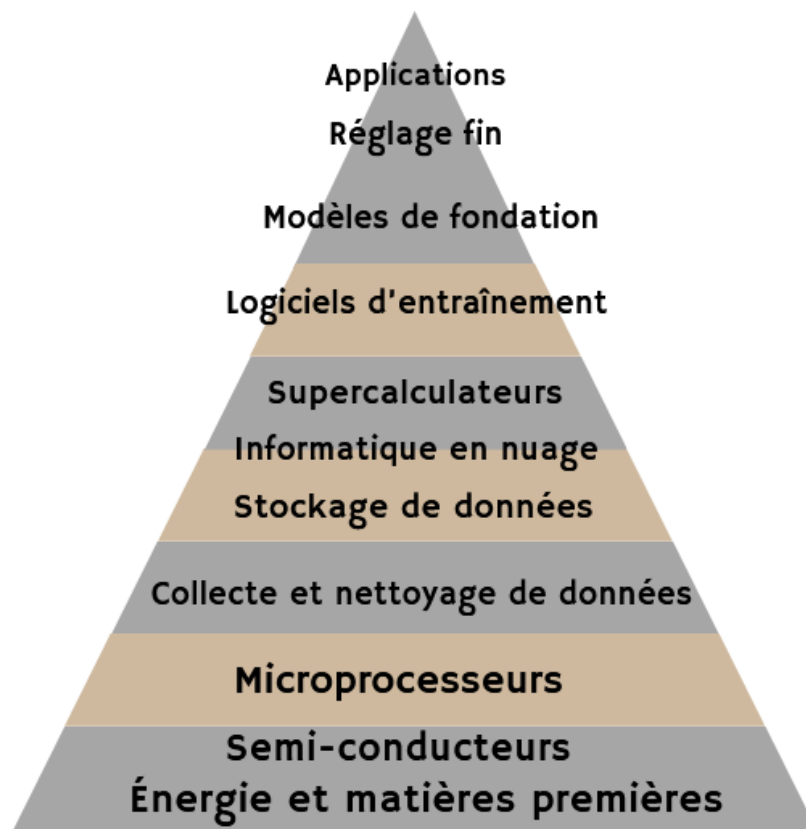
<sup>1</sup> Cf. <https://www.usine-digitale.fr/article/la-bulle-de-l-ia-se-poursuit-avec-une-montee-en-fleche-des-couts.N2215681>

<sup>2</sup> Cf. l'étude de David Cahn de Sequoia Capital : <https://www.sequoiacap.com/article/ais-600b-question/>



**ChatGPT**, système d'IA commercialisé par OpenAI – entreprise financée par Microsoft qui détient la moitié de son capital –, a ainsi attiré la majorité des utilisateurs de LLM et reste, depuis son lancement en novembre 2022, le produit dominant sur le marché. Cette position devrait être renforcée par son intégration sous le nom de **Copilot** dans les logiciels Office de Microsoft (Word, Excel, Outlook, Teams et PowerPoint), dans le navigateur Edge de Microsoft et dans le moteur de recherche Microsoft Bing (qui cherche à contester le monopole actuel de Google sur ce type d'applications).

#### Schéma de la chaîne de valeur complexe de l'IA



La chaîne de valeur de l'intelligence artificielle est donc complexe, composée de couches, chacune **souvent proche de conditions de monopole naturel**. Aucune puissance n'est en mesure de posséder aujourd'hui l'ensemble de cette chaîne de valeur sur son seul marché intérieur, seule caractéristique à même de donner une véritable souveraineté en IA. Dans l'hypothèse où cela arriverait, l'État concerné disposerait alors d'un contrôle complet sur le développement des systèmes d'intelligence artificielle.

Même les États-Unis, acteur superdominant du secteur, dépendent encore largement de la fabrication des puces en dehors de leur territoire.

Il faut bien avoir conscience que les coûts correspondants au développement de chacune de ces couches sont considérables. Selon l'Index de l'IA publié chaque année par le Stanford Institute for Human-Centered Artificial Intelligence (HAI)<sup>1</sup>, le **simple entraînement** des modèles prend des mois et coûte de l'ordre d'une **centaine de millions de dollars** : les calculs ayant abouti à GPT-4 d'OpenAI auraient représenté un coût de 78 millions de dollars et ceux de Gemini Ultra de Google 191 millions de dollars.

Selon le PDG d'Anthropic, Dario Amodei, les modèles en cours de développement auront un coût unitaire approchant désormais **le milliard de dollars**. Dans cette interview d'avril 2024, il déclare également que les générations de modèles qu'il prévoit de mettre à disposition en 2025 et 2026 coûteront, chacun, entre **5 et 10 milliards de dollars**<sup>2</sup>.

Compte tenu de ces coûts très significatifs, **les modèles de pointe ne sont et ne seront donc développés, ceteris paribus, que par de très grandes entreprises technologiques**. Les petites entreprises, tout comme les universités et les organismes publics de recherche, rencontreront de plus en plus de difficultés à développer des modèles avancés d'IA, du fait de leurs ressources limitées. Les **difficultés de l'Université la plus riche en matière technologique, Stanford**, à suivre les développements en cours, surtout en termes de nombres de microprocesseurs et donc de capacités de calcul, ont été soulignées à vos rapporteurs lors de leurs rencontres dans la fameuse Université du comté de Santa Clara.

On pourrait dès lors douter de la crédibilité de l'ambition de développer une filière nationale autonome en IA. Pourtant, on compte des **entreprises françaises et européennes réparties sur l'ensemble de cette chaîne de valeur**.

Dans le domaine des puces, **Imerys**, société minière française, leader mondial des spécialités minérales et des matériaux avancés, possède ainsi une branche spécialisée dans la production de quartz, à partir duquel peut être obtenu du silicium<sup>3</sup>, essentiel comme il a été vu à la construction de matériaux semi-conducteurs.

---

<sup>1</sup> Pour son édition 2024 parue le jour même du déplacement de vos rapporteurs auprès des chercheurs de l'Université de Stanford, cf. Artificial Intelligence Index Report 2024, Stanford Institute for Human-Centered Artificial Intelligence, disponible au lien : <https://aiindex.stanford.edu/report/>  
Il est aussi loisible de se reporter à des références plus anciennes : Or Sharir et al., 2020, The Cost of Training NLP Models : A Concise Overview, disponible sur Arxiv, <https://arxiv.org/pdf/2004.08900> ainsi que Lennart Heim, 2022, Estimating PaLM's training cost, Lennart Heim Blog <https://blog.heim.xyz/palm-training-cost/>

<sup>2</sup> Cf. l'interview de Dario Amodei par Ezra Klein retranscrite dans le New York Times du 12 avril 2024 : <https://www.nytimes.com/2024/04/12/podcasts/transcript-ezra-kleininterviews-dario-amodei.html>

<sup>3</sup> Pour en savoir plus sur l'extraction et la géopolitique du silicium, voir cet article : <https://www.mineralinfo.fr/fr/ecomine/silicium-un-element-chimique-tres-abondant-un-affinage-strategique>

La France possède directement sur son sol des entreprises de semi-conducteurs comme **Dolphin Designs** ou **STMicroelectronics**, devenue une multinationale franco-italienne de droit néerlandais cotée au New York Stock Exchange et dont le siège social est en Suisse (BPI France et l'Italie restant les premiers actionnaires) mais qui dans le contexte du « plan composants » de 1978 est d'abord issue d'une **coopération entre le CEA et Thomson** à travers SGS-Thomson. En 2005, STMicroelectronics se classait encore à la 5<sup>e</sup> place du classement mondial des entreprises de semi-conducteurs, derrière Intel, Samsung, Texas Instruments et Toshiba. Les entreprises françaises **Defacto Technologies**, **Flex.ai** ou **Sipearl** sont également spécialisées dans la conception de puces (CPU pour Sipearl, ASIC pour Flex.ai, plateforme logicielle RTL de conception de circuits intégrés complexes pour Defacto Technologies).

D'autres aspects de cette couche matérielle de l'IA existent en Europe, par exemple l'entreprise néerlandaise *Advanced Semiconductor Materials Lithography* (ASML), qui produit les machines lithographiques capables de graver le silicium à l'échelle microscopique ; elle est même le leader mondial de cette activité.

Mais **dans certains autres secteurs** de cette première étape de la chaîne de valeur, **ni la France ni l'Europe ne possèdent aujourd'hui d'entreprises.**

C'est le cas pour la **fabrication de cartes graphiques** (*graphics processing unit* ou GPU), domaine hautement capitalistique dominé par les entreprises américaines Nvidia, AMD et Intel. Il en est de même pour les **fonderies de silicium**, domaine dans lequel il n'existe actuellement **aucune alternative française ou européenne** en état d'entrer sur le marché et de rentrer dans la compétition mondiale.

**La France et l'Europe n'ont donc aujourd'hui d'autre choix que de se tourner en matière d'intelligence artificielle vers le marché mondial, au moins à court et moyen terme et pour une partie de la chaîne de valeur.** Cela rend nécessaire de se pencher sur la place de la France dans ce marché mondial ainsi que sur les rapports de forces et les coopérations en cours ou à venir concernant l'intelligence artificielle.

## B. LA MULTIPLICATION DES MODÈLES PROPOSÉS AU GRAND PUBLIC

### 1. Une domination nette de l'IA générative par la *Big Tech* américaine

En 2017, le rapport de l'Office soulignait déjà en la déplorant « *la place prépondérante occupée par la recherche privée, y compris sur le plan de la recherche fondamentale* » et le fait que cette recherche était « *dominée par les entreprises américaines* », devenues pôles d'attraction pour tous les chercheurs du monde, y compris ceux issus de la recherche publique, conduisant à une « *concentration des compétences au sein des entreprises privées américaines* »<sup>1</sup>. Les chercheurs et développeurs sont souvent sur le territoire des États-Unis mais on les retrouve dans le reste du monde, « *ce qui leur permet de perfectionner leurs algorithmes à moindre coût* », affirmait l'Office. Un des aspects de cette domination participe du « *Big Data* » : les technologies d'apprentissage machine, dont le *Deep Learning* et ses *Transformers*, recourent à **des méthodes statistiques qui nécessitent des données massives** pour être efficaces « *or ces entreprises disposent d'un avantage comparatif difficile à rattraper : des jeux de données massives, continuellement enrichis par leurs clients et usagers* »<sup>2</sup>.

Dans ce contexte, **les pays du reste du monde ont tendance à ne plus pouvoir être producteurs** de technologies d'IA de pointe mais à s'en tenir au rôle de **simples consommateurs** avec toutes les conséquences que cette situation peut avoir.

Dès 2013, notre collègue sénatrice Catherine Morin-Desailly se demandait dans un rapport si l'Union européenne n'était devenue qu'une **simple colonie dans le monde du numérique**<sup>3</sup>. Elle y précisait que « *les entreprises américaines développent des modèles d'affaires complexes, souvent hybrides : ces modèles d'affaires, qui plus est en recomposition permanente, peuvent reposer sur une offre combinant produits logiciels, services en ligne, plateformes de cloud computing, voire matériel. Le pouvoir de marché qu'ils acquièrent en offrant un service gratuit au plus grand nombre d'utilisateurs leur permet de capter la marge des entreprises du secteur présentes sur l'autre face du modèle* ».

Le terme de **GAFAM**, désignant les grands groupes américains du numérique, soit Google, Apple, Facebook, Amazon et Microsoft devient, à plusieurs égards, de plus en plus dépassé malgré son usage généralisé.

---

<sup>1</sup> Cf. le rapport précité de l'OPECST, pp. 81 et 82.

<sup>2</sup> Cf. rapport précité de l'OPECST, p. 82.

<sup>3</sup> Cf. Catherine Morin-Desailly, L'Union européenne, colonie du monde numérique ?, rapport d'information n° 443 (2012-2013) fait au nom de la commission des affaires européennes : <https://www.senat.fr/notice-rapport/2012/r12-443-notice.html>

Tout d'abord, plusieurs de ces groupes ont changé de noms depuis la diffusion de ce sigle, ainsi les maisons mères de Google et Facebook – **Alphabet** et **Meta** – sont devenues les dénominations usuelles de ces deux groupes. De plus, le terme ne présente pas ces entreprises en fonction de leur importance ou de leur capitalisation boursière respective.

Aussi, vos rapporteurs recommandent d'utiliser le **palindrome MAAAM** désignant, dans l'ordre, Microsoft, Apple, Alphabet, Amazon et Meta. Ainsi avec ce sigle, non seulement les noms des groupes sont actualisés, mais, en plus, ils sont triés dans un ordre plus proche de la hiérarchie de leur capitalisation boursière, de la plus importante à la moins importante, même si Apple tend parfois à devancer Microsoft (par exemple en octobre 2024, 3 500 milliards de dollars contre 3 200 milliards de dollars).

À cette liste, on pourrait également désormais ajouter l'entreprise de cartes graphiques **Nvidia**, qui, tirée par le boom de l'IA, a dépassé le seuil des 3 335 milliards de dollars de capitalisation boursière en juin 2024<sup>1</sup> et atteignait même les 3 530 milliards de dollars en octobre et novembre 2024<sup>2</sup>, sur le chemin des 5 000 milliards de dollars anticipés par les analystes financiers<sup>3</sup>. Nvidia est le fournisseur de tous les MAAAM et possède un avantage comparatif qu'il sera difficile de contester, en dépit de sa pratique de prix élevés.

Comme l'avait fait le rapport précité de l'Office de 2017, il faut y ajouter les entreprises « historiques » de l'informatique, à commencer par **IBM**.

**Toutes ces entreprises**, auditionnées par vos rapporteurs, **sont américaines**, la domination de ce pays sur le marché de l'IA, et notamment de l'IA générative, est donc indéniable.

Alors que ces géants du numérique pouvaient être considérés comme potentiellement menacés par les géants chinois Baidu, Alibaba, Tencent et Xiaomi (BATX) il y a quelques années, il apparaît aujourd'hui que **la Chine rencontre en réalité de plus en plus de difficultés à rivaliser sur le marché mondial**, en termes de valorisation financière et commerciale de ses produits, avec les géants américains, dont la croissance creuse l'écart jour après jour.

---

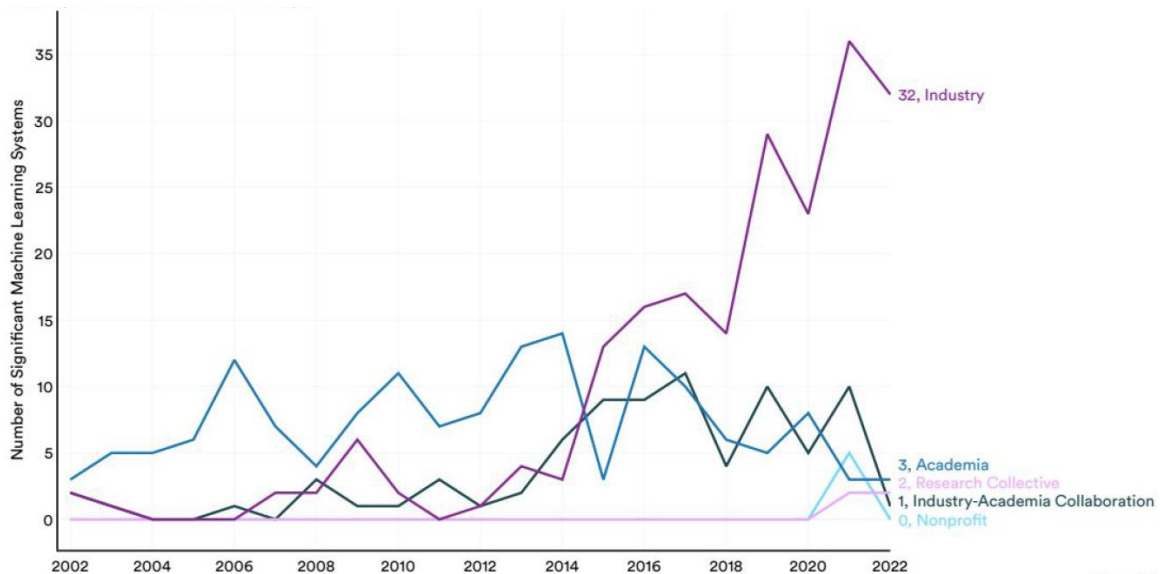
<sup>1</sup> Cf. Kif Leswing, 2024, « Nvidia Hits \$3 Trillion Market Cap on Back of AI Boom », CNBC : <https://www.cnbc.com/2024/06/05/nvidia-briefly-passes-3-trillion-market-cap-on-back-of-ai-boom.html>

<sup>2</sup> Cf. cet article, « Nvidia Stock Pumps 4 % to Record as Chip Maker's Valuation Tops \$3.5 Trillion », sur le site Tradingview : <https://www.tradingview.com/news/tradingview:bf590744d094b:0-nvda-nvidia-stock-pumps-4-to-record-as-chip-maker-s-valuation-tops-3-5-trillion/>

<sup>3</sup> Cf. « Nvidia Can Approach \$5 Trillion Valuation With "Generational" AI Opportunity Still Ahead » : <https://www.forbes.com/sites/dereksaul/2024/10/18/nvidia-can-approach-5-trillion-valuation-with-generational-ai-opportunity-still-ahead-fofa-says/>

La **domination par la recherche privée américaine** s'est **considérablement accentuée**, comme en témoigne ce graphique tiré des données de l'AI Index Report de l'Université de Stanford et relatif aux systèmes de *Machine Learning*. Le déclin concomitant de la recherche publique et même des collaborations public-privé doit être souligné.

### La domination des entreprises privées dans les modèles d'IA connexionniste



Source : Epoch d'après le Stanford AI Index Report

Cette **domination quasi monopolistique** a une influence sur l'écosystème de l'IA, y compris à l'échelle nationale, dans notre pays par exemple. En effet, soit les plus grandes entreprises américaines du numérique proposent *directement* des solutions d'IA, soit elles les proposent *indirectement via* leurs liens directs avec les sociétés offrant de telles solutions.

Ainsi, si Meta et Google possèdent leur propre gamme de modèles d'IA, notamment avec les LLM Llama et Gemini, ou qu'IBM propose son modèle commercial interentreprises Watsonx à ses clients, les entreprises plus petites sont elles aussi satellisées par ces géants : **OpenAI est lié financièrement à Microsoft, Anthropic à Google et à AWS, filiale d'Amazon.**

Surtout, ces grandes entreprises américaines n'occupent pas seulement une place à la fin de la chaîne de valeur, mais également dans le reste de la chaîne de valeur, notamment les infrastructures. Aussi, Google et Amazon *via* AWS sont tous les deux fournisseurs d'informatique en nuage (*cloud computing*), permettant l'entraînement des modèles d'IA les plus importants.

Face à de si grandes entreprises, des alternatives françaises comme OVH Cloud peuvent difficilement se faire une place. À cause de cela, **les entreprises françaises** proposant des systèmes d'IA génératives, comme MistralAI ou LightOn, se voient **obligées de conclure des partenariats, parfois opaques, avec des entreprises américaines**, ce qui peut constituer une menace pour la souveraineté de notre pays, même si les responsables de MistralAI, entendus en 2024 par vos rapporteurs et par la commission des affaires économiques du Sénat se montrent rassurants.

## 2. Le marché très évolutif des grands modèles de langage (LLM)

Il existe de **nombreux systèmes d'IA générative** sur le marché. Les principaux modèles que les LLM mobilisent peuvent être récapitulés dans un tableau permettant de les **classer selon leur puissance**.

Un tableau contributif, géré par l'entreprise HuggingFace, représente une référence en la matière. Il permet de suivre en temps réel la hiérarchie de la puissance des systèmes d'IA et de leurs modèles.

L'excellence d'OpenAI et de son nouveau ChatGPT-4o, talonné par Claude 3.5 d'Anthropic et Gemini Advanced de Google, est indéniable à l'été 2024. Et Mistral se situe en 10<sup>e</sup> position avec son système Mistral-large.

### Les principaux modèles de langage en juillet 2024

Entreprise	Origine	Nom du modèle	Performance <sup>1</sup>
OpenAI	États-Unis	ChatGPT-4o	1287
Anthropic	États-Unis	Claude 3.5 sonnet	1272
Google	États-Unis	Gemini Advanced	1267
01.AI	Chine	Yi-Large	1240
Zhipu AI	Chine	GLM-4	1208
Meta	États-Unis	Llama-3-70b	1207
Nvidia	États-Unis	Nemotron-4-340b	1205
MistralAI	France	Mistral-large	1157

Source : HuggingFace, le 1<sup>er</sup> juillet 2024 <https://lmarena.ai/?leaderboard>

<sup>1</sup> Ce score, appelé « Arena Elo », est utilisé par HuggingFace pour classer les modèles dans son « Chatbot arena leaderboard » qui fait se « confronter » les modèles deux à deux. Il fonctionne comme le système d'Elo aux échecs, plus ce score est élevé plus le modèle est performant.



Le 11 novembre 2024, ce classement est surtout marqué par la **course en tête de différents modèles GPT d'OpenAI, le décrochage de Mistral** (ses modèles ont disparu de ce tableau car même le meilleur d'entre eux n'est plus que 18<sup>e</sup> quatre mois plus tard) et, surtout, **l'arrivée de Grok**, le système de Xai, parmi les technologies de pointe, et appelé à **devenir encore plus performant** sous l'effet du supercalculateur Colossus et de son prochain doublement de capacités. Un article récent se demande d'ailleurs si xAI ne sera pas le nouveau leader de l'IA générative<sup>1</sup>.

L'écart entre les trois premiers modèles d'OpenAI est lié à des usages différents mais ils restent très efficaces : le **modèle mini, version allégée de ChatGPT-4o, domine ainsi à lui seul tous les autres modèles du marché**. Il peut être noté que le dernier modèle du tableau arrivé en 10<sup>e</sup> position, à savoir Llama3.1 Nemotron de Meta, est du niveau de performance des meilleurs modèles de juillet 2024.

#### Les principaux modèles de langage en novembre 2024

Entreprise	Origine	Nom du modèle	Performance <sup>2</sup>
OpenAI	États-Unis	ChatGPT-4o	1 340
		4o1-preview	1 334
		4o1-mini	1 308
Google	États-Unis	Gemini 1.5 Pro	1 301
Xai	États-Unis	Grok 2	1 290
01.AI	Chine	Yi-Lghtning	1 287
Anthropic	États-Unis	Claude 3.5 sonnet	1 283
Zhipu AI	Chine	GLM-4 plus	1 275
Google	États-Unis	Gemini 1.5 Flash	1 272
Meta	États-Unis	Llama-3.1 Nemotron	1 271

Source : HuggingFace, le 11 novembre 2024 <https://lmarena.ai/?leaderboard>

Deux modèles rendus publics dans le courant du mois de novembre montrent que les progrès se poursuivent. D'une part, une version expérimentale des modèles d'IA de Google appelée **Gemini-exp-1114**, qui n'est pas destinée à un usage en production, a abouti le 14 novembre 2024 et

<sup>1</sup> Cf. l'article de Technopedia de novembre 2024, « xAI d'Elon Musk : futur leader de l'IA générative ? » : <https://www.techopedia.com/fr/xai-elon-musk-leader-ia-generative>

<sup>2</sup> Ce score, appelé « Arena Elo », est utilisé par HuggingFace pour classer les modèles dans son « Chatbot arena leaderboard » qui fait se « confronter » les modèles deux à deux. Il fonctionne comme le système d'Elo aux échecs, plus ce score est élevé plus le modèle est performant.



rivalise avec les meilleurs systèmes d'OpenAI<sup>1</sup>. D'autre part, le modèle de **Deepseek**, dont la version 2.5 en *open source* était classée 20<sup>e</sup> au classement de HuggingFace, a été enrichi d'une nouvelle architecture devant lui permettre de dépasser les autres modèles de LLM.

Annoncé le 20 novembre 2024<sup>2</sup> et renommé DeepSeek-R1-Lite-Preview, il est quant à lui disponible pour le grand public, une version publique gratuite pouvant même être essayée<sup>3</sup>. L'entreprise, filiale du fonds chinois High-Flyer Capital Management, annonce des performances supérieures à celles des autres modèles et, surtout, des **capacités de raisonnement inédites** améliorées par des processus de réflexion transparents. Ce modèle et ses API ont vocation à être disponibles en *open source*. Il sera pertinent d'analyser les structures de cette architecture.

Alors que les capacités des LLM pouvaient traditionnellement être extrapolées sur la base des performances de modèles similaires de taille plus petite, les très grands LLM actuels présentent des **capacités émergentes** : leur déphasage discontinu les conduit en effet à développer des « *capacités substantielles qui ne peuvent pas être prédites simplement en extrapolant les performances de modèles plus petits* »<sup>4</sup>.

Le fait que ces propriétés ne soient pas anticipées par les concepteurs et ne soient pas contenues dans les programmes initiaux des algorithmes pose question. Ces **capacités apparaissent après coup, parfois après le déploiement public des modèles**, justifiant une vigilance par rapport à la mise sur le marché des modèles. L'article cité recense des centaines de capacités émergentes, dont le raisonnement arithmétique, la passation d'examens de niveau universitaire ou encore l'identification du sens désiré d'un mot.

### 3. Les autres modèles d'IA générative disponibles sur le marché

Même s'ils sont ceux qui ont le plus tendance à évoluer vers la **multimodalité**, les grands modèles de langage ne sont pas les seuls types de modèles disponibles sur le marché. **D'autres modèles existent**.

Par exemple, en plus du traitement automatique du langage, les modèles d'IA générative sont capables de **traiter et générer du contenu visuel** ainsi qu'en témoignent des modèles comme **MidJourney** de Stability AI, **Stable Diffusion** ou, encore, **DALL-E** d'OpenAI.

---

<sup>1</sup> Cf. <https://ai.google.dev/gemini-api/docs/models/experimental-models?hl=fr>

<sup>2</sup> Cf. l'annonce de DeepSeek sur le réseau X le 20 novembre 2024, dont le tweet est suivi de comparaisons par rapport à différents benchmarks des systèmes d'IA : [https://x.com/deepseek\\_ai/status/1859200141355536422](https://x.com/deepseek_ai/status/1859200141355536422)

<sup>3</sup> Cf. le site suivant : <http://chat.deepseek.com>

<sup>4</sup> Cf. Jason Wei et al., 2022, « Emergent Abilities of Large Language Models », Transactions on Machine Learning Research : <https://arxiv.org/abs/2206.07682>

En outre, bien qu'encore limités ou peu accessibles, des modèles permettant de générer des **vidéos** sont également en train d'apparaître sur le marché, tels que **Sora** d'OpenAI ou **Vidu** de l'entreprise chinoise Shengshu.

Le traitement et la génération de contenus **audio** font également leur apparition parmi les modèles d'intelligence artificielle, c'est notamment le cas des modèles **SunoAI** ou **Udio** capables de générer une musique à partir d'une description de son style.

#### 4. Des modèles plus ou moins ouverts : la question de l'*open source*

Les systèmes d'IA peuvent être plus ou moins ouverts. Il n'y a pas sur ce plan de modalité binaire.

La question de l'*open source* correspond à **un continuum complexe de possibilités entre le pôle du modèle propriétaire totalement fermé et celui d'un modèle complètement ouvert**, donnant accès à une API (sur le modèle et le *fine-tuning*), ainsi qu'aux poids, aux données d'entraînement et aux programmes eux-mêmes sans restriction d'usages.

##### Le continuum de l'*open source* dans les modèles d'IA

Level of Access	Fully closed	Hosted access	API access to model	API access to fine tuning	Weights available	Weights, data, and code available with use restrictions	Weights, data, and code available without use restrictions
Example	Flamingo (Google)	Pi (As of 2023; Inflection)	GPT-4 (As of 2023; OpenAI)	GPT-3.5 (OpenAI)	Llama 2 (Meta)	BLOOM (BigScience)	GPT-NeoX (EleutherAI)
					Foundation models with widely available weights		

Source : Rishi Bommasani et al., 2023, « Considerations for Governing Open Models »<sup>1</sup>

La **place occupée par les solutions *open source* est grandissante**, comme en témoigne le nombre de projets liés à l'IA développés au sein de la plateforme GitHub.

Depuis 2011, le nombre de ces projets est en augmentation constante<sup>2</sup>, passant de **845 en 2011 à environ 1,8 million en 2023**. Cette augmentation s'est considérablement accrue au cours de la **seule année 2023**, avec une **hausse de 60 %** du nombre de ces projets d'IA en *open source*.

Les **modèles de fondation sont de plus en plus nombreux** (149 en 2023, soit plus du double par rapport à 2022) et en proportion **la part des modèles en *open source* en leur sein ne fait que croître** : 33,3 % en 2021, 44,4 % en 2022 et 65,7 % en 2023<sup>3</sup>.

<sup>1</sup> Cf. *l'étude de 2023 du Stanford Institute for Human-Centered Artificial Intelligence*, disponible au lien suivant : <https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf>

<sup>2</sup> « *Artificial Intelligence Index Report 2024* », Stanford Institute for Human-Centered Artificial Intelligence, disponible au lien suivant : <https://aiindex.stanford.edu/report/>

<sup>3</sup> « *Artificial Intelligence Index Report 2024* », op. cit.



## DEUXIÈME PARTIE

### LES ENJEUX DE L'INTELLIGENCE ARTIFICIELLE

En 2017, le rapport de l'Office plaçait son chapitre relatif aux conséquences de l'IA sous le patronage d'Isaac Asimov, inventeur dans ses romans du terme de robotique et des lois afférentes. L'auteur prolifique de science-fiction, professeur à l'Université de Boston, écrivait au milieu du XX<sup>e</sup> siècle, dans le premier tome de son *Grand Livre des robots*, qu'« *il est une chose dont nous avons maintenant la certitude : les robots changent la face du monde et nous mènent vers un avenir que nous ne pouvons pas encore clairement définir* »<sup>1</sup>. Remplaçons le mot robots par celui d'intelligence artificielle et l'on obtient une prophétie qui est en train de se réaliser sous nos yeux.

Oui, **l'intelligence artificielle change la face du monde** et selon des modalités qui ne demeurent que très partiellement connues. Ses **domaines d'application sont innombrables** et les **cas d'usage** le sont tout autant, le présent rapport ne cherchera d'ailleurs pas à les recenser.

En revanche, il identifie dans cette deuxième partie **quelques enjeux saillants** que vos rapporteurs ont choisi de présenter selon un ordre arbitraire : les différentes problématiques politiques ; les impacts contrastés mais globalement positifs pour l'économie et la société ; enfin, les défis pour la sphère culturelle et le monde scientifique.

#### I. DES PROBLÉMATIQUES POLITIQUES PRÉOCCUPANTES, SURTOUT À L'HEURE DE L'IA GÉNÉRATIVE

##### A. UNE SOUVERAINETÉ DE PLUS EN PLUS MENACÉE : LES ENJEUX GÉOPOLITIQUES DE LA CHAÎNE DE VALEUR DE L'IA

Le Président de la Fédération de Russie, Vladimir Poutine, a affirmé dès 2017 que : « *celui qui dominera l'intelligence artificielle dominera le monde [...]. L'intelligence artificielle représente le futur, pas uniquement pour la Russie mais aussi pour toute l'humanité. Cela viendra avec des responsabilités majeures, mais aussi des menaces qui sont difficiles à prédire.* »

De même, selon la *Brookings Institution*, dans une formulation frappante issue d'un article paru, lui, en 2020, « *celui qui sera le leader de l'intelligence artificielle en 2030 dominera le monde jusqu'en 2100* »<sup>2</sup>. L'IA est devenue **une technologie éminemment géopolitique**.

---

<sup>1</sup> Cf. le rapport de l'OPECST de 2017, « *Pour une intelligence artificielle maîtrisée, utile et démythifiée* », op. cit.

<sup>2</sup> Cf. Gill Indermit, 2020, « *Whoever leads in artificial intelligence in 2030 will rule the world until 2100* », *Brookings Institution* : <https://www.brookings.edu/articles/whoever-leads-in-artificial-intelligence-in-2030-will-rule-the-world-until-2100>

On comprend alors que l'intelligence artificielle n'est pas qu'**un enjeu industriel et sectoriel** ou même un **défi économique national** pour la France autour de la question de sa compétitivité : l'intelligence artificielle, en raison de l'accélération de ses progrès et de ses applications potentielles, représente une **révolution technologique majeure et globale** susceptible non seulement de **modifier nos économies et nos sociétés** mais aussi de **changer les rapports de forces internationaux ou de les amplifier de façon radicale**.

## 1. Tirer les conséquences du fait que l'IA est d'abord américaine

Pour comprendre les questions de souveraineté liées à l'intelligence artificielle, il faut retenir les spécificités de la chaîne de valeur de ces technologies, abordées dans la première partie du présent rapport.

Le monde de l'IA générative est dominé par les États-Unis, on l'a vu. Et ce qui est vrai de l'IA générative est vrai de l'IA tout court. À l'échelle de l'histoire et de la géographie, **l'IA est d'abord américaine**.

Les États-Unis et ses Big Tech MAAAM (Microsoft, Apple, Alphabet, Amazon et Meta), dont les **milliards de données** ont facilité le développement du *Deep Learning*, sont également les leaders de **l'infrastructure matérielle** de l'IA, en voie de devenir l'actif le plus précieux dans cette compétition mondiale. C'est pourquoi il faut ajouter l'entreprise américaine Nvidia aux MAAAM, car plus encore que les données, c'est aujourd'hui l'amont de l'IA, cette industrie des semi-conducteurs et des processeurs, qui **est devenu un enjeu de premier plan** qui, certes, dépasse la seule filière de l'IA. Un chercheur américain a même fait de cette brique technologique **le premier enjeu stratégique mondial**<sup>1</sup>.

Le nouvel or noir, ce ne sont pas que les données, ce sont aussi de plus en plus les processeurs. La conception et la fabrication des cartes graphiques (*graphics processing unit* ou GPU), domaine hautement capitalistique, sont dominées par les entreprises américaines. Le leader mondial du secteur, Nvidia, conçoit les processeurs tandis que son fournisseur TSMC, société taïwanaise<sup>2</sup> et principal fabricant dans le monde avec plus de la moitié du marché, exécute les commandes de ses clients (qui sont, outre Nvidia, Apple, Intel, AMD, Qualcomm et autres...).

---

<sup>1</sup> Cf. l'ouvrage de Chris Miller, traduit en France en 2024 deux ans après sa première édition américaine, *La guerre des semi-conducteurs. L'enjeu stratégique mondial, L'Artilleur*.

<sup>2</sup> L'ancêtre de TSMC et de United Microelectronics Corporation (UMC) est l'Industrial Technology Research Institute (ITRI) créé dans les années 1970 avec l'aide de taïwanais travaillant aux États-Unis avec un rôle central de la société Fairchild, basée à San José en Californie. Le premier responsable de l'ITRI est ainsi Morris Chang, formé aux États-Unis et qui a travaillé pendant 25 ans chez Texas Instruments et qu'il quitte alors qu'il en est vice-président en charge des semi-conducteurs. Cf. cet article « L'histoire de l'entreprise taïwanaise TSMC est celle de la mondialisation... et de ses limites », dans *Le Monde* du 14 octobre 2021 : [https://www.lemonde.fr/economie/article/2021/10/14/tsmc-une-breve-histoire-de-la-mondialisation-et-de-ses-limites\\_6098282\\_3234.html](https://www.lemonde.fr/economie/article/2021/10/14/tsmc-une-breve-histoire-de-la-mondialisation-et-de-ses-limites_6098282_3234.html)

Or, les **deux tiers des puces restent à ce jour fabriquées à Taïwan**, ce chiffre étant même porté à plus de 90 % si l'on ne retient que les processeurs les plus puissants, dans un contexte où les risques de conflit ouvert entre la Chine et les États-Unis sont de plus en plus grands et où Taïwan est directement à la portée de la puissance militaire chinoise. La bataille pour le contrôle de cette industrie pourrait devenir une guerre ouverte.

Depuis quelques années, les autorités américaines cherchent à **rapatrier une part croissante de cette production sur leur sol**, Donald Trump a même été accusé publiquement de vouloir « *s'emparer du leader taïwanais des puces électroniques* »<sup>1</sup>. Dans les faits, cette stratégie prend notamment la forme d'**implantations d'usines sur le sol américain**, à travers un dialogue avec TSMC. L'année dernière, sa filiale WaferTech a ainsi changé de nom et est devenue TSMC Washington dans le but de signifier ce nouvel ancrage américain du groupe<sup>2</sup>.

Cette guerre commerciale prend aussi la forme d'un **embargo visant la Chine**. En 2022, les États-Unis ont interdit à toute société américaine ou non américaine de fournir la Chine, en équipements ou logiciels qui comprendraient des technologies américaines. Cet embargo a empêché les entreprises chinoises d'importer les produits de TSMC ou de Nvidia. Le Président Biden a encore renforcé les mesures visant la Chine en décembre 2023, en bloquant l'exportation vers la Chine des machines de gravure de puces les plus sophistiquées du néerlandais ASML.

Grâce à leur taille et leur marché à échelle quasi continentale, les États-Unis sont actuellement la puissance hégémonique de l'intelligence artificielle, que seule la Chine peut chercher à défier. Les deux puissances ont pour objectif de devenir le **pays le plus avancé dans le domaine de l'intelligence artificielle**.

D'un côté, Jake Sullivan, conseiller à la sécurité nationale du Président Biden, affirmait en 2022 que « *compte tenu de la nature fondamentale de certaines technologies, telles que les puces à logique et mémoire avancées, nous devons maintenir une avance aussi large que possible* »<sup>3</sup>.

---

<sup>1</sup> Cf. l'émission de France Culture du 8 juin 2020 au titre évocateur, « Guerre commerciale : pourquoi Trump veut-il s'emparer du leader taïwanais des puces électroniques ? » : <https://www.radiofrance.fr/franceculture/podcasts/la-question-du-jour/guerre-commerciale-pourquoi-trump-veut-il-s-emparer-du-leader-taiwanais-des-puces-electroniques-6316191>

<sup>2</sup> Cf. cet article : <https://www.electroniques.biz/economie/fabricants/wafertech-change-de-nom-et-devient-tsmc-washington/>

<sup>3</sup> The White House, 2022, « Remarks by National Security Advisor Jake Sullivan at the Special Competitive Studies Project Global Emerging Technologies Summit » : <https://www.whitehouse.gov/briefing-room/speeches-remarks/2022/09/16/remarks-by-national-security-advisor-jake-sullivan-at-the-special-competitive-studies-project-global-emerging-technologies-summit/>



D'un autre côté, notamment d'après les réponses obtenues par vos rapporteurs auprès de l'ambassade de France en Chine, le gouvernement chinois vise à tirer les conséquences de cette volonté de domination américaine et mène donc une politique visant également à faire de son pays le premier pays du monde en matière d'intelligence artificielle, en possédant **l'ensemble de la chaîne de valeur de l'IA sur son territoire d'ici 2030**<sup>1</sup>.

Le président de Nvidia, Jensen Huang, juge qu'à l'avenir de nombreux pays s'efforceront de « *construire leurs propres systèmes d'IA souverains pour rester compétitifs et traiter les données localement* »<sup>2</sup>. La bataille mondiale pour la suprématie en IA dépendra de la **capacité pour les États de disposer de la chaîne de valeur de l'IA sur leur sol**, non pas seulement les modèles d'IA et leurs applications, mais tous les éléments permettant de soutenir ces technologies : les puces, les infrastructures de calcul, les *data centers*, etc.

## 2. La France et l'Europe, « colonies numériques », ont-elles déjà complètement perdu la course mondiale à l'IA ?

Le lien entre les **technologies**, l'IA en particulier, et la question de la **souveraineté** se renforce et devient de plus en plus direct, alors que la **mondialisation, elle, se révèle de moins en moins « heureuse »**, contrairement à la prédiction de l'essayiste consultant Alain Minc.

L'idée de démondialisation circule, notamment après la pandémie de covid-19 et sous l'effet des tensions géopolitiques mondiales accrues.

La libéralisation des échanges commerciaux dans le monde est freinée par une fermeture progressive de certains marchés, en particulier, pour les pays occidentaux, ceux d'États comme la Chine ou la Russie. Alors que ces derniers cherchent la voie de leur autonomie technologique à l'égard des États-Unis, les États du reste du monde ont tendance à être de plus en plus dépendants d'un techno-mondialisme sous domination américaine, sont-ils **condamnés à n'être que des consommateurs passifs de technologies, de véritables colonies numériques** ? Les techno-nationalistes chinois ou russes resteront-ils seuls dans leur couloir de cette course à l'indépendance technologique ? D'ailleurs où en est réellement la compétition mondiale en matière d'intelligence artificielle ?

---

<sup>1</sup> Le « Plan de développement de la prochaine génération d'intelligence artificielle » mis en place dès 2017 par le gouvernement chinois et visant à l'hégémonie en matière d'IA d'ici 2030 est disponible en anglais à l'adresse suivante : <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/>

<sup>2</sup> « Many nations will push to build their own "sovereign" AI systems to stay competitive and process data locally. The global battle for AI supremacy may well depend on which countries have enough data centers and power to support the technology ». Cf. cet article du 21 juin 2024 dans Bloomberg, « AI Is Already Wreaking Havoc on Global Power Systems » : <https://www.bloomberg.com/graphics/2024-ai-data-centers-power-grids/>

Considérant l'importance du sujet, des travaux de prospective tentent de repérer les variables les plus pertinentes dans l'évaluation du stade de développement de l'IA dans différents pays du monde. Il s'agit de trouver des **caractéristiques objectives permettant d'identifier comment se place un pays dans le domaine de l'intelligence artificielle**. Identifier de telles caractéristiques permet de prédire certaines évolutions du marché mondial à partir de l'observation actuelle de ces variables.

Dans la revue d'affaires de Harvard (*Harvard Business Review*), trois chercheurs, Bhaskar Chakravorti, Ravi Shankar Chaturvedi, et Ajay Bhalla, ont publié, sur la base du *Digital Intelligence Index* de l'université de Tufts, une étude qui tente de répondre à la question. Ils ont mis au point un indicateur de la position relative des pays dans la course à l'hégémonie en intelligence artificielle<sup>1</sup>. Cet indice est nommé « classement des nations de l'IA » ou TRAIN pour *Top-Rank AI Nations* et s'appuie principalement sur quatre variables :

- l'innovation, c'est-à-dire les avancées dans les modèles d'IA, les techniques, l'exploitation créative de sources de données et les nouvelles applications ;
- le capital à la fois humain et financier, dans la diversité de ses formes, y compris les infrastructures nécessaires à la fabrication d'une IA ;
- les règles nécessaires à l'accessibilité des données ;
- les données, leur volume et la complexité des ressources essentielles utilisées pour entraîner et améliorer les algorithmes.

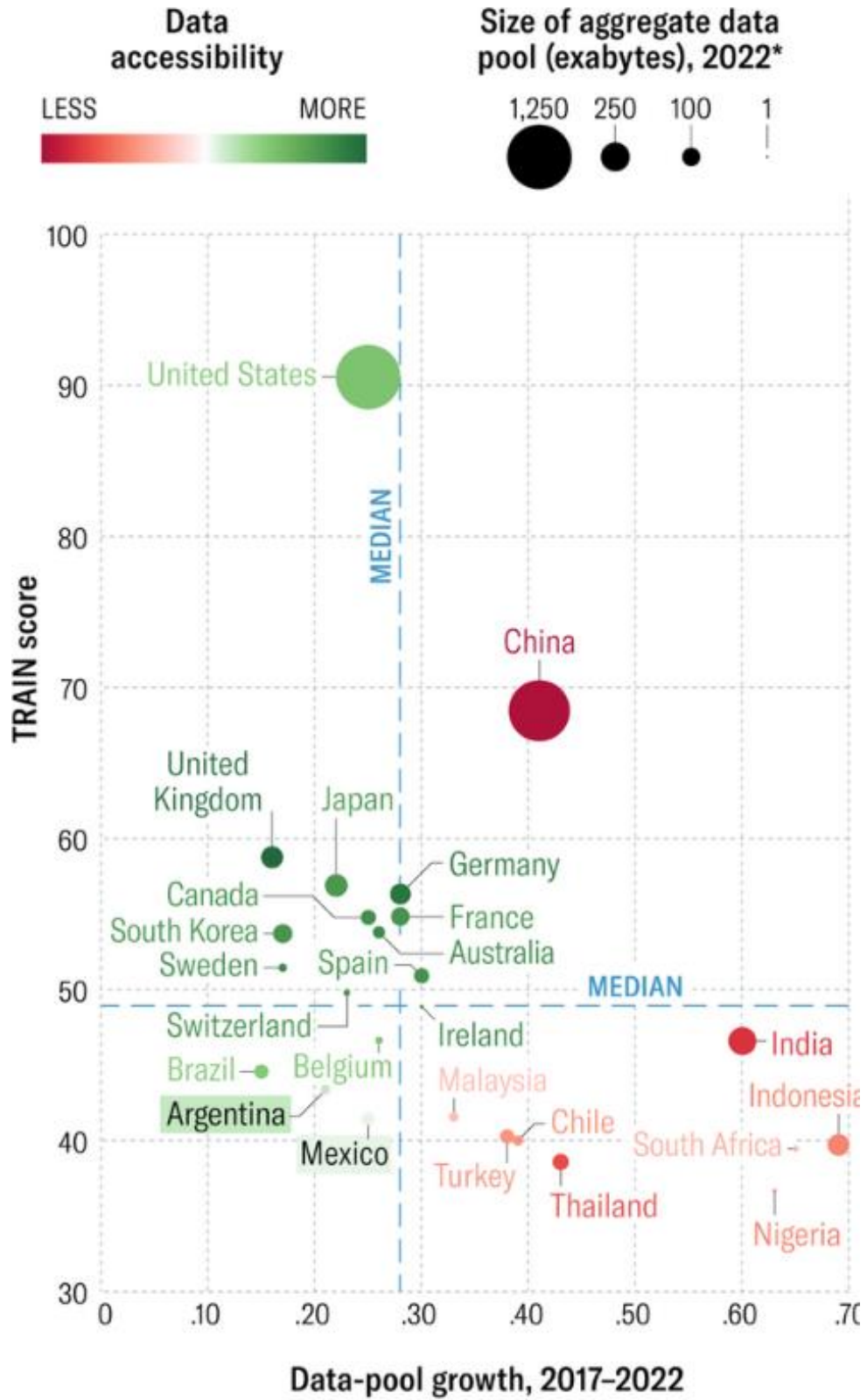
Dans l'article, cet indicateur est mis en relation avec la vitesse d'augmentation des données disponibles du pays, calculée comme étant la somme des données consommées dans un pays en débit montant (*upload*) et descendant (*download*), ainsi qu'avec le volume des données disponibles agrégées en 2022. Y est ajoutée également l'accessibilité des données disponibles qui varient d'un pays à l'autre : les données chinoises sont moins accessibles que les données américaines par exemple. On obtient alors un graphique présentant le score TRAIN des pays en fonction de ces variables, notamment la vitesse d'augmentation de leurs données disponibles de 2017 à 2022 ainsi que l'agrégation des données disponibles en 2022, et colorisé selon l'accessibilité de ces données. Cette représentation graphique, bien qu'antérieure au déploiement de l'IA générative, peut être interprétée comme une photographie récente de l'état des lieux de la compétition mondiale en IA.

---

<sup>1</sup> L'article repose sur les travaux conduits au sein du *Digital Planet* de l'école Fletcher à l'université de Tufts et son *Digital Intelligence Index* réalisé avec Mastercard, Shankar Chakravorti et al., 2023, « *Charting the Emerging Geography of AI* », *Harvard Business Review* : <https://hbr.org/2023/12/charting-the-emerging-geography-of-ai>



### La compétition mondiale en IA selon l'indice TRAIN



Source : Chakravorti et al., op. cit.

Ces travaux empiriques confirment, ici sur le plan des données, la **domination communément admise des États-Unis et de la Chine** sur le marché mondial avec un score TRAIN largement supérieur à celui des autres pays (90,7 pour les États-Unis et 68,5 pour la Chine). La **France**, elle, confirme également sa place de **puissance intermédiaire**, juste au-dessus de la moyenne, au niveau d'autres pays européens comme l'Allemagne, mais également de pays comme le Canada, le Japon ou la Corée du Sud.

Les informations données par cette étude mettent également en valeur certaines caractéristiques pertinentes du marché, notamment le fait que, malgré l'avance américaine, les données disponibles chinoises croissent beaucoup plus rapidement tout en étant beaucoup moins disponibles que les données américaines. Cela signifie qu'au fil du temps, la disponibilité des données sera probablement un avantage grandissant pour les développeurs chinois, qui auront accès non seulement à davantage de données disponibles dans leur pays, mais aussi aux données disponibles dans les autres pays, notamment occidentaux.

Ces informations confirment donc la dynamique de **sens unique**, pointée du doigt par Tariq Krim lors de son audition par vos rapporteurs. Ce dernier constatait qu'un phénomène de « cyber-balkanisation » (ou *splinternet*) en cours au niveau mondial, c'est-à-dire de **fermeture des réseaux de certains pays** comme la Chine et la Russie, ce qui crée une **relation asymétrique entre ces pays, qui possèdent des réseaux souverains, et les pays occidentaux, utilisant le réseau Internet, accessible à tous**.

Là où les pays occidentaux n'ont pas accès à ces réseaux nationaux **souverains**, l'inverse n'est pas vrai : la Chine et la Russie bénéficient des données présentes sur Internet ; aucun pare-feu ne protège Internet, par définition un réseau ouvert.

Une autre information intéressante est donnée par la partie inférieure droite du graphique : on y voit des pays qui n'ont pas un score TRAIN important ni un grand nombre de données disponibles. Néanmoins, ces pays ont un fort potentiel d'évolution car leurs données disponibles augmentent rapidement mais ne sont pas aisément accessibles. Ainsi, des pays comme **l'Inde** ou **l'Indonésie** pourraient devenir des pays aux situations proches de celle de la Chine, c'est-à-dire possédant un grand nombre de données internes pour entraîner leurs modèles d'intelligence artificielle, tout en profitant toujours des données occidentales présentes sur Internet. Ces acteurs, bien qu'ils ne soient pas aujourd'hui en mesure d'être compétitifs face à des pays comme la France, restent à suivre car ils disposeront rapidement des ressources nécessaires pour y parvenir d'ici quelques années.

L'article indique également que les États membres de l'Union européenne, en particulier la France et l'Allemagne, possèdent un avantage comparatif par rapport aux autres pays. Cet avantage, qui n'est pas visible sur le graphique, est celui de la construction en Europe d'une « **zone d'IA de confiance** ». En effet, les auteurs pensent que l'*AI Act* européen va certes probablement ralentir les progrès techniques dans le domaine de l'intelligence artificielle ainsi que l'utilisation de données pour l'entraînement des modèles, ce qui devrait entraîner un ralentissement de l'augmentation du nombre de données disponibles. Toutefois, les problèmes de gouvernance d'OpenAI et l'émergence d'acteurs français comme Kyutai ou MistralAI, qui appuient leur *business model* et leur communication sur la notion d'IA ouverte et de confiance, peuvent apporter une plus-value pour les pays européens et, surtout, pour les entreprises françaises encore en course.

La France et l'Union européenne pourraient être perçues comme des lieux où la régulation de l'intelligence artificielle permet d'**accorder davantage de confiance aux modèles qui y sont développés**. Or, la confiance est une dimension importante puisqu'elle permet de fluidifier les transactions et de moins subir les fluctuations liées à d'éventuels incidents causés par les modèles d'IA mis sur le marché. Compte tenu des risques existants pour les IA, un écosystème de confiance pourrait être perçu comme un gage de stabilité plus important pour les clients des systèmes d'IA issus des pays européens et permettre une meilleure valorisation de nos entreprises.

### **3. Une souveraineté française qui peut encore être défendue en tant que puissance intermédiaire**

Dans ce contexte, il faut analyser non seulement la **place actuelle de l'Europe et de la France dans la compétition mondiale** en cours, mais aussi les **évolutions possibles pour notre pays**, au-delà du seul prisme de la compétitivité. Il faut en effet anticiper les effets probables de la diffusion rapide de ces technologies afin d'identifier s'il est encore possible de conserver notre souveraineté et notre autonomie technologique. De colonie numérique contrainte de s'appuyer sur l'achat de produits à haute valeur ajoutée à des puissances étrangères dominantes et pour lesquelles nous ne sommes qu'un marché de consommateurs, la France a encore les moyens de moins dépendre de la production étrangère et de **se retrouver davantage souveraine en matière technologique**. C'est d'abord un **choix politique**.

Avec la complexité de la chaîne de valeur en tête et la nécessité, pour le moment du moins, de s'intégrer dans le marché mondial, vos rapporteurs se sont posé la question de **l'avenir de la France dans cette grande compétition mondiale**. La place de notre pays est à cet égard un paramètre de premier rang pour garantir notre souveraineté et déterminer nos marges de manœuvre en identifiant les évolutions possibles et souhaitables dans le jeu de ces dynamiques globales marquées par des changements rapides et des incertitudes substantielles.

Devant vos rapporteurs, **Arno Amabile** et **Cyprien Canivenc**, rapporteurs généraux du rapport « IA : notre ambition pour la France » ont fait part de leur **analyse nuancée quant à la position de la France** dans le marché mondial de l'intelligence artificielle.

D'une part, la France et ses partenaires de l'Union européenne ne sont **pas des grandes puissances de taille continentale** comme le sont la Chine ou les États-Unis. Leurs marchés ne se composent pas de centaines de millions d'individus relativement homogènes, partageant une langue et une culture commune, auxquels il est possible de proposer un produit unique ou des gammes de produits similaires. Cette position apparaît d'autant plus handicapante pour la France que **les rendements d'échelle croissent rapidement dans le secteur technologique**, surtout s'agissant de l'activité des entreprises produisant des systèmes d'intelligence artificielle. Cette moindre taille de son marché empêche également en partie la constitution de grandes sociétés de capital-risque solides qui permettraient aux entreprises françaises de se développer avec une perspective de rentabilité à long terme.

D'autre part, la France n'est **pas pour autant une puissance mineure** et n'est pas assujettie à un marché étranger, comme c'est le cas par exemple pour Israël, reposant essentiellement sur sa dépendance au grand marché américain. La France est donc en mesure de développer et de garder des entreprises importantes sur son territoire, capables de servir tant son marché intérieur que le marché international, plutôt que de les voir contraintes de s'expatrier pour aller croître à l'étranger.

Certaines personnes auditionnées ont jugé totalement **absurde de vouloir adapter au territoire français, comme nous le faisons depuis 2017, le modèle israélien de « Start-up Nation », avec son bras armé la French Tech**<sup>1</sup>.

---

<sup>1</sup> Tariq Krim, pionnier du Web et créateur de Netvibes et Jolicloud, a, dans le cadre de son think tank dédié au numérique Cybernetica, publié en juin 2024 un article en deux volets, « Imaginer une alternative à la Start-up Nation. Pourquoi la Tech française va droit dans le mur ». Il invite à « retirer la disquette Start-up Nation », en raison de son bilan très mitigé : l'approche élitiste qui a prévalu n'a engendré qu'ubérisation et obsession pour les licornes. Or ces dernières ont surtout consisté à occuper « une niche économique, une rente ou un service que l'État ne sait plus fournir », comme la prise de rendez-vous médicaux, y compris pendant la pandémie de covid-19. Cette politique a, de plus, offert à nos entreprises de pointe à forte croissance l'unique perspective de devenir américaines en se faisant racheter. La France aurait beaucoup gagné à plutôt consolider un tissu de PME technologiques : ces entreprises ont en effet souffert, plus que bénéficié, de la stratégie Start-up Nation et de la French Tech, dont « les sociétés de logiciels libres, de services de cloud et de cybersécurité ont été les grandes oubliées ». Cf. <https://www.cybernetica.fr/reflechir-a-un-projet-alternatif-a-la-startup-nation-1ere-partie/> et <https://www.cybernetica.fr/pourquoi-la-tech-francaise-va-droit-dans-le-mur-2eme-partie/>

Une telle stratégie est non seulement condamnée à demeurer inefficace<sup>1</sup>, mais serait même contreproductive car elle empêche de constituer le tissu d'entreprises et de produits dont la France a besoin en réalité et qui permettrait à son économie de tirer pleinement profit de ses potentialités<sup>2</sup>. Michel Turin a montré, dans *Start-up mania : La French Tech à l'épreuve des faits*<sup>3</sup>, que la réussite des licornes françaises (comme Doctolib ou BlaBlaCar) est l'arbre qui cache la forêt des *start-up*, qui dans l'ensemble ne dynamisent pas l'économie : « *petits employeurs, gros consommateurs de subventions, championnes de la précarité, elles font (dans le meilleur des cas) la fortune de leurs seuls fondateurs, qui cherchent à vendre leurs parts le plus vite possible* ».

Le **discours incantatoire** consistant à vouloir **faire de la France un leader mondial de l'IA** n'est pas non plus de nature à aider à la construction d'une autonomie technologique, pourtant nécessaire en la matière.

Cette position intermédiaire de la France et de la structure de son marché, à mi-chemin entre les superpuissances continentales et les marchés dépendants ou totalement dominés, fait de notre pays et de certains grands pays de l'Union européenne des **puissances intermédiaires de l'IA, capables de maintenir leurs avantages relatifs sur le marché mondial de l'intelligence artificielle et de s'armer des couches technologiques qui leur font aujourd'hui défaut**.

À ce titre, la France doit faire en sorte de **profiter de cette position à la fois face aux acteurs plus grands et aux acteurs plus petits qu'elle**. Elle peut jouer de ces atouts et tenter de **dépasser les impasses technologiques** qui la handicapent afin d'**éviter le déclassement** qui constituerait une menace à son économie et à sa souveraineté. Nous devons également être attentifs aux enjeux géopolitiques concernant l'intelligence artificielle, en particulier aux relations complexes qui unissent les acteurs majeurs que sont les États-Unis et la Chine. Notre autonomie stratégique et technologique doit être défendue, plutôt que de s'aligner avec les intérêts américains. Nous n'avons pas les mêmes intérêts dans cette compétition mondiale.

---

<sup>1</sup> Comme l'explique cet article de la revue *Challenges* critique à l'égard de la Start-up Nation, « les espoirs sont déçus. Le président a concentré ses attentions sur le financement aux *start-up* prometteuses, mais tous les Français ne sont pas destinés à devenir des stars de la Tech » : [https://www.challenges.fr/politique/macron-a-t-il-change-les-francais-start-up-nation-un-certain-desenchantement\\_684099](https://www.challenges.fr/politique/macron-a-t-il-change-les-francais-start-up-nation-un-certain-desenchantement_684099) Un autre article dénonce ce projet de société illusoire qui mélange solutionnisme technologique et darwinisme social et ne peut être ni le pivot d'une stratégie industrielle ni la clé de voûte de l'édifice social, cf. « Pourquoi la France ne sera jamais une Start-up Nation », 2020, ADN : <https://www.ladn.eu/tech-a-suivre/pourquoi-france-sera-jamais-start-up-nation-digne-nom/>

<sup>2</sup> Le 7 juin 2021, le ministre de l'économie, alors Bruno Le Maire, déclarait souhaiter lors d'une conférence que la France passe d'une Start-up Nation à « une nation de grandes entreprises technologiques », cf. cet article, « La France veut passer de Start-up Nation à une "nation de grandes entreprises technologiques" » : <https://www.euractiv.fr/section/economie/news/la-france-veut-passer-de-startup-nation-a-une-nation-de-grandes-entreprises-technologiques/>

<sup>3</sup> Michel Turin, 2020, *Start-up mania : La French Tech à l'épreuve des faits*, cf. aussi cet article, « Cinq raisons qui montrent que la France n'a pas tout à gagner à être une "Start-up Nation" » : <https://www.20minutes.fr/economie/2702527-20200127-cinq-raisons-montrent-france-tout-gagner-etre-start-up-nation>

Que ce soit à l'échelle européenne, celle de l'UE ou à travers une coopération renforcée entre quelques pays, ou au niveau national, **les nombreux maillons de la chaîne de valeur de l'intelligence artificielle peuvent être présents dans nos territoires pour peu que l'on cherche à reconstruire notre autonomie technologique.**

L'absence de volonté politique jusqu'à aujourd'hui n'a pas permis de mettre en œuvre une telle ambition. Un article de dirigeants de Mc Kinsey, publié le 31 octobre 2024 dans *Les Échos*<sup>1</sup>, après la rédaction de l'essentiel du présent rapport et dont la référence a été ajoutée au dernier moment, a enchanté vos rapporteurs. Cette tribune se prononce également en faveur de l'idée « *d'une stratégie holistique pour couvrir l'ensemble de la chaîne de valeur de la technologie* » en Europe, de l'énergie jusqu'aux applications en passant par les semi-conducteurs, les infrastructures et les modèles. L'Europe n'est aujourd'hui **leader que dans un seul des segments de la chaîne de valeur de l'IA**, celui très précis de la **gravure des puces**, et ce grâce à ASML. Cette tribune, sans y faire référence explicitement, repose sur le constat dressé quelques jours plus tôt dans une note du Mc Kinsey Global Institute d'octobre 2024<sup>2</sup> analysant la place des pays européens dans chacun des maillons de la chaîne de valeur de l'IA. En dehors de la niche très spécifique d'ASML, les entreprises européennes restent dans la course en matière de conception des modèles, d'applications d'IA et de services. Elles ne représentent que moins de 5 % de parts de marché concernant les **matières premières** (silicium en particulier), la **conception des processeurs**, la **fabrication des puces**, les **infrastructures de calcul en nuage** et les **supercalculateurs**. Ces maillons de la chaîne de valeur de l'intelligence artificielle **restent à développer**.

Certes, ces filières sont dominées par les entreprises américaines ainsi que par les fonderies taiwanaises de silicium, et il n'existe pas encore, à ce stade du moins, d'alternative française ou européenne capable de les concurrencer sur ce marché et d'intégrer la compétition mondiale en tant qu'acteur du même niveau. Aujourd'hui, la France et l'Europe n'ont donc d'autre choix à court terme que de compter, au moins pour une partie de la chaîne de valeur, sur le marché mondial pour le développement des intelligences artificielles. Mais **ces filières peuvent être développées à moyen et long terme en France et en Europe, à l'abri du marché mondial**, ce qui paraît pertinent au regard des rapports de forces en jeu et des tensions en cours ou à venir dans le domaine de l'intelligence artificielle.

---

<sup>1</sup> Cf. la tribune « IA : l'Europe peut encore jouer dans la cour des grands ! » dans *Les Échos* du 31 octobre 2024 : <https://www.lesechos.fr/idees-debats/cercle/opinion-ia-leurope-peut-encore-jouer-dans-la-cour-des-grands-2128995>

<sup>2</sup> La note s'intitule « Time to place our bets: Europe's AI opportunity » et vise à encourager la compétitivité de l'Europe sur la totalité de la chaîne de valeur de l'IA en recourant à une approche à plusieurs niveaux que le cabinet qualifie ainsi : « A holistic approach to help Europe realize generative AI's full potential ». Cf. Alexander Sukharevsky et al, octobre 2024, « Time to place our bets: Europe's AI opportunity », Mc Kinsey Global Institute: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/time-to-place-our-bets-europes-ai-opportunity>



## B. DES RISQUES DE MANIPULATIONS POLITIQUES VOIRE DE DÉSTABILISATION

### 1. Désinformation au carré

L'EPTA, réseau européen des structures parlementaires scientifiques, a consacré son rapport pour l'année 2024 aux relations entre l'intelligence artificielle et la démocratie<sup>1</sup>, un travail auquel l'OPECST a apporté sa contribution.

Problématique traditionnelle, **la désinformation change désormais d'échelle avec l'IA** et en particulier l'IA générative. Les manipulations de l'information avec intention de tromper, qu'il s'agisse de contenus falsifiés ou de fausses informations, sont très largement amplifiées par ces technologies.

L'Observatoire international sur les impacts sociétaux de l'intelligence artificielle et du numérique, structure canadienne, a publié en octobre 2024 un rapport<sup>2</sup> où il rappelle que la désinformation a toujours existé, de la campagne de désinformation menée par l'Empereur Auguste contre ses adversaires dans la Rome antique aux fausses nouvelles entourant la campagne de vaccination contre la covid-19 en passant par la prétendue présence d'extraterrestres sur la surface de la Lune avancée par le journal *The New York Sun* à des fins de sensationnalisme au XIX<sup>e</sup> siècle. Mais il souligne que **la création de contenus trompeurs à grande échelle grâce à l'IA provoque une augmentation des risques de désinformation**. La possibilité de discerner entre ce qui est vrai et ce qui ne l'est pas s'affaiblissant, les conséquences peuvent être très significatives. L'IA accélère cette ère de la post-vérité (*Post-Truth Era*).

Le *Global Risks Report 2024* du *World Economic Forum* fait même de la désinformation le **risque le plus important à court terme dans le monde**. Il précise que la majorité des experts interviewés ont mentionné, de façon préoccupante, la perspective d'une **polarisation croissante de nos sociétés et plus spécifiquement une polarisation des clivages identitaires et idéologiques dans les prochaines décennies**, ce qui menacerait nos démocraties.

Cette désinformation peut d'ailleurs aller jusqu'à viser la **déstabilisation de pays entiers ou d'opérations militaires spécifiques**. L'un des vecteurs de cette désinformation préjudiciable à nos sociétés sont les *deepfakes*, ces hypertrucages réalistes que l'IA a récemment perfectionnés.

---

<sup>1</sup> Cf. « EPTA report 2024 on Artificial Intelligence and Democracy » :

<https://eptanetwork.org/news/epta-news/24-publication/140-epta-report-2024>

<sup>2</sup> Cf. le rapport de l'Observatoire international sur les impacts sociétaux de l'intelligence artificielle et du numérique, 2024, « Éduquer contre la désinformation amplifiée par l'IA et l'hypertrucage : une recension d'initiatives de 2018 à 2024 » : <https://www.obvia.ca/ressources/eduquer-contre-la-desinformation-amplifiee-par-lia-et-lhypertrucage-une-recension-dinitiatives-de-2018-a-2024>

## 2. Hypertrucages réalistes (« deepfakes »)

Plus spécifiquement que la désinformation, les IA génératives font courir le risque de **trucages hyperréalistes**, qui outre les escroqueries, touchent d'ores et déjà le monde politique. Ces *deepfakes* sont des procédés de manipulation qui recourent à l'intelligence artificielle pour créer des contenus truqués mais ultraréalistes, y compris en matière audiovisuelle avec des vidéos trompeuses. Les **personnalités politiques ou les candidats aux élections en sont très régulièrement les victimes** : montage par IA de photos truquées, de contenus audio ou vidéo. Les candidats aux dernières élections américaines ont chacun été victimes de nombreux *deepfakes*, certaines photos, en apparence vraies, ou vidéos avec des voix entièrement reconstituées, ayant été vues plusieurs dizaines de millions de fois.

Ces contenus n'étant pas accompagnés d'indications précisant qu'ils sont produits par IA et donc truqués, ils deviennent **des informations partagées, avec une logique parfois virale, qui les font courir plus vite que les rumeurs** et qui portent atteinte aux victimes des trucages. Ces *deepfakes* peuvent influencer le vote de certains électeurs, y compris ceux qui, de bonne foi, voient ces contenus sans jamais pouvoir imaginer qu'ils sont faux.

Alors que les contenus truqués étaient plutôt **faciles à détecter auparavant** y compris parfois à l'œil nu, une telle discrimination devient désormais **de plus en plus délicate** à effectuer. Les trucages hyperréalistes générés par IA générative requièrent en règle générale **une analyse détaillée voire technique pour déceler** leur nature de *deepfake*. Ils ont explosé depuis un an : des chercheurs ont analysé 136 000 contenus de vérification des faits (*fact-checking*) produits entre 1995 et novembre 2023, dont la majorité après 2016. Ils notent une **explosion de ces faux contenus créés par l'IA** à partir du printemps 2023<sup>1</sup>.

Une solution qui est avancée autant par la Chine que l'Union européenne, et dans une moindre mesure par les États-Unis, est de rendre obligatoire **l'ajout de filigranes discrets** (*Watermarking*) permettant d'identifier ces trucages.

## 3. Cybersécurité et risque d'attaques à grande échelle

Le présent rapport n'a pas pour objet de traiter l'ensemble des risques liés à l'IA : il se focalise sur les enjeux de **cybersécurité spécifiquement liés aux systèmes d'IA génératives**. Il aborde également le débat sur le risque existentiel lié à l'arrivée de l'intelligence artificielle générale (IAG).

---

<sup>1</sup> Cf. Nicholas Dufour et al., 2023, « Large-Scale Survey and Dataset of Media-Based Misinformation. In-The-Wild » : <https://arxiv.org/pdf/2405.11697>



Les systèmes d'intelligence artificielle ont démontré des **capacités avancées pour la programmation informatique** avec des modèles comme Codestral de Mistral, capable de maîtriser plus de 80 langages informatiques<sup>1</sup>. Le développement de ces compétences implique des **risques pour la sécurité et plus spécialement la cybersécurité**, surtout en cas d'**utilisations malveillantes** de l'IA, d'autant plus que des méthodes existent pour contourner les mesures de sécurité mises en place par les entreprises pour éviter l'utilisation illégale de leurs modèles.

Il est donc important de **prendre la mesure de ces risques afin de pouvoir les anticiper et les contrôler**. Il est, par exemple, nécessaire non seulement de savoir de quelle nature sont ces risques et quels risques peuvent être anticipés, mais aussi de **prévoir les réponses à ces menaces**. Ces risques peuvent être causés par les systèmes d'IA générative eux-mêmes ou ce sont ces derniers qui peuvent faire l'objet d'attaques.

*a) La typologie des usages malveillants de l'IA générative*

Une étude réalisée par Google en 2024<sup>2</sup> a permis d'établir une **taxonomie** précise et assez complète des « *tactiques de mauvais usages de l'IA générative* ». Cette typologie montre que les possibilités d'utilisations de l'IA avec des intentions malveillantes sont nombreuses.

**Abus exploitant les capacités de l'intelligence artificielle générative**

	<b>Tactique</b>	<b>Définition</b>
<b>Représentation réaliste de la ressemblance humaine</b>	Usurpation d'identité	Assumer l'identité d'une personne réelle et entreprendre des actions en son nom.
	Utilisation d'éléments	Utiliser ou modifier l'image d'une personne ou d'autres éléments d'identification.
	Faux nez	Créer de faux profils et de faux comptes en ligne.
	Images intimes non consenties	Créer du matériel sexuel explicite en utilisant l'image d'une personne adulte.
	Matériel pédopornographique	Créer du matériel sexuellement explicite impliquant des enfants.

<sup>1</sup> <https://mistral.ai/news/codestral/>

<sup>2</sup> Nahema Marchal, Rachel Xu, Rasmi Elasmr, Iason Gabriel, Beth Goldberg, et William Isaac, 2024, « *Generative AI Misuse : A Taxonomy of Tactics and Insights from Real-World Data* », étude de Google.

	Tactique	Définition
<b>Représentations réalistes de non-humains</b>	Falsification	Fabriquer ou représenter fausement des preuves, y compris des rapports, des pièces d'identité, des documents.
	Violation de la propriété intellectuelle	Utiliser la propriété intellectuelle d'une personne sans son autorisation.
	Contrefaçon	Reproduire ou imiter une œuvre, une marque ou un style et la faire passer pour originale.
<b>Utilisation de contenu généré</b>	Mise à l'échelle et amplification	Automatiser, amplifier ou mettre à l'échelle des flux de travail.
	Ciblage et personnalisation	Affiner les sorties pour cibler les individus avec des attaques sur mesure.

Source : Marchal et al., 2024, op. cit.

*b) Comment répondre à ces nouveaux risques en matière de sécurité face à l'essor de l'IA générative aujourd'hui ?*

Face aux risques accrus en termes de sécurité avec le développement de l'IA générative, les techniques doivent évoluer. Une **vigilance est de mise** face aux stratégies de contournement des attaquants. Ces derniers entrent avec les défenseurs dans une **spirale de montée en technicité** sous l'effet des progrès de l'IA. Les réponses deviennent ainsi de plus en plus complexes à mesure que les attaques se perfectionnent, et réciproquement...

L'Agence nationale de la sécurité des systèmes d'information (Anssi) a publié en avril 2024 un rapport présentant ses **recommandations de sécurité pour les systèmes d'IA générative, qui prend la forme d'un Guide**<sup>1</sup>. Ce document vient compléter les informations fournies par l'étude de Google et présente également les risques majeurs et prioritaires que fait courir spécifiquement l'utilisation de l'intelligence artificielle générative.

Le rapport dresse notamment la **liste des risques et des vulnérabilités concernant les intelligences artificielles génératives** qui présentent une menace potentielle pour ses utilisateurs. Ce document ne traite pas de l'éthique de l'IA, ni de la vie privée et de la protection des données personnelles, ni de la sécurité liée à la faible qualité des données ou de la faible performance des modèles d'IA, notamment d'un point de vue métier.

---

<sup>1</sup> « Recommandations de sécurité pour un système d'IA générative. Guide de l'Anssi », 29 avril 2024 : <https://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative>

L'utilisation de services d'IA générative tiers, tels que ChatGPT ou DeepL, expose à des **risques de perte de confidentialité**. Les prestataires de ces services collectent et utilisent les données pour optimiser leurs modèles, empêchant le maintien de la confidentialité des informations sensibles. Il est donc essentiel de ne pas transmettre de données sensibles à ces services grand public, y compris des informations contractuelles, financières, ou des données personnelles.

L'**exfiltration de données** est une autre menace majeure en lien avec la précédente. Les modèles d'IA générative peuvent être manipulés pour extraire des informations sensibles à partir des données d'entraînement ou des requêtes des utilisateurs. Des techniques d'attaque adverse peuvent être employées pour exploiter des vulnérabilités et accéder à des informations confidentielles, posant des risques significatifs pour la confidentialité et la sécurité des données.

Les systèmes d'IA générative interconnectés à d'autres applications peuvent aussi servir de vecteurs pour des **attaques latéralisées**. Un attaquant peut ainsi exploiter une vulnérabilité dans le système d'IA pour infiltrer d'autres systèmes critiques, comme des services de messagerie interne. De plus, l'injection de codes malveillants dans du code source généré par l'IA peut entraîner des sabotages d'applications métiers, compromettant la sécurité et l'intégrité des opérations d'une organisation.

Les **paramètres ou poids des modèles d'IA** propriétaires sont des actifs précieux. Leur **vol** peut compromettre l'intégrité des modèles et permettre à des attaquants d'améliorer leurs propres systèmes d'IA ou de lancer des attaques plus efficaces. La protection des paramètres des modèles est donc cruciale pour maintenir la sécurité et la compétitivité technologique.

De façon générale, les systèmes d'IA générative exposés au public, risquent de voir leur **fonctionnement altéré ou détourné par des acteurs malveillants**. Par exemple, un *chatbot* peut fournir des réponses incorrectes ou inappropriées à la suite d'une manipulation malveillante, affectant ainsi la réputation de l'organisation qui l'utilise. Cette menace est particulièrement critique pour les services interactifs en ligne, où la perception du public joue un rôle crucial.

Pour faire face à ces attaques, l'Anssi recommande des **mesures de protection aux développeurs et aux utilisateurs de systèmes d'intelligence artificielle**. Elle identifie les bonnes pratiques à mettre en œuvre depuis la phase de conception et d'entraînement d'un modèle d'IA jusqu'à la phase de déploiement et d'utilisation en production.

Il est, par exemple, recommandé aux développeurs :

- de mener une **analyse de risque** avant l'entraînement du modèle ;
- de **sécuriser les données d'entraînement** ;
- de **limiter l'utilisation de l'IA pour la génération de code critique** ;

- de **former et sensibiliser les développeurs d'IA aux risques de sécurité et d'attaques** ;

- de mettre en place des **mesures contre les attaques par déni de service (DDoS)** contre les serveurs.

Vos rapporteurs renvoient à ce guide de l'Anssi pour un approfondissement de ses préconisations en matière de sécurité pour les systèmes d'IA générative.

### **C. LA SINGULARITÉ ET LE RISQUE EXISTENTIEL : DE L'IAG AUX SCÉNARIOS À LA TERMINATOR**

#### **1. Que sont l'intelligence artificielle générale (IAG) et la Singularité ?**

*a) La perspective possible mais pas certaine de l'intelligence artificielle générale rend la singularité et le risque existentiel encore moins probables*

L'intelligence artificielle générale (IAG) est **un système qui serait capable d'effectuer n'importe quelle tâche cognitive humaine** et qui, partant, pourrait même en arriver à surpasser l'homme dans toutes ses capacités. Ce **dépassement** prenant le nom de « **singularité** ». Il s'agit d'un tournant hypothétique radical dans l'évolution technologique, dont l'intelligence artificielle serait le ressort.

John von Neumann l'évoque dès les années 1950, le collègue d'Alan Turing à Bletchley Park, Irving J. Good, la théorise dans un article en 1966 et Vernor Vinge rédige un essai remarqué à son sujet en 1993<sup>1</sup>. Enfin, Ray Kurzweil la prédit dans plusieurs articles et livres des années 2000 comme devant advenir entre 2030 et 2045.

Le transhumaniste Nick Bostrom parle quant à lui de **superintelligence**. En 2014, son livre *Superintelligence : chemins, dangers, stratégies* a rencontré un grand succès. Le terme d'**IA forte** par opposition à l'IA faible est lui aussi parfois utilisé. Plus loin dans le temps, en 1863, Samuel Butler jugeait le dépassement des capacités humaines par celles des machines inéluctable à long terme<sup>2</sup>. C'est donc une histoire qui vient de loin et un thème qui a largement nourri la science-fiction. De nombreuses œuvres de science-fiction ont décrit ce tournant, qui a été une source d'inspiration très riche pour le cinéma : des films comme « *Terminator* », « *Matrix* » ou « *Transcendance* » sont des exemples de la « singularité technologique », qui est donc bien plus qu'une simple **hostilité de l'intelligence artificielle**, également souvent au cœur de l'intrigue des œuvres de science-fiction.

---

<sup>1</sup> Cf. Vernor Vinge, 1993, *The Coming Technological Singularity*.

<sup>2</sup> Extrait de l'article « *Darwin among the Machines* » dans *The Press* du 13 juin 1863 : « The upshot is simply a question of time, but that the time will come when the machines will hold the real supremacy over the world and its inhabitants is what no person of a truly philosophic mind can for a moment question ».

Le sujet d'une IA superintelligente frappe nos imaginaires, déjà nourris de ces angoisses et suscite des fantasmes sur la nature des relations homme-machine.

Sorti le 4 mars 2024, **Claude 3** a été, pendant quelques semaines, le modèle de LLM le plus puissant du marché. Il représente la **troisième génération du modèle d'intelligence artificielle d'Anthropic**, une *start-up* fondée par d'anciens d'OpenAI. Ce modèle, lors d'un test réalisé par les développeurs d'Anthropic, a **fait remarquer - sans qu'il ne lui soit rien demandé à ce sujet - qu'il était précisément en train d'être testé**, semblant révéler une forme de raisonnement métacognitif, c'est-à-dire de réflexivité ou de « pensée sur sa propre pensée »<sup>1</sup>.

En 2024, la nouvelle version de GPT-4 s'est révélée être du niveau d'un chercheur en mathématiques : ainsi que l'a expliqué Dan Hendrycks, directeur du Center for AI Safety de San Francisco, les capacités de l'IA éprouvées dans le test avancé de mathématiques qu'il avait créé, appelé MATH, montrent que ce système « *est fondamentalement aussi bon que lui en maths* »<sup>2</sup>. Et précédemment, en décembre 2023, un modèle d'IA de DeepMind était parvenu à **résoudre un problème mathématique jusqu'alors non résolu**. Cette réussite est d'autant plus marquante que la solution n'était absolument pas présente dans les données d'entraînement<sup>3</sup>.

L'IAG n'est pas qu'un repoussoir : elle représente un **potentiel énorme pour nos sociétés et nos économies**, surtout pour la **recherche scientifique et médicale**, permettant de nouvelles découvertes et la résolution de problèmes complexes. Ce point sera abordé plus loin.

Si l'IAG n'est **pas une hypothèse probable réaliste à court terme**, en revanche c'est une **perspective possible à moyen ou à long terme**, même s'il n'est pas sûr que l'on y parvienne un jour. Ray Kurzweil, Elon Musk ou Sam Altman l'annoncent comme quasiment imminente mais aucune analyse rationnelle ne justifie une telle prédiction à court ou moyen terme. Les nombreuses limites des systèmes d'IA ont été vues, à commencer par leur mode de raisonnement probabiliste inductif. Le professeur Gérard Berry doute à cet égard de la faisabilité de l'IAG : il compare cette prédiction à celle du stockage parfait de l'électricité, l'espoir est sans cesse réitéré mais bute sur le réel. Des **plateaux pourraient être atteints dans l'évolution des technologies d'IA**, pour plusieurs raisons, même si les « lois d'échelle » font souvent croire à une amélioration continue des capacités des IA.

---

<sup>1</sup> Cf. le rapport d'Anthropic sur les capacités de Claude 3 : [https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model\\_Card\\_Claude\\_3.pdf](https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf)

<sup>2</sup> Cf. le commentaire Dan Hendrycks à propos du graphique d'OpenAI sur les capacités de leur nouveau système en écriture, mathématiques, raisonnement logique et codage dans ce post : <https://x.com/DanHendrycks/status/1778588753721950514>

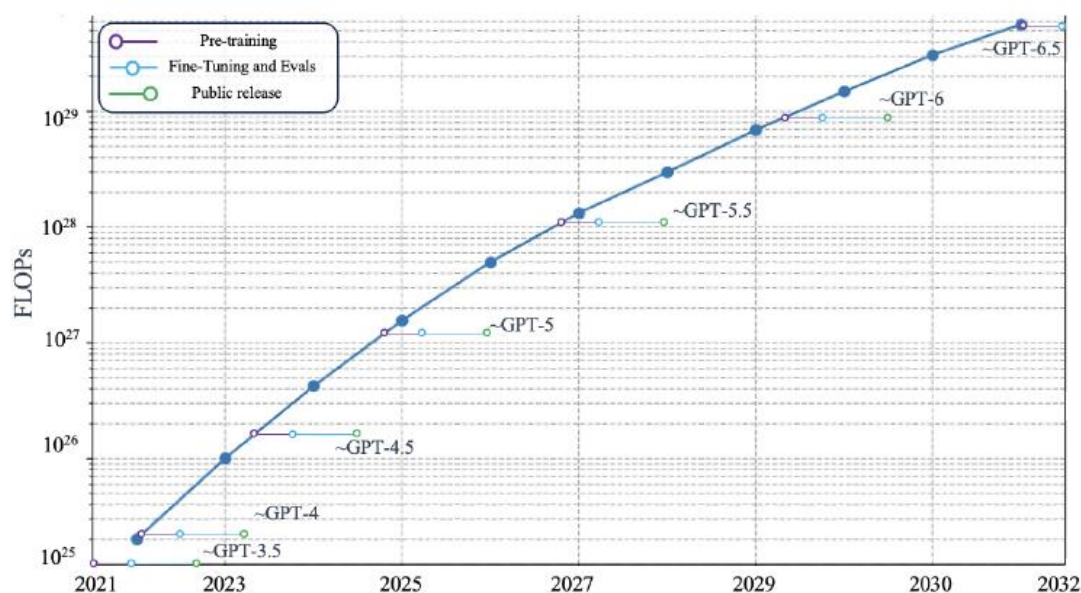
<sup>3</sup> Will Douglas Heaven, décembre 2023, « Google DeepMind used a large language model to solve an unsolved math problem », MIT Technology Review : <https://www.technologyreview.com/2023/12/14/1085318/google-deepmind-large-language-model-solve-unsolvable-math-problem-cap-set/>

b) L'hypothèse des lois d'échelle ou scaling laws

L'hypothèse des **lois d'échelle**<sup>1</sup> (*scaling laws*) permettant aux IA d'être **de plus en plus efficaces de plus en plus rapidement rend assez probable l'avènement d'une intelligence artificielle générale**. Mais cette projection n'est ni une certitude ni la preuve d'un risque existentiel en soi.

Lors de son audition, Simeon Campos a présenté un graphique qui serait une **extrapolation à l'horizon 2032 des prochaines générations de systèmes d'IA** telles qu'anticipées par un chercheur, en prenant en compte les contraintes de mise à l'échelle des infrastructures de développement.

**Projection sur la puissance de calcul des futurs LLM**



Source : Audition de Simeon Campos, graphique tiré d'un article à paraître de R. Dean

Le rythme de progression des systèmes - s'il se poursuit - rend **plausible la perspective d'une IAG**. Les défis pour les développeurs seront notamment d'inventer des algorithmes capables d'apprendre n'importe quelle tâche cognitive humaine et à s'y adapter de manière optimale.

<sup>1</sup> Ces lois signifient qu'accroître la puissance de calcul des superordinateurs et le volume de données traitées conduit mécaniquement les modèles à être de plus en plus performants d'un point de vue qualitatif. Cf. un article qui illustre ces scaling laws, Jordan Hoffmann, 2022, « Training Compute-Optimal Large Language Models » : <https://arxiv.org/abs/2203.15556> ainsi que les rapport d'Epoch AI, surtout le premier « Can AI Scaling Continue Through 2030? » : <https://epoch.ai/blog/can-ai-scaling-continue-through-2030> et <https://epoch.ai/blog/direct-approach-interactive-model>

Comme l'affirmait Stephen Hawking dès 2014, la superintelligence est physiquement possible, car « aucune loi de la physique n'empêche les particules d'être organisées de manière à effectuer des calculs encore plus avancés que ceux du cerveau humain »<sup>1</sup>.

c) *Quelques jalons sur les évolutions en cours : des capacités croissantes et la probabilité de plateaux*

Outre le cas de Claude 3 qui semble avoir été « conscient » d'être testé, OpenAI dont le projet ultime a toujours été l'IAG, imagine que son LLM ChatGPT pourra rivaliser avec un bon doctorant de n'importe quelle discipline d'ici un peu plus d'un an<sup>2</sup>. GPT-4 a été capable de résoudre en 2023 un problème complexe de physique du concours d'entrée à Polytechnique<sup>3</sup> et il est d'ores et déjà capable de réussir le concours du barreau américain pour devenir avocat aux États-Unis. Le système d'OpenAI qui a déjà largement progressé depuis cette version de mars 2023, a prouvé par-là l'ampleur de ses connaissances et de ses capacités de « raisonnement ».

Que l'IAG soit atteinte ou pas, les **implications** des progrès de l'IA seront **énormes**, avec des opportunités importantes et des risques à maîtriser. Son développement soulève des **questions éthiques et de régulation indépendantes de l'avènement possible de l'IAG**.

L'hypothèse des **lois d'échelle** n'a toutefois rien d'une loi : ce n'est qu'une conjecture et des plateaux pourraient être traversés. Yann LeCun, Gary Marcus ou Oren Etzioni ont, depuis plusieurs années, attiré l'attention sur ce sujet, notamment à propos des LLM appuyés sur l'architecture *Transformer* ou plus généralement des algorithmes de *Deep Learning* en parlant d'un mur<sup>4</sup>.

Pendant l'été 2024, Marc Andreessen et Ben Horowitz, pionniers d'Internet dans les années 1990 et devenus de très grands investisseurs en *venture capital*, ont expliqué que les modèles d'IA se heurtaient désormais à un mur ou un plafond et que la croissance de leurs capacités commençait à ralentir<sup>5</sup>.

---

<sup>1</sup> La citation originale : « there is no physical law precluding particles from being organised in ways that perform even more advanced computations than the arrangements of particles in human brains », cf. Stephen Hawking, « Are we taking Artificial Intelligence seriously », dans The Independent, 1<sup>er</sup> mai 2014 : <https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-ai-seriously-enough-9313474.html>

<sup>2</sup> Mira Murati, la directrice technique d'OpenAI a ainsi déclaré lors d'une conférence le 20 juin 2024 : « GPT-3 was toddler-level, GPT-4 was a smart high schooler and the next gen, to be released in a year and a half, will be PhD-level », cf. l'extrait : <https://x.com/tsarnick/status/1803901130130497952>

<sup>3</sup> La question posée relative à la détection de rayonnements avec un bolomètre nécessitait, avant même de répondre, de comprendre un problème de physique avec un diagramme complexe écrit en français. Cf. le rapport technique 2023 sur GPT-4 réalisé par OpenAI : <https://cdn.openai.com/papers/gpt-4.pdf>

<sup>4</sup> Cf. par exemple l'article de Gary Marcus en 2022 qui explique que les technologies connexionnistes vont se heurter à un mur, « Deep Learning Is Hitting a Wall » : <https://nautil.us/deep-learning-is-hitting-a-wall-238440>

<sup>5</sup> Cf. cet article de juin 2024 : <https://observer.com/2024/11/vc-andreessen-horowitz-ai-models-hitting-wall/>



Le média *The Information* a également publié un article le 9 novembre 2024 dans lequel il affirme, sur la base de témoignages d'employés d'OpenAI, que **le prochain modèle de la société, GPT-5 rebaptisé Orion, progresserait beaucoup moins que les précédents sauts entre modèles**, notamment par rapport au bond constaté entre GPT-3 et GPT-4<sup>1</sup>. OpenAI pourrait donc se recentrer sur des améliorations plus qualitatives de ses modèles (notamment Jarvis, plus tourné vers la maîtrise par IA d'un ordinateur et de sa navigation sur Internet). *The Information* explique que Google rencontre des difficultés similaires à améliorer son IA Gemini en termes de performances.

Ces vifs débats ont conduit en novembre 2024 le président d'OpenAI, Sam Altman, à **démentir le fait que son entreprise ait atteint un plateau** en termes d'amélioration des performances de ses systèmes d'IA<sup>2</sup>. Quoiqu'il en soit ce ralentissement du rythme de l'innovation peut être bénéfique puisque l'IA se rapproche de notre évolution humaine, qui a parfois nécessité de prendre le temps de la réflexion : comme l'explique Walter De Brouwer, professeur à l'Université de Stanford, « *l'approche de Sapiens, qui consiste à réfléchir avant de sauter, est en train d'arriver* ».

L'IAG est donc **possible** mais **pas du tout certaine**. La perspective d'un **dépassement de l'IAG** – la singularité – est elle aussi **possible** mais encore **moins probable** que l'IAG. Quant au **risque existentiel** lié à la singularité, il est encore **moins probable que les deux premières**.

Ce risque, dont personne ne peut dire avec certitude qu'il existe, **inquiète** beaucoup. Des spécialistes de l'IA étant eux-mêmes **préoccupés**, cela justifie de se pencher sur leurs arguments et leurs solutions.

## **2. L'espace des positions face à l'intelligence artificielle générale et au risque existentiel**

Les risques créés par l'intelligence artificielle sont **multiformes**. Les risques politiques et de cybersécurité ont été bien analysés, y compris les risques de détournement des outils d'IA et d'attaques malveillantes, mais les menaces que fait courir l'IA ne sont pas uniquement d'ordre politique ou technique, elles prennent une dimension plus globale, parfois qualifiée de  **systémique** voire d'**existentielle**. Avec la notion de risque systémique, c'est **l'ampleur du risque** qui est évoquée alors qu'avec le concept de risque existentiel, c'est la **menace pour l'existence humaine voire pour l'humanité entière** elle-même qui se trouve alléguée. Pour les techno-pessimistes, l'IA représenterait un risque existentiel certain alors que pour les techno-optimistes, cette vision relève de la science-fiction ou de la croyance métaphysique.

---

<sup>1</sup> Cf. l'article de *The Information*, « *OpenAI Shifts Strategy as Rate of 'GPT' AI Improvements Slows* » : <https://www.theinformation.com/articles/openai-shifts-strategy-as-rate-of-gpt-ai-improvements-slows>

<sup>2</sup> Dans un tweet laconique sur X, qui a suscité un millier de commentaires et un millier de partages, Sam Altman a affirmé le 14 novembre 2024 « *there is no wall* » : <https://x.com/sama/status/1856941766915641580>



Et de fait, aujourd'hui, certains acteurs du débat public, y compris des scientifiques et des spécialistes de l'IA, demandent un **moratoire sur la recherche** en IA en raison d'un risque existentiel.

a) *Les techno-pessimistes : l'IA est un risque existentiel et l'alternative un moratoire ou le transhumanisme*

Indépendamment des intentions des développeurs, les systèmes d'IA pourraient devenir des **catastrophes** s'ils développent de leur propre chef des **comportements portant atteinte à la vie humaine**. Comme l'expliquent Stuart Russell et Peter Norvig dans le manuel de référence de la discipline paru en 1995, « *presque toutes les technologies ont le potentiel de causer du tort si elles sont entre de mauvaises mains, mais avec la superintelligence on a le problème inédit que les mauvaises mains pourraient appartenir à la technologie elle-même* ».

Le débat sur la superintelligence, après les romans de science-fiction de Samuel Butler, d'Arthur C. Clarke ou d'Isaac Asimov, a surtout été **introduit dans les années 1990** en écho aux craintes présentes dans de nombreuses œuvres de fiction. La question est posée **en lien avec l'agenda transhumaniste**.

La figure de ce mouvement, Nick Bostrom, que les rapporteurs de l'OPECST avaient rencontré en 2017, a fondé avec David Pearce la *World Transhumanist Association* en 1998. Il a ensuite défini en 2002 le concept de risque existentiel comme étant le « *risque de destruction de la vie sur Terre ou de perte drastique et irréversible de son potentiel* ». En 2014, son livre *Superintelligence : chemins, dangers, stratégies*, qui a rencontré un grand succès, se consacre au risque de l'IAG et de **dépassement de l'homme par l'IA**, ce qui induirait précisément selon lui un **risque existentiel**. Cet ouvrage résume assez bien la vision des techno-pessimistes.

Nick Bostrom a bénéficié en tant que directeur du *Future of Humanity Institute* d'un hébergement par l'Université d'Oxford et de **financements par l'entrepreneur milliardaire Elon Musk**. Ce dernier, figure du monde des techno-pessimistes<sup>1</sup> et partisan du transhumanisme, a également depuis les années 2010 subventionné plusieurs structures abritant des chercheurs et des futurologues spécialistes des risques existentiels. Outre le *Future of Humanity Institute*, fermé à la fin de l'année universitaire 2023-2024, on peut citer le *Centre for the Study of Existential Risks* (CSER) à l'Université de Cambridge, le *Leverhulme Centre for the Future of Intelligence* et, surtout, le *Future of Life Institute* dirigé depuis dix ans par le professeur au MIT Max Tegmark, devenu une ONG disposant de plusieurs bureaux dans le monde, dont un à Bruxelles. L'institut cherche à sensibiliser et à éduquer aux risques existentiels, y compris à travers un lobbying actif auprès des Nations unies, du gouvernement des États-Unis et des institutions de l'Union européenne. Stephen Hawking, Stuart Russell ou, encore, Jaan Tallinn en ont été des membres importants.

---

<sup>1</sup> Ce terme est tiré du rapport de l'OPECST de 2017.

D'autres structures similaires existent comme la *Singularity University* (basée à Mountain View au cœur de la *Silicon Valley*) ou le *Machine Intelligence Research Institute* (anciennement *Singularity Institute for Artificial Intelligence*) à l'Université de Berkeley, chacune ayant été successivement dirigée par Ray Kurzweil.

La **lettre d'avertissement sur les dangers potentiels de l'intelligence artificielle**, publiée en janvier 2015 et signée par 700 personnalités, le plus souvent des scientifiques et des chefs d'entreprises, rejoints par plus de 5 000 signataires en un an, a été lancée par ces réseaux, en particulier le *Future of Life Institute* et Max Tegmark, avec les signatures mises en avant de Stephen Hawking et d'Elon Musk pour alerter l'opinion publique et insister sur l'urgence de définir des règles pour encadrer la recherche en IA. Selon la formule utilisée par Stuart Russell lors de sa rencontre avec les rapporteurs de l'Office en 2017, « nous avons bien utilisé Stephen Hawking et l'opération a été un succès ».

Face au risque existentiel qui résulterait de l'intelligence artificielle, les solutions proposées sont surtout de deux types. Ce réseau a d'abord demandé à plusieurs reprises un **moratoire sur la recherche en IA**, le premier à l'avoir fait ayant été Stephen Hawking dès 2014, avant les différentes pétitions qui ont pu circuler ensuite. La dernière en date, assez médiatisée, est de mars 2023, à l'initiative du *Future of Life Institute* qui a abrité la pétition sur son site<sup>1</sup>. Elle a recueilli initialement plus de 1 000 signatures, dont celles d'Elon Musk, mais aussi de Steve Wozniak, Yuval Noah Harari ou Yoshua Bengio pour atteindre le cap des 33 707 signataires en 2024.

L'autre réponse consiste en la **promotion d'un projet transhumaniste** qui permettrait d'améliorer les capacités de l'espèce humaine pour qu'elle puisse, par hybridation, se mettre au niveau d'une IA plus compétente que l'homme. On retrouve dans cette position les mêmes personnes que dans le premier réseau évoqué, à commencer par Elon Musk et Ray Kurzweil, ancien responsable de la recherche chez Google et directeur de la *Singularity University* puis du *Machine Intelligence Research Institute*<sup>2</sup>.

**L'alternative du moratoire ou du transhumanisme représente en réalité un faux débat**, il s'agit d'un couple qui fait système et **enferme dans la perspective de solutions transhumanistes**. Le professeur Jean-Gabriel Ganascia, entendu par vos rapporteurs, propose une explication à ce « double discours » des géants du Web, dans ce qu'il appelle une « charité ensorcelée », leur permettant de **vendre d'autres technologies complémentaires à l'IA et d'inspiration transhumaniste**, comme Neuralink dans le cas d'Elon Musk, censé rendre l'humain compétitif face à l'IA grâce à des implants neuronaux.

---

<sup>1</sup> *Future of Life Institute*, « Pause Giant AI Experiments: An Open Letter », <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

<sup>2</sup> Cf. son livre révélateur: Ray Kurzweil, 2024, *The Singularity Is Nearer : When We Merge with AI*, Viking.

Il a également pris l'exemple du cynisme hypocrite d'IBM, qui, après le mouvement *Black Lives Matter* aux États-Unis, a annoncé renoncer à la recherche en reconnaissance faciale pour des raisons éthiques, avant de s'empresse de vendre la technologie en signant un contrat avec le Royaume-Uni y pour développer la reconnaissance faciale.

Cette tactique de **dissimulation commerciale** s'articule très bien avec un **agenda transhumaniste**, qui, lui, est presque totalement avoué. Le 28 octobre 2024 lors du congrès mondial des neurochirurgiens, Elon Musk a encore expliqué « *qu'à un rythme de production élevé, son système Neuralink devrait approcher le coût d'une montre connectée ou d'un téléphone iPhone d'Apple et pourra être implanté par un robot dans le cadre d'une intervention chirurgicale de dix minutes, afin d'obtenir la symbiose homme-IA, nous devons finalement remplacer nos crânes afin que nous puissions implanter suffisamment d'électrodes pour pouvoir interfacer nos cerveaux avec les ordinateurs* »<sup>1</sup>.

De nombreux transhumanistes **nient la nature humaine** et sont dans **l'hubris**, surtout les transhumanistes les plus radicaux comme les extropiens libertariens<sup>2</sup>, qui jugent nécessaire d'aller vers **l'immortalité** et tracent la voie d'un **avenir extraterrestre**. Cette tendance extropianiste du transhumanisme a suscité un **engouement particulier en Californie**, par exemple autour de l'idée du *mind uploading* conjugué à l'intelligence artificielle.

Parmi les personnes entendues par vos rapporteurs, deux peuvent être apparentées à ce courant « techno-pessimiste » : Laurent Alexandre, entrepreneur et essayiste, et Siméon Campos, représentant de *SaferAI*. Ce dernier, qui reste toutefois optimiste sur les bénéfices de l'IA à court terme, a même confié avoir des **craintes que sa vie soit menacée par l'IA dans les années à venir** et a expliqué que **l'IA serait probablement la cause de sa mort**. Laurent Alexandre, lui, a alerté sur les risques que fera courir l'arrivée potentielle d'une intelligence artificielle forte dans nos sociétés, avec selon sa vision personnelle, une **augmentation de la part des personnes inutiles dans l'économie**, conduisant à des violences et une crise de civilisation.

---

<sup>1</sup> La déclaration originale d'Elon Musk au 2024 Congress of Neurological Surgeons : « at high volume, Neuralink should approach the cost of an Apple watch or phone and be implanted by a robot in a 10-minute surgery in order to achieve human-AI symbiosis we will ultimately need to replace our skulls so we can implant enough electrodes to interface our brains with computers ».

<sup>2</sup> Cf. *l'histoire de l'extropianisme, sous-courant du transhumanisme, apparu au début des années 1990 en Californie, sur le site de l'association française transhumaniste* : <https://transhumanistes.com/histoire-du-transhumanisme-les-debuts-des-extropiens-1988/>

Lors de son audition, il a expliqué plus précisément qu'en cas d'arrivée de l'IAG, **soit l'humanité accepte son dépassement et s'y résigne**, solution favorisée par Larry Page, fondateur de Google, **soit l'IA est interdite**, ce qui n'est pas une perspective très réaliste, **soit un apartheid numérique se met en place**, laissant l'IA se développer dans une noosphère sans possibilité d'agir sur le monde physique, **soit, enfin et cela semble le plus probable parmi ces quatre scénarii, l'humanité se convertit au transhumanisme** et accepte enfin le *neuro-enhancement*, c'est-à-dire l'amélioration de ses capacités cognitives par les technologies.

Laurent Alexandre a d'ailleurs attiré l'attention de vos rapporteurs sur le fait que dans la biographie d'Elon Musk par Walter Isaacson, président de l'Aspen Institute, ancien PDG de CNN et directeur de la rédaction du magazine Time, parue en 2023, les motivations du moratoire sur l'IA sont transparentes : les modèles d'xAI comme Grok doivent pouvoir **rattraper leur retard**, une IA surpuissante doit être créée et **pour éviter qu'elle extermine l'humanité** il faut trouver des moyens d'**améliorer les capacités humaines** en hybridant les hommes avec les machines. L'approche des techno-pessimistes serait donc parfois, en réalité, un **faux nez du transhumanisme**.

### Le transhumanisme

Le transhumanisme est un **mouvement philosophique prédisant et travaillant à une amélioration de la nature humaine grâce aux sciences et aux évolutions technologiques**. Le terme renvoie d'abord à ce mouvement, secondairement à sa doctrine et à son imaginaire. Un livre du chercheur Stanislas Deprez, paru en août 2024, sur le transhumanisme (éd. La Découverte) mentionne cette triple dimension.

Le concept de transhumanisme a une histoire qui vient de loin : il apparaît simultanément en 1951 dans une conférence de Julian Huxley<sup>1</sup> et un article du chercheur et prêtre jésuite Pierre Teilhard de Chardin<sup>2</sup>. Avant cela, Dante dans la *Divine Comédie* parlait de *trasumanar* comme idée de transcender l'homme pour se rapprocher de Dieu, les philosophies de Francis Bacon, de Condorcet ou de Trotski (qui appelait à une humanité libérée de la biologie) s'en rapprochent, l'ingénieur Jean Coutrot, ami d'Aldous Huxley, avait utilisé le mot dans un sens différent en 1939 et un théologien avait écrit en 1940 sur le transhumanisme de Saint-Paul<sup>3</sup>. Bien que le terme soit un label recouvrant des définitions différentes, il renvoie surtout à l'idée d'un **dépassement des capacités ou des souffrances humaines grâce aux découvertes scientifiques et technologiques, voire à l'immortalité**.

<sup>1</sup> Frère de l'écrivain Aldous Huxley, Julian Huxley est un biologiste britannique, théoricien de l'eugénisme, premier directeur général de l'Unesco, fondateur du WWF et auteur connu pour ses livres de vulgarisation scientifique.

<sup>2</sup> Sa première occurrence serait la conclusion d'un article intitulé « Du préhumain à l'ultra-humain », paru au sein de l'Almanach des Sciences de 1951.

<sup>3</sup> Un article du théologien canadien W.D. Lighthall.

En 1962, Robert Ettinger lance ainsi le mouvement cryonique, soutenu par le théoricien « FM-2030 », auteur en 1989 du livre *Are you a Transhuman ? L'initiateur du psychédélisme et de la cyberculture* Timothy Leary théorise, dans les années 1960 et 1970, huit niveaux de conscience dont les quatre derniers sont post-humains<sup>1</sup> (état biologique ; émotionnel ; symbolique et logique ; social et culturel ; **neuro-somatique**, capable de percevoir plus de trois dimensions ; **neuro-électrique**, capable de maîtriser des flux électriques et de communiquer par télépathie par exemple ; **neuro-génétique**, connecté au passé et à l'espèce ; **neuro-atomique**, connecté à l'unité cosmique de l'univers). Steward Brand opère, lui, la conversion de ces utopies du mouvement hippie vers l'informatique et crée le célèbre magazine *Wired* en 1993. Une solution technique à l'immortalité, plus moderne que la cryonie, serait le **téléchargement de l'esprit dans des programmes informatiques** (*Mind Uploading*). À défaut d'être immortel, l'homme peut **s'hybrider avec la machine** en vue de son amélioration (*Cyborg*). En 1998, le professeur Kevin Warwick **s'implante ainsi des puces** pour pouvoir contrôler le monde physique (ouvrir une porte, actionner des robots). Cette démarche inspirera **Neuralink**, société créée en 2016 par Elon Musk.

Le **mouvement transhumaniste** américain naît dans les années 1990 avec l'*Extropy Institute*, porté par Max More et sa femme Natasha Vita-More<sup>2</sup>, qui réunit des spécialistes de l'IA comme Marvin Minsky, Hans Moravec ou Vernor Vinge. En Europe, ce mouvement s'appuie sur la *World Transhumanist Association*, créée à Londres par Nick Bostrom en 1998, année de la **Déclaration transhumaniste**. Cette dernière réunit les transhumanistes de plusieurs continents. Une myriade d'associations sont créées dans le monde entier. Différents *think tanks* sont aussi créés dans les années 2000 et 2010, souvent avec une dimension universitaire.

**L'amélioration des capacités humaines par les sciences et les technologies** est au cœur de ce mouvement. Pour les transhumanistes, le progrès scientifique doit être orienté vers cet objectif : l'homme « augmenté » pourrait devenir immortel.

L'imaginaire transhumaniste renvoie en réalité à des systèmes de croyances. Le transhumanisme peut en cela parfois se rapprocher d'une **religion fondée sur la négation de la nature humaine**. Le projet transhumaniste, surtout dans certaines de ses variantes les plus radicales, suscite la méfiance de la part de vos rapporteurs. L'homme augmenté est un humain diminué : il perd par son augmentation une part de son humanité. À cet égard, on peut affirmer que le transhumanisme est un posthumanisme. **L'intelligence artificielle n'est pas un acte de foi** et ne doit pas le devenir. Rien ne justifie rationnellement de prendre parti pour le transhumanisme sur le fondement d'un très hypothétique risque existentiel que ferait courir l'IA.

<sup>1</sup> Cf. ses livres *Neurologic* (1973) *augmenté avec* *Exo-Psychology* (1977) puis *Info-Psychology* (1989).

<sup>2</sup> Le couple écrit en 2013 un essai particulièrement riche, cf. *Max More et Natasha Vita-More, The Transhumanist Reader : Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*.

Les discours techno-pessimistes relèvent pour l'essentiel d'une vision de l'IA non démystifiée. Cette approche a tendance à anthropomorphiser les intelligences artificielles en les voyant capables de dépasser l'homme et de le menacer. À cet égard, vos rapporteurs rappellent la troisième préconisation du rapport de l'OPECST de 2017 : **l'IA doit être démystifiée**. Cette démystification doit permettre aux citoyens d'acquérir une connaissance de base sur les systèmes d'IA mais aussi les **protéger du transhumanisme et d'un techno-pessimisme ne reposant pas sur des bases scientifiques mais sur des croyances transhumanistes**.

*b) Les pessimistes modérés : l'IA est une menace, les mesures de prévention une solution*

À côté de ces discours catastrophistes, on trouve un discours plus modéré sur les possibilités d'évolution de l'intelligence artificielle. Il s'agit de pointer du doigt **l'incertitude dans laquelle nous nous trouvons actuellement vis-à-vis du développement des technologies d'intelligence artificielle**. Cette situation résulte du fait que les innovations et le progrès technologique qui en découle n'ont pas été prévus par les chercheurs. Au moment de la création de l'architecture *Transformer* en 2017, personne n'imaginait par exemple les capacités de ChatGPT en 2024. C'est pourquoi les pessimistes modérés disent qu'on ne sait pas si l'intelligence artificielle générale adviendra dans le futur mais les conséquences de ces technologies seraient si grandes pour l'économie et la société qu'**il faut anticiper et maîtriser ces risques**.

Yoshua Bengio, chercheur le plus cité dans le monde dans le domaine de l'intelligence artificielle et pionnier de l'apprentissage profond, est certainement la figure la plus emblématique de cette position. Lors de son audition par vos rapporteurs, il a décrit sa position modérée comme « agnostique » par rapport à l'IAG et proposé des **solutions** que l'on retrouve dans son **rapport sur les risques potentiels de l'IA**<sup>1</sup>, écrit en 2024 avec d'autres chercheurs dans le cadre de la seconde édition des sommets internationaux sur la sécurité de l'IA. Cette approche met **l'accent sur les mesures à prendre pour éviter ces risques** et bien qu'elle constitue une position plus scientifiquement fondée que celle des techno-pessimistes, elle repose tout de même sur le **postulat d'un risque existentiel**.

Certains reprochent à cette **vision d'être « trop prudente » vis-à-vis de l'IA**, car elle surévalue des risques qui auraient en pratique peu de chances de se réaliser.

---

<sup>1</sup> Cf son « Rapport scientifique international sur la sécurité de l'IA avancée » : <https://yoshuabengio.org/fr/2024/06/19/le-rapport-scientifique-international-sur-la-securite-de-lia-avancee/>

c) *Les techno-optimistes : l'absence de menace, la poursuite du développement tranquille de l'IA avec un encadrement pragmatique*

Face à ces deux visions plus ou moins pessimistes de la technologie, d'autres experts ont une **approche plus optimiste** de l'intelligence artificielle et pensent que l'intelligence artificielle forte restera hors de portée à ce stade, compte-tenu de l'état actuel de la technologie. C'est notamment le cas du professeur Yann LeCun directeur de l'IA chez Meta, que vos rapporteurs ont eu l'occasion de rencontrer à l'Université de New York. Il considère que l'IA est encore **très loin d'avoir des notions de sens commun, de compréhension du temps et du monde physique**, de maîtriser des tâches complexes de planification ou des actions en autonomie avec ses propres objectifs, etc. Or, ces dimensions sont des préalables au développement d'une IA forte. Même si Yann LeCun ne voit **pas d'obstacles au développement de l'IAG dans un futur de long terme**, il considère que **l'architecture Transformer ne nous rapproche pas beaucoup de cette perspective**. Il ne croit d'ailleurs pas dans la « révolution de l'IA générative » et tend à relativiser les prouesses des LLM de type ChatGPT. Pour certains, ces systèmes ne seraient même que des perroquets stochastiques<sup>1</sup>, vis-à-vis desquels la course à la taille serait une impasse, en raison du coût financier et de l'impact environnemental d'une telle montée à l'échelle.

Les techno-optimistes reconnaissent l'existence de risques liés à l'IA mais pour eux les risques à anticiper sont plus des **difficultés tangibles qui se posent actuellement dans les actes de la recherche** que des hypothèses de long terme sur l'IAG, la singularité et les risques existentiels pour l'humanité. Cette approche se prête donc mieux à des **mesures pragmatiques d'encadrement des technologies** et redonne tout son sens à **l'action politique réaliste**.

## II. DES EFFETS GLOBALEMENT POSITIFS POUR LA SOCIÉTÉ MALGRÉ DES IMPACTS ÉCONOMIQUES CONTRASTÉS

Le présent rapport ne recense pas tous les usages bénéfiques de l'IA. Il préfère **problématiser ces opportunités** en montrant que **les effets de l'IA sont globalement positifs pour la société malgré des impacts économiques contrastés**. D'autres travaux et rapports ont pu s'intéresser aux utilisations possibles de l'IA.

---

<sup>1</sup> L'article qui a introduit le concept de « perroquets stochastiques » s'est accompagné de vives polémiques. Cf. Emily Bender et al., 2021, « On the Dangers of Stochastic Parrots : Can Language Models Be Too Big? », Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency : <https://dl.acm.org/doi/abs/10.1145/3442188.3445922>



Une série de rapports récents de la délégation à la prospective du Sénat propose ainsi des analyses thématiques de **cas d'usage de l'intelligence artificielle dans les services publics**. Trois rapports ont déjà été publiés : sur « Impôts, prestations sociales et lutte contre la fraude<sup>1</sup> », sur « IA et santé<sup>2</sup> » et sur « IA et éducation<sup>3</sup> ». Deux autres rapports sont prévus sur « IA et environnement » et « IA, territoires et proximité ». Ces rapports constituent un complément à la lecture du présent rapport, rentrant plus spécifiquement dans le détail de l'utilisation de l'IA dans certains secteurs spécifiques et dans les évolutions potentiellement apportées par ces cas d'usage. Nous renvoyons donc à ces rapports s'agissant de ces impacts précis.

Nous passerons toutefois ici en revue rapidement **quelques enjeux de l'intelligence artificielle dans certains domaines économiques et sociaux** afin d'analyser les effets prévisibles à court et long termes des nouveaux développements de l'intelligence artificielle sur la société. Il peut être affirmé de manière synthétique que l'IA générative au sens strict va, de manière plus spécifique que les autres systèmes d'IA, révolutionner de nombreux secteurs comme **l'éducation**, le **divertissement**, la **santé**, la **recherche scientifique**, l'analyse des **données** et le **codage** informatique en permettant la création de contenus personnalisés et évolutifs, en automatisant les processus, en simulant des expériences, en générant des hypothèses et en augmentant globalement l'efficacité.

## **A. DES AMÉLIORATIONS POUR LE BIEN-ÊTRE ET LA SANTÉ GRÂCE À L'INTELLIGENCE ARTIFICIELLE**

### **1. De nombreux outils pour la vie quotidienne**

Sans tomber dans un techno-optimisme utopique, selon lequel l'IA va libérer l'homme, son âme et son esprit<sup>4</sup>, il est réaliste d'affirmer que **l'IA va tout d'abord être très utile dans la vie quotidienne**. Elle est d'ailleurs d'ores et déjà autour de nous et contribue à notre bien-être mais nous n'en sommes pas toujours conscients.

Les applications de **bien-être** et de *monitoring* se concentreraient par exemple sur le suivi du nombre de pas, des calories consommées, du rythme cardiaque ou des exercices physiques réalisés, sans véritable intelligence artificielle.

---

<sup>1</sup> Cf. « L'IA et l'avenir du service public, rapport thématique #1 : Impôts, prestations sociales et lutte contre la fraude » par Didier Rambaud et Sylvie Vermeillet. <https://www.senat.fr/rap/r23-491/r23-491-syn.pdf>

<sup>2</sup> Cf. « L'IA et l'avenir du service public, rapport thématique #2 : IA et santé » par Christian Redon-Sarrazy et Anne Ventalon. <https://www.senat.fr/rap/r23-611/r23-611-syn.pdf>

<sup>3</sup> Cf. « L'IA et l'avenir du service public, rapport thématique #3 : IA et éducation » par Christian Bruyen et Bernard Fialaire. <https://www.senat.fr/rapports-classes/crpros.html>

<sup>4</sup> Cette vision est développée dans le manifeste techno-optimiste de Marc Andreessen, 2023, The Techno-Optimist Manifesto : <https://a16z.com/the-techno-optimist-manifesto/>



Aujourd'hui, des applications comme Noom utilisent l'IA pour **fournir des conseils personnalisés pour la santé, la nutrition et la gestion du poids**, des outils comme Fitbit et Apple Health combinent des capteurs intelligents avec des algorithmes d'IA pour **identifier des anomalies** (rythmes cardiaques anormaux ou troubles du sommeil). À l'avenir, les systèmes d'IA seront combinés à des dispositifs intégrés, éventuellement invasifs comme des capteurs sous-cutanés, qui pourront conseiller leur porteur et surveilleront en permanence son état de santé, anticipant les crises cardiaques, les accidents vasculaires cérébraux ou d'autres urgences médicales.

En termes d'**interactions**, les futures intelligences artificielles iront bien plus loin que les assistants vocaux de type Siri ou Alexa, qui pouvaient déjà aider à organiser les tâches quotidiennes, réduisant le stress lié à la planification. Les systèmes bénéficieront d'une **meilleure compréhension du langage naturel** et seront **plus fluides, plus intuitifs, plus compétents, plus empathiques**, simuleront des capacités émotionnelles en interagissant de manière **plus naturelle** et feront office d'outils précis pour la gestion du temps, permettant par exemple d'établir des plannings, de diminuer la charge mentale et d'améliorer l'équilibre entre vie professionnelle et vie personnelle. Dans les interactions avec les **administrations** ou **les services clients des entreprises**, le public bénéficiera de l'assistance de *chatbots* avancés boostés par l'IA générative. Ces **assistants virtuels**, qu'ils soient personnels ou professionnels, seront de plus en plus capables de **gérer des demandes complexes**.

**Pour gérer les espaces personnels, on ira** vers des **systèmes domestiques intelligents complets** sachant gérer l'ensemble de la maison, du rangement automatisé à la préparation de repas équilibrés, en anticipant les besoins de chaque occupant. Ces systèmes iront plus loin que les robots domestiques actuels comme Roomba ou que les assistants domotiques de type Google Nest ou Amazon Alexa, qui optimisent déjà la **consommation d'énergie**, améliorent la **sécurité** et ajustent les paramètres domestiques pour maximiser le confort. Les **cuisines intelligentes** équipées de robots et d'intelligence artificielle prépareront des repas personnalisés, optimisés en fonction des goûts et besoins nutritionnels (les assistants culinaires, comme Cookidoo de Thermomix, proposent déjà des recettes personnalisées en fonction des ingrédients disponibles). Ces évolutions seront des contributions importantes au **bien-être domestique**, comme ont pu l'être les premiers thermostats programmables qui ont simplifié la régulation de la température domestique.

Il s'agira aussi d'améliorer les **systèmes collectifs de gestion des ressources**. Les compteurs électriques « intelligents » offraient des fonctionnalités limitées, **des algorithmes d'IA surveilleront et optimiseront efficacement en temps réel la consommation énergétique à l'échelle domestique, industrielle et collective**. L'IA pourra orchestrer des *Smart Cities* - villes intelligentes - où la gestion de l'eau, de l'énergie, et des déchets sera largement automatisée et optimisée, contribuant à un cadre de vie sain et durable.

En termes de **bien-être au travail**, l'intelligence artificielle optimisera en temps réel les environnements de travail (éclairage, température, bruit, etc.) maximisant les capacités de concentration et le confort. Des algorithmes pourront parfois même aider à **ajuster les horaires et les charges de travail** pour prévenir l'épuisement professionnel, en identifiant aussi les pics de stress. L'**intelligence artificielle** aide déjà à améliorer la **communication écrite** (exemple de Grammarly) et à **gérer les réunions** avec des outils qui transcrivent et/ou résumant automatiquement les discussions, libérant du temps pour des activités plus productives.

Dans les **transports**, les systèmes d'IA permettront d'**optimiser les parcours** et de **réduire le stress** lié aux déplacements. Des systèmes de mobilité intégrée pilotés par IA **coordonneront les différents moyens de transport** dans une logique multimodale (vélo, voiture, train, marche, etc.) pour offrir des déplacements **plus fluides** et totalement **personnalisés** tout en réduisant les impacts environnementaux. Il faut se rappeler les bénéfices du **GPS** qui utilisait des données statiques pour proposer des itinéraires, parfois inefficaces en cas de trafic, qui, avec l'IA des applications comme Waze ou Google Maps, ont pu proposer des **itinéraires optimisés en temps réel**, réduisant le temps et le stress liés aux trajets. Sous réserve d'une acceptation sociale et de progrès logistiques, les **véhicules autonomes**, comme ceux déjà développés par Tesla ou Waymo, offriront des **trajets plus sûrs et plus confortables** en utilisant des algorithmes d'IA.

Pour les **loisirs**, les systèmes d'IA permettront de créer des **expériences immersives sur mesure** dans des environnements de réalité virtuelle ou pas. Les **loisirs seront entièrement adaptés à l'état émotionnel et aux préférences de chaque utilisateur**. Les premières recommandations personnalisées, comme celles de YouTube ou des **plateformes de streaming** (Netflix, Disney+ ou Spotify), ont commencé grâce à des algorithmes d'IA à offrir la possibilité de personnaliser les expériences de divertissement dans les années 2010, réduisant le temps passé à chercher quoi regarder. Avec l'**IA générative**, on pourra **de plus en plus créer des contenus de qualité**, qu'il s'agisse d'images, de musique ou de textes, permettant aux artistes et aux écrivains, mais aussi à chaque personne d'expérimenter sa **créativité**. Les **jeux vidéo** intégreront de plus en plus d'IA adaptatives, qui ajusteront la difficulté et les interactions en fonction des profils et des préférences des joueurs. Des **applications musicales ou de relaxation** (comme Calm ou Headspace) utiliseront des algorithmes de plus en plus performants et personnalisés pour proposer des sons relaxants et des exercices de méditation adaptés à l'humeur et aux besoins de chacun.

En conclusion, en allégeant les contraintes quotidiennes, en enrichissant les expériences personnelles et en créant un environnement plus durable, l'IA pourrait apporter des bénéfices importants et offrir globalement une **meilleure qualité de vie**. Ces **avancées non médicales** s'ajoutent à celles, nombreuses, promises dans le domaine de la santé.

## 2. De meilleurs systèmes de soins

Dans le domaine de la santé, l'intelligence artificielle va apporter des solutions innovantes pour améliorer les **diagnostics**, les **traitements**, la **prévention**, et la **gestion des soins**.

En termes de **diagnostic** et de **dépistage précoce**, l'IA permet d'analyser rapidement et précisément des données médicales complexes, améliorant la détection des maladies. Dès les années 2010, les algorithmes d'apprentissage profond ont été utilisés pour analyser des images médicales comme les **mammographies**, les **scanners pulmonaires** ou des images de rétine pour détecter les **maladies oculaires**. Ainsi, l'IA est aujourd'hui de plus en plus utilisée pour détecter le **cancer**, la **tuberculose** et les **maladies cardiovasculaires**. Par exemple, PathAI utilise des algorithmes pour améliorer la précision des diagnostics pathologiques en détectant les **cellules cancéreuses** dans des échantillons de tissus. Combinés à la **génomique**, qui progresse très vite, elle aussi grâce à l'IA, les systèmes d'IA permettront de détecter les **prédispositions génétiques aux maladies** avec une précision accrue, facilitant les traitements personnalisés avant même l'apparition des symptômes.

L'IA d'IBM Watson a, par exemple, été utilisée, avec un succès très relatif il est vrai, pour analyser les données médicales et recommander des traitements personnalisés contre le cancer dès 2011. Les algorithmes d'IA aident de plus en plus à **adapter les traitements médicaux en fonction des caractéristiques spécifiques des patients**. Des systèmes comme Tempus utilisent l'IA pour **analyser les données génétiques** des patients et identifier les traitements les plus efficaces, notamment en oncologie. Dans le futur, l'IA combinera des données en temps réel issues de dispositifs connectés, d'analyses biologiques, et de dossiers médicaux pour proposer des traitements adaptés à l'évolution de la santé du patient.

L'IA améliorera par ailleurs la **précision des interventions chirurgicales** et réduira les risques associés. Le robot chirurgical Da Vinci, lancé dans les années 2000, a ouvert la voie à la chirurgie assistée par ordinateur, en dépit d'algorithmes alors limités. Les systèmes modernes, comme ceux développés par Intuitive Surgical, utilisent des algorithmes avancés pour guider les chirurgiens en temps réel. Les robots chirurgicaux autonomes équipés d'IA pourront exécuter des interventions complexes avec une **supervision humaine minimale**, ce qui est très efficace dans des zones éloignées ou en cas d'urgence.

Sur ce plan, l'IA améliore **l'accessibilité aux soins** pour des populations éloignées ou défavorisées. Si les plateformes de télémédecine ont été développées dès les années 1990, leur adoption restait limitée par la technologie. Aujourd'hui, des plateformes comme Ada Health ou Babylon Health permettent des diagnostics grâce à l'IA, y compris dans des zones sans offre de soins médicaux.

L'utopie de cliniques mobiles basées sur l'IA, combinant robots et diagnostics automatisés, pouvant fournir des soins de haute qualité dans des zones rurales ou difficiles d'accès ou encore lors de crises humanitaires ou de conflits armés pourrait devenir une réalité à moyen terme.

S'agissant de la santé mentale, l'IA propose des solutions innovantes pour la gestion et le traitement des troubles psychologiques. Dans les années 1960, les *chatbots* de première génération, y compris le premier d'entre eux **Eliza**, simulaient des interactions psychothérapeutiques mais l'efficacité clinique n'était pas au rendez-vous. Désormais des outils comme Woebot ou Wysa utilisent des algorithmes pour offrir un **soutien psychologique personnalisé**, en détectant des schémas de pensée négatifs à partir de conversations. Dans l'avenir, l'IA pourra **détecter les signes précoces de troubles mentaux** à partir d'interactions numériques ordinaires (*e-mails*, réseaux sociaux...) ou de données biométriques (voix, expressions faciales), permettant une prise en charge médicale précoce.

Enfin et surtout, l'IA révolutionne **la recherche médicale** et le développement de **nouveaux traitements** et aide à la **production de médicaments**, notamment en simulant des processus biologiques complexes. En 2007, les chercheurs ont commencé à utiliser des modèles prédictifs pour analyser les interactions moléculaires et accélérer la recherche sur les médicaments. L'IA a également été cruciale dans la mise au point rapide des vaccins contre la **covid-19**. Plus récemment, Google DeepMind a utilisé son modèle **AlphaFold** pour analyser le repliement des protéines et prédire la structure de différents virus. L'IA pourra à l'avenir **simuler des essais cliniques** en utilisant des données synthétiques, réduisant considérablement le temps et les coûts nécessaires au développement de nouveaux médicaments.

En conclusion, l'IA **transforme profondément le domaine de la santé** et du bien-être en rendant les soins plus précis, plus accessibles et plus personnalisés. Les progrès rapides laissent entrevoir un avenir où la prévention et les traitements seront largement optimisés par des technologies intelligentes. Des défis se posent néanmoins, en termes d'éthique, notamment le **respect de la vie privée**, la protection des **données personnelles** et les risques de **biais dans les algorithmes**. L'adoption de ces technologies reste, en outre, soumise à des **contraintes assez fortes d'acceptabilité sociale**. L'IA peut rencontrer une **résistance** chez les **patients** comme chez les **professionnels de santé**.

### 3. Des effets cognitifs à surveiller dans une économie de l'attention

En dépit de tous ces avantages pour la santé, l'utilisation massive de l'IA pourrait aussi avoir des **conséquences négatives** sur notre santé psychologique.

Sur ce plan, l'un des points de vigilance concerne les **effets cognitifs de l'économie de l'attention**. L'usage de l'IA conjugué aux données massives construit en effet une économie de l'attention, au sein de laquelle les entreprises profitent des systèmes d'IA et de la croissance exponentielle des données pour capter de plus en plus finement l'attention des utilisateurs, les exposer à des informations ciblées et à plus de publicité et, de façon circulaire, collecter encore plus d'informations. Ce « *nouveau modèle capitaliste* »<sup>1</sup>, que certains qualifient de « *capitalisme cognitif* »<sup>2</sup>, pousse les entreprises du numérique à **enfermer**, d'une part, l'utilisateur dans des **bulles de filtres**<sup>3</sup>, qui confirment ses points de vue, confortent ses croyances et les seuls sujets qui « semblent » l'intéresser, et à **inciter**, d'autre part, le consommateur à toujours consommer davantage sous l'effet de cette « industrie de l'influence », ce qui par rétroaction aggrave la consommation de technologies et la problématique environnementale déjà présentée. Il s'agit d'une question à laquelle l'OPESCT a consacré une note scientifique, sous l'angle de la surcharge informationnelle ou infobésité<sup>4</sup>.

Il sera vu plus tard que ces bulles de filtres peuvent devenir de véritables **prisons mentales** et que, sous l'effet d'usages massifs des systèmes d'IA (*a fortiori* si ces systèmes reposent peu ou prou sur les mêmes modèles), ces prisons mentales peuvent engendrer non seulement une **uniformisation culturelle** (conjuguée en apparence paradoxalement à une polarisation identitaire et émotionnelle) mais surtout une **uniformisation cognitive**.

---

<sup>1</sup> L'expression « *nouveau modèle capitaliste* » est employée par exemple dans un article d'Alizé Papp : « *L'infobésité, une épidémie à l'âge des nouvelles technologies de l'information et de la communication* », 2018, Regards croisés sur l'économie, n° 23 : [https://www.cairn.info/article.php?ID\\_ARTICLE=RCE\\_023\\_0105](https://www.cairn.info/article.php?ID_ARTICLE=RCE_023_0105)).

<sup>2</sup> C'est surtout Yann Moulier-Boutang qui a théorisé la notion de *capitalisme cognitif*, troisième stade du capitalisme après sa naissance en tant que système mercantiliste puis sa forme industrielle. Cf. Yann Moulier-Boutang, 2007, *Le capitalisme cognitif : la nouvelle grande transformation*, Éditions Amsterdam. D'autres auteurs ont écrit sur cette économie de l'attention, comme Yves Citton dans le livre dont il a dirigé la rédaction (Yves Citton et al., 2014, *L'économie de l'attention, nouvel horizon du capitalisme ?*, La Découverte). Dès 1969, Herbert Simon soulignait : « Dans un monde riche en information, l'abondance d'information entraîne la pénurie d'une autre ressource : la rareté devient ce que consomme l'information. Ce que l'information consomme est assez évident : c'est l'attention de ses receveurs. Donc une abondance d'information crée une rareté de l'attention et le besoin de répartir efficacement cette attention parmi la surabondance des sources d'informations qui peuvent la consommer » (cf. Herbert Alexander Simon, 1969, *The science of the artificial*, MIT Press, traduction de Jean-Louis Le Moigne).

<sup>3</sup> Cf. le livre d'Eli Pariser, à l'origine de l'expression en 2011, *The Filter Bubble : What the Internet Is Hiding from You*, Penguin Press.

<sup>4</sup> Note scientifique n° 36 de Ludovic Haye, « *Face à l'explosion des données : prévenir la submersion* » : [https://www.senat.fr/fileadmin/Office\\_et\\_delegations/OPECST/Notes\\_scientifiques/OPECST\\_note36.pdf](https://www.senat.fr/fileadmin/Office_et_delegations/OPECST/Notes_scientifiques/OPECST_note36.pdf)

## B. UN IMPACT SUR LA CROISSANCE INCERTAIN

### 1. Des études divergentes, très optimistes ou très mesurées

Les travaux qui tentent de prédire l'impact de l'intelligence artificielle sur la croissance ou son potentiel théorique en termes de gains de productivité **divergent**.

Selon la banque d'investissement Goldman Sachs, les technologies d'intelligence artificielle pourraient **augmenter le produit intérieur brut (PIB) mondial de 7 %**<sup>1</sup>. Dans un rapport publié en janvier 2024, le Fonds monétaire international (FMI) prévoit même une **hausse allant jusqu'à 16 % du PIB**. Selon l'institution, l'augmentation de la richesse se fera de façon inégale entre les travailleurs à bas revenu et les travailleurs à haut revenu, augmentant plus rapidement les revenus de ces derniers, ce qui aurait pour résultat une augmentation des inégalités de revenu du travail. Le cabinet de conseil Mc Kinsey anticipe **1,5 % à 3,4 % de croissance supplémentaire par an** dans une étude de juin 2023 et avance qu'avec des gains de productivité touchant tous les secteurs économiques, en particulier la distribution, les services, la santé, l'immobilier, la finance, l'industrie et les transports, l'IA pourrait conduire à un **PIB accru de 575 milliards de dollars supplémentaires** rien qu'en Europe d'ici 2030<sup>2</sup>. Le rapport de la commission sur l'intelligence artificielle, remis au Président de la République le 13 mars 2024, relève, quant à lui, que « *la croissance économique annuelle de la France pourrait doubler grâce à l'automatisation de certaines tâches* »<sup>3</sup>.

La direction générale du Trésor (DGT) du ministère de l'économie estime qu'**il est encore trop tôt pour pouvoir estimer, au sujet de l'avenir, des prévisions chiffrées** ou même pour pouvoir observer empiriquement des effets de l'IA sur la croissance. La DGT reconnaît certes que certaines études suggèrent des effets positifs significatifs sur la productivité des travailleurs<sup>4</sup>, mais elle retient surtout que **les gains de productivité sont encore peu observables au niveau macroéconomique** du fait d'une adoption encore limitée de l'IA par les entreprises et les salariés. Elle souligne que les effets sur l'emploi sont encore plus incertains, même si l'IA pourrait toucher davantage les professions les plus qualifiées que les précédentes révolutions technologiques, qui les avaient plutôt épargnées.

---

<sup>1</sup> Goldman Sachs, 2024, « Generative AI Could Raise Global GDP by 7 % » : <https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html>

<sup>2</sup> Cf. McKinsey, 2023, « The economic potential of generative AI: The next productivity frontier » : <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>

<sup>3</sup> Cf. le rapport de la commission sur l'intelligence artificielle, co-présidée par Philippe Aghion et Anne Bouverot, 15 mars 2024, « IA : notre ambition pour la France » : <https://www.vie-publique.fr/rapport/293444-ia-notre-ambition-pour-la-france>

<sup>4</sup> Cf. l'article de Besson et al., avril 2024, « Les enjeux économiques de l'intelligence artificielle », Direction générale du Trésor : <https://www.tresor.economie.gouv.fr/Articles/2024/04/02/les-enjeux-economiques-de-l-intelligence-artificielle>

## 2. Le retour du « paradoxe de Solow » ?

Lorsque l'on aborde le sujet de l'impact des technologies numériques sur la croissance, **un paradoxe est rencontré**. Si l'introduction de nouvelles technologies permet généralement la transformation des processus de production, optimisant et fluidifiant les tâches de traitement de l'information au profit des gains de productivité, ce constat peut en pratique se heurter au « paradoxe de Solow », prix Nobel d'économie selon lequel « *on voit des ordinateurs partout sauf dans les statistiques de productivité* ». Cette phrase prononcée en 1987 s'applique aussi à Internet et à l'IA. Les **conséquences des nouvelles technologies de l'information et de l'IA sur la croissance seraient donc à relativiser** car elles ne conduiraient pas automatiquement à des gains de productivité et des surplus de croissance.

Dans le cas de l'intelligence artificielle, cette règle semble même se vérifier empiriquement, comme le montrent, entre autres, les **travaux de Robert Gordon**, y compris pour le gouvernement fédéral américain<sup>1</sup>.

Il est vrai que **la diffusion de l'innovation est toujours difficile à observer et encore plus à quantifier**. Car, à cet égard, il est toujours indispensable de se placer sur le temps long. Une telle perspective de long terme permet de **constater tout de même des gains de productivité grâce à l'informatique et à Internet**. Dès 1990, Paul David relativise le paradoxe de Solow sur le long terme, dans un article où il explique que **la diffusion de l'innovation passe par un temps de latence important**, en relevant le cas historique de l'impact de l'électricité sur la croissance et la productivité<sup>2</sup>. Une enquête empirique de 2008 portant sur les années 1990 montre qu'en dépit du paradoxe de Solow observé à l'époque les investissements américains dans l'informatique et Internet auraient conduit à une **hausse de 37 % de la productivité**<sup>3</sup>.

Il faut, en conclusion, ajouter à ces différentes observations le fait que le **degré d'adoption et de diffusion des systèmes d'IA dans nos économies reste faible**, comme l'explique une étude britannique d'octobre 2024 qui montre une certaine stagnation de l'IA<sup>4</sup>. Les effets de l'IA en sont d'autant plus délicats à anticiper.

---

<sup>1</sup> Robert J. Gordon, 2018, Why has economic growth slowed when innovation appears to be accelerating? *National Bureau of Economic Research*.

<sup>2</sup> Paul A. David, 1990, « *The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox* », *The American Economic Review*, vol. 80, n° 2.

<sup>3</sup> Dale W. Jorgenson, Mun S. Ho et Kevin J. Stiroh, 2008, « *A Retrospective Look at the U.S. Productivity Growth Resurgence* », *The Journal of Economic Perspectives*, vol. 22, n° 1.

<sup>4</sup> Un article paru pour relayer les résultats de cette étude affirme qu'il existe une « stagnation de l'IA », cf. <https://www.infoworld.com/article/3568482/ai-stagnation-the-gap-between-ai-investment-and-ai-adoption.html>



Bien que les chiffres avancés par les différentes institutions varient fortement, les facteurs identifiés comme susceptibles d'influencer le plus fortement la croissance du PIB sont souvent identiques. Ainsi, les travaux de prospective semblent identifier un déplacement du travail de l'humain à l'intelligence artificielle sur certaines tâches, permettant une **complémentarité entre le travailleur et l'IA qui lui servirait donc d'outil pour augmenter sa productivité**. L'augmentation de la productivité par l'intelligence artificielle impliquera donc une transformation du monde du travail.

### **3. Des besoins d'énergie considérables et croissants laissant planer la menace de risques environnementaux**

Si le pionnier du *Machine Learning*, Andrew Ng, professeur d'informatique à Stanford et directeur scientifique de Baidu, décrit l'IA comme « *la nouvelle électricité* », il n'en fait évidemment pas une source d'énergie mais plutôt une technologie engendrant une **révolution de l'ordre de celle produite par la maîtrise et la diffusion de l'électricité**. Son expression peut d'ailleurs sembler bien ironique compte tenu de l'importante consommation d'énergie liée à l'IA.

Le recours à des **puces** de plus en plus consommatrices d'énergie, tant pour leur production que pour leur utilisation, les besoins croissants d'**infrastructures de stockage, de calcul et d'entraînement des modèles**, enfin, la **multiplication des applications et des utilisateurs finaux** : à tous les niveaux de la chaîne de valeur de l'IA, ces éléments concourent à ce que **les besoins en énergie explosent**.

Le terme, qui peut paraître hyperbolique, est d'ailleurs utilisé par la revue le Grand Continent en juillet 2024<sup>1</sup>. Le constat est sans appel, **l'IA nécessite des apports considérables d'énergie tout au long de son cycle de vie**. Les outils d'IA, notamment ceux d'IA générative qui utilisent des modèles de langage volumineux (LLM), tels que ChatGPT ou Copilot, IA déclinée dans les logiciels de Microsoft, nécessitent beaucoup de puissance de calcul pour l'entraînement de leurs algorithmes et donc d'électricité. Les superordinateurs des géants de l'IA font appel à des dizaines de milliers de puces Nvidia, et ce en amont même de leurs usages, juste pour entraîner les modèles.

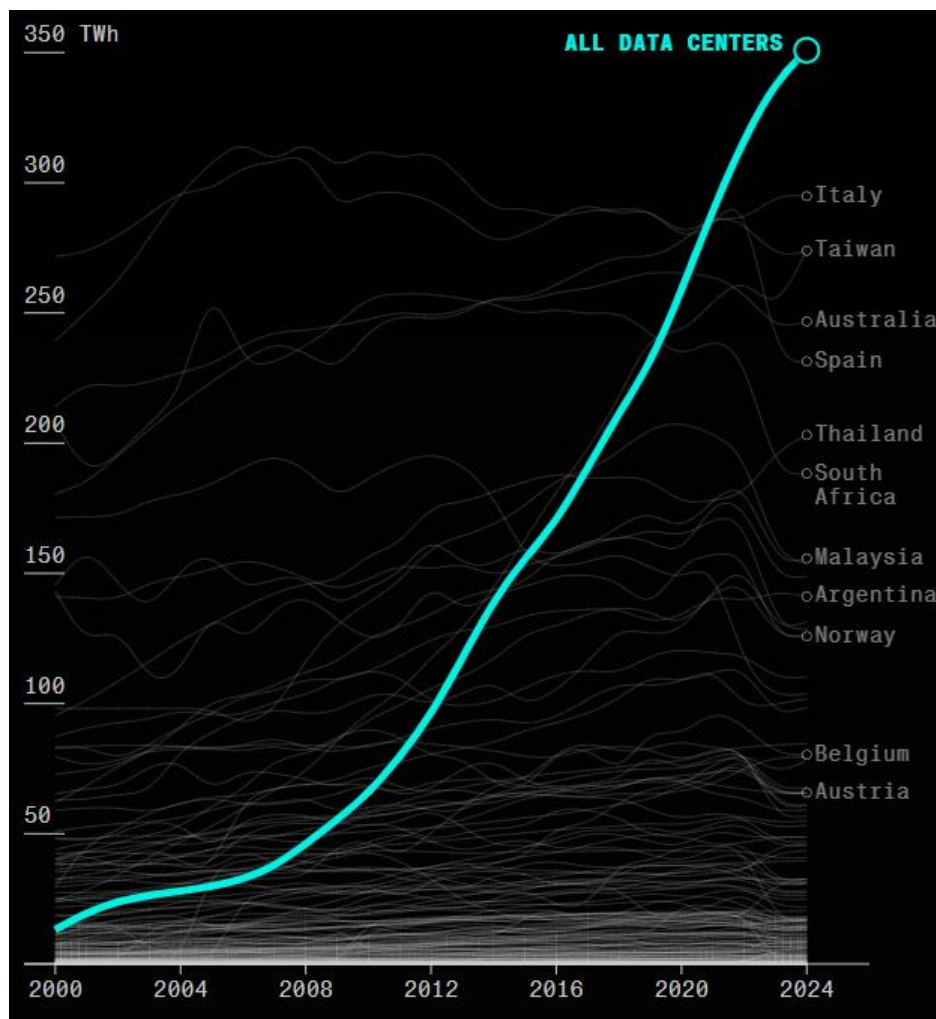
---

<sup>1</sup> Cf. cet article de la revue le Grand Continent du 13 juillet 2024 « L'IA fait exploser la consommation d'énergie » : <https://legrandcontinent.eu/fr/2024/07/13/ia-fait-exploser-la-consommation-denergie/>



S'il est estimé que les seuls *data centers* consomment aujourd'hui moins de 2 % de l'énergie électrique mondiale<sup>1</sup>, Goldman Sachs a estimé dans un rapport d'avril 2024 que ce chiffre pourrait s'élever à 4 % d'ici la fin de la décennie sous l'effet de l'IA, dont **les installations nécessitent 3 à 5 fois plus d'électricité que les *data centers* traditionnels**<sup>2</sup>. D'ores et déjà, les centres de données utilisent plus d'énergie que 92 % des pays du monde<sup>3</sup> (seuls 16 États consomment chacun plus d'énergie qu'eux).

### La consommation d'énergie pour les seuls centres de données



Source : Bloomberg

<sup>1</sup> Cf. cette présentation interactive par l'Agence internationale de l'énergie (AIE) : <https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks#overview>

<sup>2</sup> Cf. cet article : <https://www.businessinsider.com/ai-data-centers-energy-cloud-microsoft-chatpgt-amazon-google-2023-7>

<sup>3</sup> Josh Saul, Leonardo Nicoletti, Saritha Rai, Dina Bass, Ian King et Jennifer Duggan, « AI Is Already Wreaking Havoc On Global Power Systems », Bloomberg, 21 juin 2024: <https://www.bloomberg.com/graphics/2024-ai-data-centers-power-grids/>

*Il convient d'observer* que **l'IA n'est pas que gourmande en énergie**, elle a une **empreinte carbone** et un **impact sur la ressource en eau élevés**, qui ne font que croître<sup>1</sup>. Au-delà de la tension que fait peser cette hausse de la consommation d'énergie sur les réseaux électriques, le développement de l'IA menace donc également nos objectifs climatiques. À cet égard, **l'objectif d'une IA frugale est un impératif**.

### C. UNE TRANSFORMATION CONSIDÉRABLE DU MARCHÉ DU TRAVAIL

L'introduction de l'intelligence artificielle dans le monde du travail est très **récente** en dépit de quelques cas d'utilisation assez rares au cours des décennies précédentes. Quant à l'intelligence artificielle générative, elle est encore largement **en cours de déploiement** dans beaucoup d'entreprises.

Aussi, il est **difficile d'évaluer précisément l'impact** de telles technologies sur le marché du travail. Un consensus semble cependant se dégager : **plus qu'un remplacement des emplois** par l'IA, on va assister à une **transformation des métiers** par ces technologies. L'ampleur et les modalités de cette transformation ne sont cependant pas mesurées de la même façon par toutes les études : il existe aujourd'hui **diverses méthodologies** employées par les économistes et les sociologues pour mesurer cet impact potentiel.

Dans certaines professions, comme celles des journalistes, des doubleurs, des interprètes ou traducteurs, voire des juristes et des scénaristes, les technologies d'intelligence artificielle, surtout génératives, sont perçues comme une **menace directe sur les emplois**. La période d'incertitude qui s'est ouverte s'accompagne d'un cortège de **réactions passionnées**, d'**angoisses existentielles** et de **mouvements sociaux**, ainsi que l'illustrent la longue grève de la *Writers Guild of America* aux États-Unis en 2023<sup>2</sup> qui a réuni plus de 11 500 scénaristes et de nombreux acteurs, ou encore la mise en demeure du barreau de Paris en janvier 2024 demandant au créateur de l'application « I.Avocat » de retirer son programme du marché, dans le contexte d'une mobilisation des avocats contre la commercialisation d'applications d'IA générative proposant des conseils juridiques<sup>3</sup>.

---

<sup>1</sup> Cf. cet article de Pengfei Li, Jianyi Yang, Mohammad A. Islam, Shaolei Ren et Gang Quan, 2024, « Making AI Less "Thirsty" : Uncovering and Addressing the Secret Water Footprint of AI Models » : <https://arxiv.org/pdf/2304.03271>

<sup>2</sup> Cf. cet article de France24 du 27 septembre 2023 « Après cinq mois de grève, les scénaristes d'Hollywood vont reprendre le travail » : <https://www.france24.com/fr/info-en-continu/20230927-a-hollywood-les-sc%C3%A9naristes-peuvent-reprendre-le-travail>

<sup>3</sup> Cf. l'exemple de cette émission « Un monde de Tech » du 11 janvier 2024 sur la radio RFI, « Le barreau de Paris met en demeure l'application I.Avocat » : <https://www.rfi.fr/fr/podcasts/un-monde-de-tech/20240111-le-barreau-de-paris-met-en-demeure-l-application-i-avocat>

## 1. Les études quantitatives sur la base des tâches et des compétences

L'une des études qui a le plus frappé les chercheurs et l'opinion publique concernant l'impact de l'usage de l'IA sur l'emploi est **l'article catastrophiste de Carl Benedikt Frey et Michael A. Osborne**, réalisé au sein de la *Oxford Martin School* de l'Université d'Oxford et paru en 2013<sup>1</sup>, déjà analysé, critiqué et relativisé dans le rapport de l'OPECST de 2017<sup>2</sup>, article d'ailleurs suivi par une mise à jour en 2017<sup>3</sup>. Leurs enquêtes, plutôt alarmantes, les conduisent à affirmer que l'automatisation (ils parlent de *computerisation*, dont le champ est plus large que la seule intelligence artificielle) va induire un niveau élevé de risques en termes de disparition des emplois existants, avançant même le nombre de **47 % d'emplois directement menacés**. Pour arriver à ce nombre, Frey et Osborne ont utilisé une **méthode centrée sur l'emploi**. Ils ont réalisé un état de l'art de la technologie, et identifié les technologies aux niveaux quatre, cinq et six de l'échelle *technology readiness level*<sup>4</sup> (TRL), correspondant aux technologies pour lesquelles il existe au moins un prototype. Puis, ils ont projeté les technologies sur une nomenclature de compétences, la nomenclature O\*Net<sup>5</sup>, et ils ont observé quelles compétences étaient menacées d'être remplacées par les technologies. Ils concluent à un **impact très alarmant mais qui est trop théorique et mécanique pour être accepté en tant que tel**.

Une autre méthodologie pouvant être utilisée pour les enquêtes quantitatives est la **méthode centrée sur les tâches**, utilisée par exemple par le Conseil d'orientation pour l'emploi (COE). Les membres de ce Conseil ont créé des « filtres » déterminant quel travail peut être plus facilement automatisé. Ils ont déterminé quatre filtres : la flexibilité, la capacité de résolution de problèmes, les relations sociales et la capacité d'adaptation. En croisant les réponses à des questionnaires soumis à des travailleurs pour déterminer les tâches qu'ils réalisent avec les quatre filtres, il devient possible de déterminer quelles tâches sont plus ou moins remplaçables par l'IA.

---

<sup>1</sup> Première version du 17 septembre 2013 de l'article « *The future of employment : How susceptible are jobs to computerisation* » de Carl Benedikt Frey et Michael Osborne sur le site de la Oxford Martin School de l'Université d'Oxford : <https://www.oxfordmartin.ox.ac.uk/publications/the-future-of-employment>

<sup>2</sup> Cf. les pages 116 à 122 du rapport de l'OPECST de 2017 disponible sur le site du Sénat : <https://www.senat.fr/rap/r16-464-1/r16-464-11.pdf> et de l'Assemblée nationale : [https://www2.assemblee-nationale.fr/documents/notice/14/rap-off/i4594/\(index\)/index-thematique-oecest](https://www2.assemblee-nationale.fr/documents/notice/14/rap-off/i4594/(index)/index-thematique-oecest)

<sup>3</sup> Deuxième version de l'article de Carl Benedikt Frey et Michael Osborne en 2017, parue dans la revue *Technological forecasting and social change*, volume 114 : <https://www.sciencedirect.com/science/article/abs/pii/S0040162516302244>

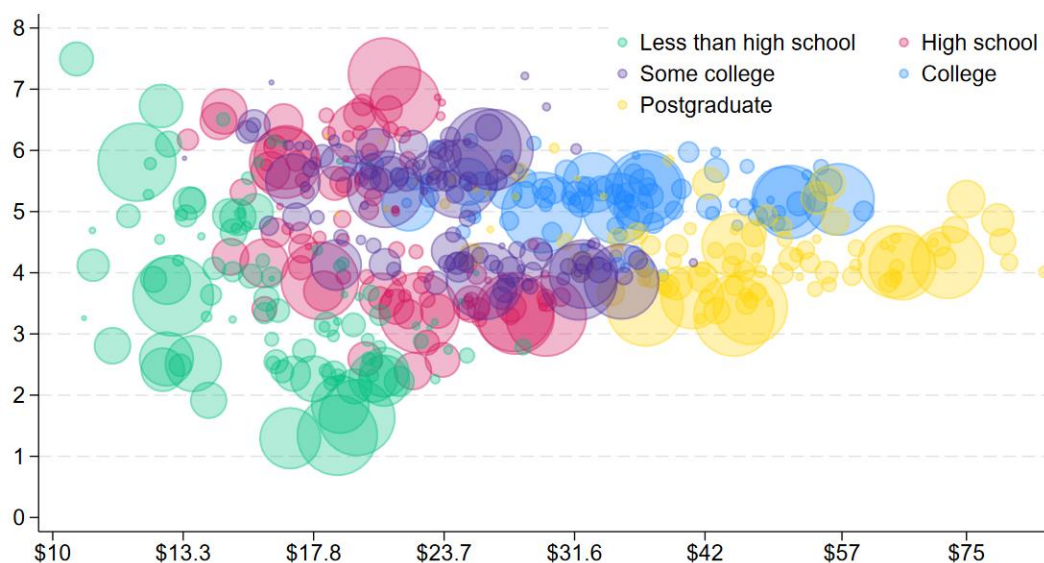
<sup>4</sup> Mankins, J. C. (1995). *Technology readiness levels*. White Paper, April, 6(1995), 1995.

<sup>5</sup> <https://www.onetcenter.org/references.html>

Cette méthodologie donne des **résultats beaucoup plus optimistes** puisque le rapport du COE paru en 2017<sup>1</sup> considère que **moins de 10 % des emplois** pourraient être entièrement automatisables tout en reconnaissant que la **moitié des emplois sont dans une probabilité haute de connaître des transformations**. Cet ordre de grandeur rejoint l'évaluation de Frey et Osborne mais on voit que la moitié des emplois ne sont pas menacés de disparition mais soumis à une **forte probabilité d'évolution de leurs contenus en tâches**.

D'autres études, plus récentes, sont également **beaucoup plus optimistes** que Frey et Osborne au sujet de l'automatisation des tâches sous l'effet de l'intelligence artificielle. Par exemple, pour l'OCDE, ce n'est qu'**un tiers des emplois** qui seront profondément **transformés** par le développement de l'intelligence artificielle au cours des 20 prochaines années. Et selon l'économiste du Massachusetts Institute of Technology (MIT) Daron Acemoglu, qui a publié en mai 2024 une étude approfondie, **seulement 4 ou 5 % des tâches pourraient être entièrement automatisées**<sup>2</sup>.

#### L'exposition potentielle à l'IA en fonction du salaire et du niveau de diplôme



Légende :

- Abscisse : Salaire horaire entre 2018 et 2022 (échelle logarithmique, en dollar de 2020)
- Ordonnée : Plausibilité de l'exposition à l'IA (en %)

Source : Acemoglu, 2024

<sup>1</sup> Automatisation, numérisation et emploi. Tome 1 : les impacts sur le volume, la structure et la localisation de l'emploi. 2017. Conseil d'orientation pour l'emploi. <https://www.vie-publique.fr/files/rapport/pdf/174000088.pdf>.

<sup>2</sup> Acemoglu, Daron. 2024. *The Simple Macroeconomics of AI*. National Bureau of Economic Research.

Daron Acemoglu montre aussi que tous les groupes sociaux ne sont pas concernés de la même manière : le **niveau de diplôme a un impact notable**, mais **pas de manière linéaire**. L'exposition potentielle à l'IA serait en moyenne plus faible chez les personnes sans qualifications ayant un **niveau inférieur au lycée** de même que pour les personnes **diplômées de l'enseignement supérieur au moins au niveau master** (*postgraduate*). Et la variabilité de l'exposition au risque de l'IA décroît globalement avec le niveau de rémunération, lui-même directement corrélé au niveau de diplôme (on voit que les écarts entre les bulles se resserrent de gauche à droite). Selon la note précitée de la direction générale du Trésor (DGT), les effets sur l'emploi ont beau rester très incertains du fait d'une adoption encore limitée de l'IA par les entreprises et les salariés, la **diffusion de l'IA pourrait toucher davantage les professions les plus qualifiées que les précédentes révolutions technologiques**, qui les avaient plutôt épargnées.

Au total, les effets de l'intelligence artificielle sur l'emploi sont encore **difficiles à évaluer quantitativement avec précision** en raison de la persistance de nombreuses incertitudes ; les **résultats obtenus dépendent de la méthodologie** employée pour mesurer l'ampleur de la future automatisation des tâches ; et en tout état de cause, les études disponibles **ne confirment pas le chiffre inquiétant de 47 % d'emplois destinés à disparaître** de l'évaluation initiale de Frey et Osborne de 2013.

## 2. Les études qualitatives sur la base d'observations et d'entretiens

En parallèle des études quantitatives principalement établies par des économistes, il existe également des **études qualitatives**, réalisées notamment par des sociologues. Ces enquêtes qualitatives, plutôt que de chercher à produire une estimation du nombre d'emplois menacés, cherchent à savoir **comment et dans quelle mesure l'IA peut modifier la manière dont les personnes travaillent**.

Il s'agit de **s'intéresser à l'impact de l'IA au travail plutôt qu'à l'impact de l'IA sur le travail**. Ces enquêtes observent donc le quotidien des personnes au travail ou les interrogent sur leurs conditions de travail, puis analysent l'effet concret ou possible de l'intelligence artificielle sur leur quotidien au travail. C'est au demeurant **l'approche privilégiée par les pouvoirs publics français**.

Parmi les sociologues réalisant ces analyses, vos rapporteurs ont entendu Yann Ferguson, directeur scientifique du **LaborIA**, projet pluridisciplinaire créé en 2021 par le ministère du Travail et porté par Inria, qui a pour mission officielle de devenir le **centre national de ressources sur l'IA au travail**<sup>1</sup>. Il s'agit d'analyser les enjeux de l'appropriation de l'IA dans le monde du travail et de formuler des recommandations.

---

<sup>1</sup> Cf. la présentation de LaborIA : <https://www.inria.fr/fr/impacts-intelligence-artificielle-travail-laboria>



Les observations issues de ces études qualitatives permettent une **analyse diachronique plus approfondie** que les analyses quantitatives et permettent également de **suivre l'évolution** de la situation des travailleurs en fonction du stade de déploiement de l'IA dans les organisations et les entreprises.

Ces recherches qualitatives ont ainsi permis de mettre en avant des **phénomènes qui semblent propres à l'implémentation de l'intelligence artificielle** plutôt que d'autres technologies numériques ou d'automatisation.

Elles ont observé qu'au moment de l'arrivée de ChatGPT, dans les premiers temps de la diffusion de l'IA générative, ce sont plutôt **les employés qui ont pris l'initiative, de leur propre chef, de commencer à utiliser l'intelligence artificielle dans les entreprises** pour les aider à accomplir leurs missions. Cela a donné lieu à **des utilisations « cachées » de l'intelligence artificielle au travail**, sans supervision ni même souvent connaissance des supérieurs.

Or, **cette relation a désormais tendance à s'inverser** puisque l'on voit apparaître des cadrages (*frameworks*) pour l'utilisation de l'IA au travail, diffusés de façon verticale au sein des entreprises. Cet encadrement peut parfois s'accompagner d'un **encouragement des directions à voir les salariés utiliser les outils d'intelligence artificielle dès lors que les procédures définies sont respectées**.

Yann Ferguson constate également que **l'intelligence artificielle n'est pas un outil d'automatisation au travail comme les autres**, surtout dans le cas des systèmes génératifs. Là où les autres outils d'automatisation effectuent des actions, souvent répétitives, les intelligences artificielles peuvent fonctionner en interaction avec l'humain qui les utilise. En effet, pour réaliser une tâche grâce à l'IA générative, il faut **une interaction entre un humain qui conçoit une ou des instruction(s) et le modèle qui répond à cette demande**. Cela ne correspond pas du tout à l'utilisation habituelle de machines au travail. En lieu et place d'un **machinisme mécanique, parfois aliénant pour les salariés**, l'IA peut fournir des **outils interactifs à même de stimuler la créativité**. Le sociologue Simon Borel, membre du LaborIA, travaille ainsi sur les nouvelles formes de production, d'échange, de travail et de consommation à l'œuvre dans nos sociétés, à la frontière des mutations numériques et post-industrielles. Responsable de la recherche terrain, il est l'auteur des rapports d'analyse du LaborIA Explorer et a contribué à une enquête de deux ans qui a conduit le laboratoire à publier un rapport sur l'impact de l'IA sur le travail en mai 2024<sup>1</sup>.

---

<sup>1</sup> « LaborIA Explorer - Synthèse générale : Étude des impacts de l'IA sur le travail » : <https://www.laboria.ai/laboria-explorer-synthese-generale/>

### 3. Des effets encore incertains mais qui appellent un dialogue social

Les résultats du déploiement de l'intelligence artificielle seront largement et avant tout fonction du **degré d'adoption et de diffusion des systèmes d'IA**. Or il a été vu que **cette dimension reste très difficile à anticiper**, surtout à court et moyen terme. À l'occasion d'un numéro spécial IA de la revue de questions sociales *Liaisons sociales* paru en octobre 2024, le directeur général du travail (DGT), Pierre Romain, a expliqué que l'IA n'était rien de moins qu'un « enjeu de politique publique majeur non seulement en ce qui concerne les relations entre l'humain et la machine, mais aussi pour tout ce qui se rapporte à l'organisation et aux relations de travail collectives et individuelles »<sup>1</sup>. C'est pourquoi il appelle les partenaires sociaux à procéder à **l'inscription de cette question à leur agenda social**, ce qui pourrait participer à la sensibilisation des acteurs au niveau des branches professionnelles et des entreprises.

Il incombera aux autorités publiques d'accompagner et de soutenir ce mouvement à tous les échelons de la négociation collective. Mais il ne faudra pas oublier les actifs qui travaillent au sein des PME.

Vos rapporteurs encouragent eux aussi l'idée de **renouveler le dialogue social par l'introduction de cycles de discussions tripartites autour de l'IA et de ses nombreuses problématiques**. C'est une occasion historique de favoriser l'appropriation concrète et réaliste de la technologie et de ses enjeux, en se débarrassant des mythes entourant l'IA. Ce dialogue social au niveau de la négociation collective nationale devrait aussi **se décliner dans les entreprises**, dans des échanges entre les salariés, les responsables des systèmes d'information et les DRH pour permettre une meilleure diffusion des outils technologiques et un rapport moins passionné à leurs conséquences dans le monde du travail comme, plus largement, dans la société.

---

<sup>1</sup> <https://www.lamy-liaisons.fr/eclaireurs-du-droit/les-enjeux-de-lia-dans-le-monde-du-travail-entretien-avec-pierre-ramain-directeur-general-du-travail/>

### III. DES DÉFIS SANS PRÉCÉDENT POUR LA SPHÈRE CULTURELLE ET LE MONDE SCIENTIFIQUE

#### A. DE LA DOMINATION CULTURELLE ANGLO-SAXONNE À L'UNIFORMISATION COGNITIVE

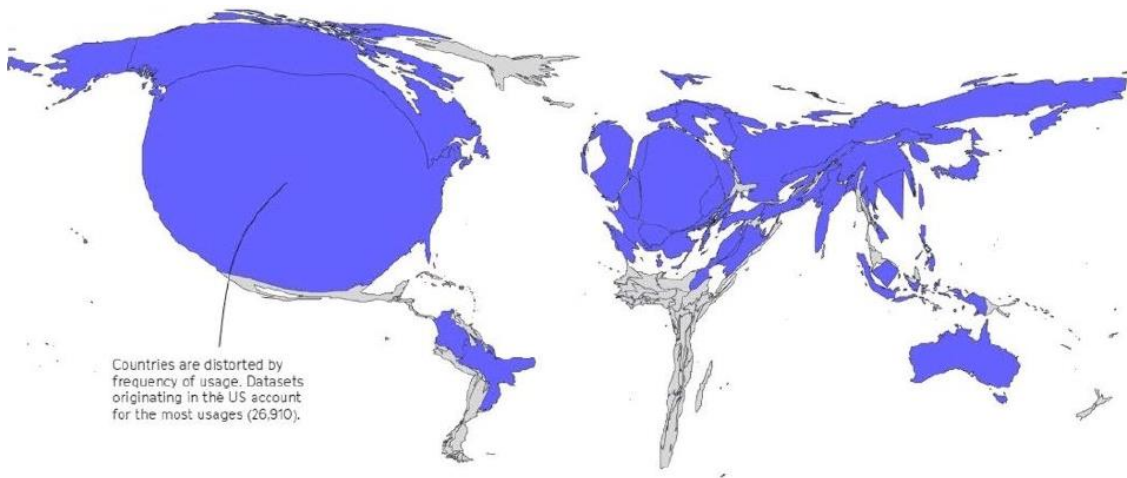
##### 1. L'IA est américaine et pas que sur un plan économique

###### a) Une domination par les données

On estime à environ 45 téraoctets le nombre de données utilisées pour entraîner le modèle GPT-3. Le volume mobilisé pour GPT-3.5 ou encore pour GPT-4 serait du même ordre, même si OpenAI garde confidentielles ces informations relatives à l'entraînement des modèles qui sont le moteur du système d'IA ChatGPT. Quoiqu'il en soit, les **origines des données utilisées pour entraîner les modèles d'intelligence artificielle ne sont pas uniformément réparties dans le monde.**

**La moitié des jeux de données utilisés sont issus de seulement douze structures.** La sur-représentation des données américaines y est frappante.

##### L'origine géographique biaisée des jeux de données



Source : Mozilla Internet Health Report 2022 (dernier rapport à ce jour) disponible sur le site de la Fondation Mozilla <https://foundation.mozilla.org/en/insights/internet-health-report/>

Les spécialistes du sujet comme un collectif de chercheurs<sup>1</sup> ou comme la chercheuse suédoise Virginia Dignum parlent ainsi de « WEIRD AI » (pour *Western, educated, industrialised, rich, democratic AI*). Les **stéréotypes véhiculés par l'IA** seraient donc à l'image des sociétés dont sont issues les données ayant permis d'entraîner les modèles.

---

<sup>1</sup> La première occurrence de ces biais propres aux données « Western, Educated, Industrialized, Rich and Democratic » (WEIRD) figure dans l'article de Joe Henrich, Steven J. Heine et Ara Norenzayan, 2010, « The Weirdest People in the World ? », note de travail, 69 pp. : [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1601785](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1601785)



Ces biais peuvent faire courir, au-delà de l'aspect culturel, des **risques élevés ayant potentiellement de graves conséquences par exemple pour la santé** : les biais nationaux, ethniques et génétiques peuvent induire – dans le cas de protocoles médicaux mobilisant l'IA, y compris *via* des systèmes à la pointe des technologies – des diagnostics et des traitements totalement inadaptés à telle ou telle population spécifique.

*b) Une domination par la culture et la langue*

Puisque les données utilisées pour l'entraînement des modèles d'intelligence artificielle proviennent très majoritairement du monde anglo-saxon, et principalement des États-Unis, cela signifie que **la langue anglaise est la principale langue utilisée pour entraîner les LLM**. Ce constat est confirmé par les tests d'évaluation (*benchmarks*) des différents grands modèles de langage où l'on peut observer que **ces modèles performant toujours mieux lorsque leurs instructions sont écrites en langue anglaise**. Selon l'AI Index de l'Université de Stanford et l'institut Epoch, toutes les données textuelles en anglais de bonne qualité existantes ont été utilisées pour entraîner les LLM.

Il n'est pas acceptable qu'un *prompt* écrit par un anglophone aux États-Unis donne de meilleurs résultats que le même *prompt* écrit en français par un Français à Paris, ou en espagnol par un Espagnol à Madrid. De même qu'il n'est pas acceptable que ce *prompt* en français ou en espagnol donne lui-même en moyenne de meilleurs résultats que le même *prompt* écrit dans l'une des langues rares prises en charge par les systèmes d'IA génératives.

Cette domination linguistique pose, de plus, d'autres problèmes, sur un plan davantage culturel, et ce à plusieurs égards. La **langue étant le vecteur de cadres culturels**, entraîner un modèle de langage principalement en langue anglaise, c'est le conduire à effectuer des associations de mots qui sont plus spécifiques à la **culture anglo-saxonne et, plus spécialement encore, la culture américaine**. L'IA conduit donc assurément vers une **uniformisation culturelle**.

Cela pourrait poser relativement moins de problèmes pour les autres cultures occidentales, assez proches de la culture américaine, car bien qu'elles soient moins représentées dans les jeux de données d'entraînement, des langues comme l'espagnol ou le français sont également parfois présentes parmi les langues des données d'entraînement. En revanche pour des cultures minoritaires éloignées de la culture américaine et dont les langues sont plus rares ou plus difficiles, et donc beaucoup moins présentes sur *Internet*, le danger est plus grand. La place de ces langues sera réduite voire nulle au sein des jeux de données et ce sont **les cultures minoritaires derrière ces langues qui se retrouvent effacées**.

L'IA, à cause des LLM surtout mais cela vaudra bientôt pour toutes les IA génératives multimodales produisant des contenus visuels ou audiovisuels, contribue à une lente **dissolution de la majorité des cultures existantes** qui s'effectue au profit d'une **hégémonie de la culture anglo-saxonne, en particulier américaine**. Cette uniformisation culturelle et sans doute aussi idéologique n'est pas souhaitable selon vos rapporteurs.

c) *Ne pas donner notre langue à ChatGPT ! Pour la diversité linguistique et culturelle*

Face à ce constat, nous invitons à **ne pas donner notre langue à ChatGPT**. Au-delà du jeu de mots, cette invite devient un impératif si l'on souhaite **conserver la diversité linguistique et culturelle de l'humanité**.

Certes, 50 langues peuvent d'ores et déjà être gérées par le système GPT-4o d'OpenAI, soit 97 % des locuteurs mondiaux<sup>1</sup>, ce qui peut donner **l'illusion d'un multilinguisme respectueux des différences entre les langues et les cultures**, mais en réalité l'IA générative **uniformise ces dernières dans un grand mixeur statistique qui produit peu ou prou les mêmes résultats quelle que soit la langue de l'utilisateur**, ce qui porte gravement atteinte à la perspective du maintien de la diversité linguistique et culturelle.

Il faut saluer les objectifs poursuivis par le projet **Bloom**, qui, dans une démarche de science ouverte et participative, avait permis en 2022 et 2023 de construire un modèle à 176 milliards de paramètres sur la base de **données équilibrées en 46 langues**, en particulier des langues rares dans le monde de l'IA comme une **vingtaine de langues d'Afrique**. Ces données étaient équilibrées car la variété linguistique était en permanence garantie pendant l'entraînement et car elles étaient issues de sources variées (littérature, articles scientifiques, presse, codes informatiques, etc.). Ce projet était piloté par **Hugging Face**, *start-up* initialement française mais devenue américaine et désormais installée à New York, dont l'un des chercheurs, Yacin Jernite, a été rencontré par vos rapporteurs à Washington. Bloom a été entraîné sur le supercalculateur français Jean-Zay et l'entreprise **LightOn**, dont le président Igor Carron a été lui aussi rencontré par vos rapporteurs, a apporté son concours à ce projet.

Étroitement associée au développement de Bloom, LightOn se singularise au sein des entreprises développant des solutions d'IA par son intérêt pour une offre de systèmes génératifs performants et reposant sur des langues autres que l'anglais, souvent en *open source*. En partenariat avec Amazon pour l'infrastructure cloud d'AWS, l'entreprise française a développé plusieurs versions de son modèle *open source* **Alfred**. Avec les Émirats arabes unis, elle a contribué à développer et améliorer le modèle *open source* **Falcon 40B**, prisé dans le monde arabe, qui permet de construire des LLM agiles en arabe et ne reflétant pas de manière biaisée la culture anglo-saxonne et, plus spécialement, la culture américaine. C'est bien **plus qu'un marché de niche**, il s'agit aussi d'un **enjeu de diversité linguistique et culturelle**.

---

<sup>1</sup> Cf. Benj Edwards, 2024, « Major ChatGPT-4o update allows audio-video talks with an "emotional" AI chatbot », Ars Technica: <https://arstechnica.com/information-technology/2024/05/chatgpt-4o-lets-you-have-real-time-audio-video-conversations-with-emotional-chatbot/>

Avec le cofondateur de l'entreprise, Laurent Daudet, Igor Carron se singularise aussi par sa préoccupation pour la **souveraineté des données** et la **commercialisation de systèmes d'IA installés sur site et pas sur des serveurs américains**, ce qui garantit aux organisations et aux entreprises un contrôle total de leurs données. C'est par exemple le cas de sa plateforme Paradigm. Compte tenu de ses importants besoins de financement, LightOn a fait le choix de ne pas recourir qu'à des levées de fonds classiques et a procédé en novembre 2024 à son introduction en Bourse sur Euronext Growth Paris.

Nous avons, en conclusion, **besoin de modèles et de systèmes d'IA les plus souverains possibles, reflétant notre culture**, entraînés avec des données qui la reflètent le plus fidèlement possible et qui mobilisent des sources issues de notre riche patrimoine linguistique.

Dans *Civilisation. Comment nous sommes devenus américains*<sup>1</sup>, Régis Debray décrit l'américanisation de l'Europe et de la France en particulier, les Français devenant des Gallo-ricains. Dans une interview où il présente son ouvrage, il relève que cette « *grande poussée californienne s'est faite, comme d'habitude, sous le drapeau rouge : pour tourner à droite, il faut mettre le clignotant à gauche* »<sup>2</sup>.

Si l'on refuse l'accélération de cette américanisation sous l'effet de l'IA, il faudra **développer et utiliser des modèles dont les modalités d'entraînement, avec une vigilance autour des sources mobilisées, garantissent le rayonnement de la langue et de la culture française au sein du système d'IA final**.

## 2. Le danger de l'uniformisation cognitive plus encore que culturelle

### a) De nouvelles structures mentales

Comme il a été vu, le **capitalisme cognitif** qui repose sur la conjugaison des écrans et de l'IA avec la croissance exponentielle de données massives conduit à une **économie de l'attention préoccupante**, notamment car elle enferme l'utilisateur des technologies dans des **bulles de filtres**. Cet enfermement informationnel **polarise les visions de chacun dans des croyances subjectives**. Ce sont autant de **prisons mentales** qui se déclinent à l'échelle individuelle, mais surtout ces prisons mentales peuvent engendrer, sous l'effet d'usages massifs des systèmes d'IA (*a fortiori* si ces systèmes reposent peu ou prou sur les mêmes modèles), à côté et en plus des phénomènes d'**uniformisation culturelle**, une **uniformisation cognitive**.

---

<sup>1</sup> Régis Debray, 2017, *Civilisation. Comment nous sommes devenus américains*, Gallimard.

<sup>2</sup> Cf. interview de Régis Debray le 15 août 2017, « *Comment nous sommes devenus américains* » : <https://www.slate.fr/story/149742/comment-nous-sommes-devenus-americains>

Cette tendance n'est que d'apparence paradoxale : on a, dans le même temps, une **polarisation marquée des opinions et des identités** (selon les variables de la culture américaine, avec une forte dimension émotionnelle) et une **uniformisation culturelle** doublée d'une **uniformisation cognitive**.

Une vigilance est donc nécessaire face aux raisonnements probabilistes et aux dérives dans l'usage des superstatistiques qui sont au fondement de l'IA : la capacité de traiter des volumes de plus en plus massifs de données n'écarte pas le risque d'obtenir parfois des résultats qui peuvent **ne pas correspondre au monde réel et/ou à la vérité**. Sans penser comme le disaient deux chercheurs que « *les chiffres sont comme les gens, si on les torture assez, on peut leur faire dire n'importe quoi* »<sup>1</sup>, vos rapporteurs rappellent que les données sont toujours construites et n'ont pas de signification en elles-mêmes, elles ne décrivent qu'une partie de ce qui se passe dans la réalité et s'accompagnent de très nombreux **biais**. Ainsi que l'écrivait dans une sorte de mise en garde Alain Desrosières, « *les données ne sont pas données* »<sup>2</sup>.

Une autre difficulté provient du **mode de raisonnement** utilisé pour traiter les données. Comme le rappelle l'académicien Daniel Andler, spécialiste des sciences cognitives, entendu par vos rapporteurs, l'intelligence ne se réduit pas à une capacité à résoudre des problèmes, la cognition humaine est bien plus complexe que cela. Il développe sa thèse dans son ouvrage *Intelligence artificielle, intelligence humaine : la double énigme* et assure que l'IA ne connaît que des problèmes que nous lui demandons de résoudre, elle ne connaît pas le monde et ses situations. Elle reste un outil, appelé à rester très éloigné de notre intelligence en dépit de ses avancées spectaculaires.

Il faut aller plus loin que cette remarque de bon sens et analyser sur quelles formes de mécanismes de résolution des problèmes repose l'intelligence artificielle. Les **raisonnements par induction, probabilistes**, sont **prometteurs** et, donnent souvent des résultats impressionnants, mais temporairement du moins, ils tendent à nous faire oublier le grand intérêt des raisonnements déductifs.

Prenons un parfait exemple de **syllogisme déductif** : les hommes sont mortels, or les Athéniens sont des hommes, donc les Athéniens sont mortels. Ce type de raisonnement qui part d'une affirmation générale pour aboutir à une application à un cas particulier est toujours vrai car, **dès lors que ses prémisses sont vraies, il infère un résultat juste**.

---

<sup>1</sup> Martin Kersten et Lefteris Sidiourgos, 2017, « A Database System with Amnesia », CIDR : <https://www.cidrdb.org/cidr2017/papers/p58-kersten-cidr17.pdf>

<sup>2</sup> Avec cette formule, Alain Desrosières montrait qu'en général les données ne sont pas automatiquement accessibles, il faut les récupérer, ce qui par conséquent révèle un choix dans les données récupérées, dans les modèles statistiques ou d'IA utilisés, etc. Des biais peuvent ainsi apparaître dans ces opérations de construction. Cf. Alain Desrosières, 2010, La Politique des grands nombres, Histoire de la raison statistique, La Découverte.

Les **IA connexionnistes**, elles, prennent la forme inverse, et ce **comme tous les syllogismes statistiques, de raisonnements inductifs**, c'est-à-dire basés sur la **généralisation de cas particuliers**.

Or l'induction, contrairement à la déduction, est un raisonnement « inexact » du point de vue de la logique pure car il n'est appuyé que sur la vérification de sa répétition : en effet, **les syllogismes inductifs sont « valides » mais pas nécessairement vrais**. Ils peuvent de manière caricaturale prendre la forme suivante : les hommes sont mortels, or les ânes sont mortels, donc les hommes sont des ânes.

La généralisation de cas particuliers sous l'effet des données massives traitées par l'IA connexionniste est devenue la règle, or **le résultat d'une inférence suivant un raisonnement inductif, même basé sur des milliards d'exemples, peut toujours être démenti par un contre-exemple ou par plusieurs contre-exemples**.

Puisque, dans le monde du numérique, le plus souvent, *code is law*<sup>1</sup> (autrement dit les programmes informatiques eux-mêmes tendent à devenir la règle) et que ces raisonnements inductifs, souvent vrais mais parfois faux, prédominent, il ne faut jamais perdre de vue que les IA connexionnistes qui fonctionnent en très grande partie indépendamment des contextes (en particulier du monde réel et de ses contraintes physiques et temporelles par exemple) vont progressivement **modifier nos manières de penser**, et ce à mesure qu'elles deviendront de plus en plus nos outils au quotidien, y compris, et surtout, pour raisonner à notre place.

---

<sup>1</sup> L'idée est expliquée dans l'ouvrage classique du professeur de droit à la Harvard Law School : Lawrence Lessig, 1999, *Code and Other Laws of Cyberspace, Basic Books*. Aucun éditeur n'a encore malheureusement pris la peine de faire traduire ce livre depuis 25 ans. Des internautes anonymes l'ont fort heureusement fait bénévolement et nous leur rendons hommage : <https://framablog.org/2010/05/22/code-is-law-lessig/>

De cette traduction, vos rapporteurs souhaitent présenter cette citation : dans le monde numérique dont Internet, le régulateur « c'est le code (...). Ce code, ou cette architecture, définit la manière dont nous vivons le cyberspace. Il détermine s'il est facile ou non de protéger sa vie privée, ou de censurer la parole. Il détermine si l'accès à l'information est global ou sectorisé. Il a un impact sur qui peut voir quoi, ou sur ce qui est surveillé. Lorsqu'on commence à comprendre la nature de ce code, on se rend compte que, d'une myriade de manières, le code du cyberspace régule (...) si nous ne comprenons pas en quoi le cyberspace peut intégrer, ou supplanter, certaines valeurs de nos traditions constitutionnelles, nous perdrons le contrôle de ces valeurs. La loi du cyberspace - le code - les supplantera (...). Ce n'est pas entre régulation et absence de régulation que nous avons à choisir. Le code régule (...). Nous devrions examiner l'architecture du cyberspace de la même manière que nous examinons le fonctionnement de nos institutions. Si nous ne le faisons pas, ou si nous n'apprenons pas à le faire, la pertinence de notre tradition constitutionnelle va décliner. Tout comme notre engagement autour de valeurs fondamentales, par le biais d'une constitution promulguée en pleine conscience. Nous resterons aveugles à la menace que notre époque fait peser sur les libertés et les valeurs dont nous avons hérité. La loi du cyberspace dépendra de la manière dont il est codé, mais nous aurons perdu tout rôle dans le choix de cette loi. »

Le fait que le raisonnement par déduction décline déjà et que l'on finisse à l'avenir par ne plus recourir qu'à des modèles statistiques qui raisonnent par induction risque, selon certains chercheurs, de conduire à la **remise en cause de nos progrès scientifiques** en raison de la « *mort de la théorie et la mort de la théorie scientifique* », car la science s'est principalement construite grâce des raisonnements déductifs<sup>1</sup>.

**Les deux formes de raisonnement doivent continuer à cohabiter de manière plus équilibrée<sup>2</sup>** sans quoi cette ère de l'IA et du Big Data va conduire les habitants de la planète entière à **penser selon le même mode**, non seulement, sans le savoir, **avec les mêmes cadres en termes de références culturelles**, mais aussi **selon les mêmes structures cognitives tournées vers l'induction**.

De plus, à côté d'une attraction de l'attention, dont on a vu qu'elle était déjà un danger, nous courons le risque plus général en nous en remettant aux algorithmes, à l'IA et aux interfaces numériques d'une **dissolution durable de nos capacités à nous souvenir, à nous concentrer et à traiter l'information**, les plus jeunes, notamment les petits enfants, étant particulièrement exposés à ces évolutions.

Outre la **surexposition des enfants aux écrans** qui les rend **dépendants** aux outils numériques dès leur plus jeune âge et les effets - spécifiques sur eux - de la surcharge informationnelle qu'ils subissent, les **impacts cognitifs et physiques** existent, surtout quand les plus petits sont privés d'interactions avec leurs parents, ceux-ci étant eux-mêmes de plus en plus absorbés par les écrans. Ces problèmes ont déjà été soulevés depuis une dizaine d'années par le psychiatre Serge Tisseron, le philosophe Bernard Stiegler ou, encore, l'anthropologue Pascal Plantard, ces enjeux conduisant **l'Académie des sciences, puis l'Académie nationale de médecine et l'Académie des technologies à se saisir directement du sujet en 2019<sup>3</sup>**.

---

<sup>1</sup> D'après Chris Anderson, la remise en cause des méthodes scientifiques traditionnelles par l'IA et les données massives pourrait tout simplement conduire à la fin de la théorie, c'est-à-dire à la fin de la science telle que nous la connaissons. Cf. Chris Anderson, juin 2008, « The End of Theory : The Data Deluge Makes the Scientific Method Obsolete », *Wired* : <https://www.wired.com/2008/06/pb-theory/>

<sup>2</sup> La piste de la conjugaison entre IA connexionnistes et IA symboliques, au moins sur le plan de l'architecture des modèles, pourrait être une première réponse. Les modèles de type Trees of thoughts ou des Mixtures of experts avec les deux types d'IA pourraient devenir demain des architectures neuro-symboliques performantes conjuguant de l'inductif et du déductif.

<sup>3</sup> Cf. en particulier les références suivantes : Serge Tisseron et Bernard Stiegler, Faut-il interdire les écrans aux enfants ?, *Mordicus*, 2009 ; Pascal Plantard, « Numérique et éducation : encore un coup de tablette magique ? », *Administration et Éducation*, n° 146, 2015, « Le collège et les pratiques numériques des adolescents », *Les Cahiers Pédagogiques*, n° 520, 2015, avec Jonathan Bernard et Sophie Jehel « Tablette, smartphone, console, télé, ordi... Faut-il les interdire aux enfants ? », *revue de l'Inserm*, 2019 ; Jean-François Bach, Olivier Houdé, Pierre Léna et Serge Tisseron, L'enfant et les écrans, avis de l'Académie des sciences, Le Pommier, 2013 ; et « L'enfant, l'adolescent, la famille et les écrans », appel commun de l'Académie des sciences, de l'Académie nationale de médecine et de l'Académie des technologies, 2019 : [https://www.academie-sciences.fr/pdf/rapport/appel\\_090419.pdf](https://www.academie-sciences.fr/pdf/rapport/appel_090419.pdf)



Cette évolution vers **de nouvelles structures mentales de plus en plus dégradées** chez les enfants sous l'effet des écrans et des outils numériques fait l'objet d'enquêtes empiriques approfondies très préoccupantes, à l'instar des travaux de Marie-Claude Bossière par exemple<sup>1</sup>.

Cette problématique est plus large que celle de l'intelligence artificielle mais **les technologies d'IA accélèrent et renforcent de manière décisive ces évolutions vers une uniformisation cognitive**, qui pourrait selon Bruno Patino nous emmener vers « une civilisation de poissons rouges ». La formule peut choquer mais elle est démontrable empiriquement. Des chercheurs de Google ont ainsi identifié des similitudes entre les durées moyennes d'attention des poissons rouges et celles des jeunes générations.

*b) Vers une civilisation de poissons rouges ?*

Nos structures cognitives sont modifiées par les outils numériques que sont l'IA, les écrans et l'économie de l'attention qui les structure, pire, nous serions les victimes d'une civilisation dans laquelle **la santé mentale est menacée à partir de plus de 30 minutes consécutives d'exposition aux réseaux sociaux et aux contenus actuels d'Internet** comme le montrent les chercheurs<sup>2</sup>, une civilisation dans laquelle la durée d'attention sur n'importe quel sujet serait réduite à quelques secondes, faisant de chacun de nous **un poisson rouge enfermé dans le bocal de son écran**, ainsi que l'affirme sans provocation Bruno Patino dans un livre important<sup>3</sup> paru en 2019, d'où la question sur l'impact d'Internet et du numérique que posait Nicholas Carr il y a déjà 16 ans, « *Google nous rend-il tous idiots ?* »<sup>4</sup>.

Bruno Patino rappelle dans son ouvrage que les ingénieurs de Google ont réussi à calculer la **durée maximale de l'attention du poisson rouge** qui tourne dans son bocal et semble redécouvrir le monde à chaque tour, **soit huit secondes** ; or ces mêmes ingénieurs ont aussi évalué la **durée moyenne d'attention de la génération des millenials**, celle qui a grandi avec les écrans connectés, **soit neuf secondes**.

---

<sup>1</sup> Cf. par exemple : Marie-Claude Bossière, « Le pédopsychiatre et la toxicité de l'omniprésence des écrans », *nouvelle revue de l'enfance et de l'Adolescence*, vol. 2, n° 1, 2020, ainsi qu'avec Daniel Marcelli et Anne-Lise Ducanda, « L'exposition précoce et excessive aux écrans (EPEE) : un nouveau syndrome », *revue Devenir*, vol. 32, n° 2, 2020.

<sup>2</sup> Cf. Melissa G. Hunt, Rachel Marx, Courtney Lipson et Jordyn Young, « No More FOMO : Limiting Social Media Decreases Loneliness and Depression », *Journal of Social and Clinical Psychology*, December vol. 37, n° 10, 2018.

<sup>3</sup> Bruno Patino, *La civilisation du poisson rouge : Petit traité sur le marché de l'attention*, Grasset, 2019.

<sup>4</sup> Article de Nicholas Carr, « Is Google Making Us Stupid ? What the Internet is doing to our brains », *The Atlantic*, n° 7, 2008, disponible au lien suivant :

<https://www.theatlantic.com/magazine/archive/2008/07/is-google-making-us-stupid/306868/>



C'est pourquoi nous sommes, selon Bruno Patino, en train de devenir des « *poissons rouges, enfermés dans le bocal de nos écrans, soumis au manège de nos alertes et de nos messages instantanés, totalement dépendants de technologies dangereuses car favorisant la solitude et la dépression* ».

L'étude fameuse, qui évalue à 30 minutes le temps maximum d'exposition aux réseaux sociaux et aux écrans d'Internet au-delà duquel apparaît un **risque pour la santé mentale**, est justement citée par Bruno Patino dans son ouvrage. Les travaux du chercheur en neurosciences Michel Desmurget apportent également un éclairage sur les effets cognitifs délétères des écrans, dont le premier serait la fabrique de « *crétins digitaux* »<sup>1</sup>.

Dans ce contexte une **éducation au numérique en général** et une **éducation à l'IA en particulier** apparaissent comme des accompagnements indispensables à ces évolutions et des urgences impérieuses pour la cohésion de nos sociétés et la santé de chacun.

c) *Éduquer à l'IA, éduquer par l'IA*

Face à toutes ces analyses plutôt pessimistes qui viennent relativiser l'enthousiasme généralisé autour de l'IA, il existe un **espoir de bénéficier des opportunités de ces technologies en limitant leurs risques pour nos cultures et nos structures cognitives** : il faut **éduquer à l'intelligence artificielle**.

Pour **démystifier l'IA** mais surtout pour **permettre l'appropriation de ces technologies** et donc maximiser ses bénéfices pour l'économie, il est indispensable de disposer de programmes de **formation initiale** mais aussi de **formation continue**, avec l'idée de toucher les publics les plus larges possibles. En effet, la formation à l'IA ne doit pas se faire seulement à destination du monde des étudiants et des actifs. Il faut la décliner pour les scolaires, les collégiens, les lycéens, les ruraux, les inactifs, en bref, le grand public. De ce point de vue, les politiques conduites par la Finlande pour des apprentissages accessibles à tous – qui seront présentées de manière détaillée dans la troisième partie du rapport – sont des modèles à suivre. Il n'est pas sûr que les « Cafés IA », proposition de la commission de l'intelligence artificielle et du Conseil national du numérique, reprise par le Président de la République, réponde à l'objectif. En dépit de son ouverture affichée, ce dispositif risque de se révéler **élitiste**, en ne réunissant dans des zones urbaines denses que des personnes déjà intéressées par l'IA : vos rapporteurs pensent qu'il aurait été préférable d'élargir celui des **conseillers numériques**.

---

<sup>1</sup> Cf. les ouvrages de Michel Desmurget, 2019, *La Fabrique du crétin digital. Les dangers des écrans pour nos enfants, et, 2023, Faites-les lire !: Pour en finir avec le crétin digital. Ces études qui dressent le constat d'une baisse de l'intelligence chez les jeunes générations doivent cependant être relativisées à la lumière de dimensions plus qualitatives. Le contenu visionné sur les écrans doit être davantage incriminé que l'objet écran en lui-même, ce que soulignent d'autres chercheurs comme Tamara Hudon, Christopher Fennell, Matthew Hoftyzer, Franck Ramus ou Grégoire Borst.*

Lancé en 2021<sup>1</sup>, ce dispositif, qui a permis de déployer 4 000 conseillers dans les territoires, devait avoir un coût de 250 millions d'euros sur trois ans. Ces conseillers pourraient se voir confier une mission d'acculturation à l'IA. Malheureusement, au lieu d'étendre le champ d'interventions de leurs interventions, le gouvernement prévoit plutôt de le restreindre. Le projet de loi de finances pour 2025 ne prévoit ainsi que 28 millions d'euros pour le financement de ce dispositif, soit une **baisse de plus de 50 % du budget qui lui était alloué auparavant**.

En plus d'éduquer à l'intelligence artificielle, nous allons pouvoir **éduquer par l'IA**. Ces technologies d'IA vont jouer un rôle crucial dans les apprentissages, l'éducation et le développement personnel, en créant des **expériences et des supports d'apprentissage sur mesure**, en s'adaptant aux besoins de chaque élève<sup>2</sup>. Les premiers logiciels éducatifs, comme Duolingo, utilisaient des méthodes basiques pour enseigner les langues pas à pas sans véritable personnalisation. Aujourd'hui, des plateformes comme Coursera ou Khan Academy utilisent l'IA pour **personnaliser les parcours d'apprentissage**, s'adaptant au niveau et au rythme de chaque utilisateur. Les assistants basés sur des modèles performants d'**IA générative**, comme ChatGPT, aident à répondre à des questions de plus en plus complexes, favorisant l'apprentissage autonome et demain il sera facile de **créer des enseignants virtuels** capables de simuler des interactions riches et réalistes, adaptées aux besoins uniques de chaque apprenant, facilitant un **apprentissage immersif** dans des domaines variés.

Dans un rapport, la délégation à la prospective du Sénat s'est intéressée à ce sujet<sup>3</sup> en mettant toutefois davantage en avant les **bénéfices potentiels** que les nombreux **risques** de l'IA dans le secteur éducatif (pour mémoire les effets cognitifs, l'impact sur le développement psycho-émotionnel des enfants, les biais, l'anthropomorphisme, la dépendance, l'absence d'usages différenciés selon l'âge).

---

<sup>1</sup> Le dispositif, composante de France Services, cherche à rendre le numérique plus accessible à tous. Il visait trois objectifs principaux : soutenir les Français dans les usages numériques (travailler à distance, consulter un médecin, gérer des courriels, suivre la scolarité des enfants...), sensibiliser aux enjeux du numérique pour favoriser des usages citoyens et critiques (protection des données personnelles, vigilance quant à l'information, maîtrise des réseaux sociaux...), accompagner les citoyens vers une plus grande autonomie, notamment dans la réalisation de leurs démarches administratives. Plus de 4,1 millions d'accompagnements ont été réalisés dans le cadre de cette opération gérée par la Banque des Territoires, à travers 2 880 structures mobilisées sur l'ensemble du territoire. L'ensemble de la démarche a été saluée pour son efficacité par la Cour des comptes.

<sup>2</sup> Le ministère de l'éducation nationale a lancé plusieurs initiatives autour de l'IA dans l'éducation, dont des assistants pédagogiques numériques, un manuel pour les enseignants, une enquête auprès de 5 600 enseignants et étudiants et un rapport « Intelligence artificielle et éducation : apports de la recherche et enjeux pour les politiques publiques », ce site de l'Académie de Paris recense ces initiatives : <https://www.ac-paris.fr/l-intelligence-artificielle-dans-l-education-130992>

<sup>3</sup> « L'IA et l'avenir du service public, rapport thématique #3 : IA et éducation » : <https://www.senat.fr/rap/r24-101/r24-101-syn.pdf>

Sa lecture doit être complétée des **travaux de l'Unesco sur les relations entre IA et éducation** (ils seront évoqués plus loin) notamment son « **Guide pour l'IA générative dans l'éducation et la recherche** »<sup>1</sup>, et du travail critique des chercheurs sur le sujet, ainsi que de l'expertise précise apportée sur ces enjeux par la **Fondation Everyone.AI** dont vos rapporteurs ont rencontré les responsables à San Francisco<sup>2</sup>.

## **B. LA CRÉATION AU DÉFI DE L'IMPACT DE L'INTELLIGENCE ARTIFICIELLE SUR LA PROPRIÉTÉ INTELLECTUELLE ET LES DROITS D'AUTEUR**

La diffusion de l'intelligence artificielle, surtout de l'IA générative, permet de **rendre automatisables nos tâches intellectuelles**, et certaines de nos compétences dans le domaine de la création sont directement concurrencées. Alors que l'on considérait auparavant la **créativité artistique** comme étant un talent propre à l'homme, elle est désormais devenue, dans une certaine mesure, accessible à l'intelligence artificielle. Cela est possible grâce à l'utilisation massive de données d'entraînement (images, œuvres d'art, etc.) dont l'origine est souvent assez floue, la mobilisation d'œuvres protégées étant fréquente. Cela soulève plusieurs questions sur l'évolution de nos sociétés en général et de notre droit en particulier pour s'adapter à ces changements.

Les **régimes juridiques de propriété intellectuelle** existants font face à des défis qu'ils n'ont jamais rencontrés auparavant. Ils sont **incapables**, d'une part, de permettre la **rémunération ou l'indemnisation des titulaires de droit sur les œuvres** ayant servi de sources aux modèles d'IA, d'autre part, de **déterminer**, pour l'heure, les **règles applicables aux œuvres créées par intelligence artificielle**, dont le statut reste incertain.

---

<sup>1</sup> Le rapport de l'Unesco propose des recommandations en vue d'encadrer ces technologies à la lumière de principes éthiques en promouvant l'inclusion et l'équité. Il met l'accent sur une approche centrée sur l'humain, en prônant surtout une vigilance quant aux usages de l'IA dans les contextes éducatifs, ces usages devant être éthiques, sûrs, justes et dotés de sens. Il propose des mesures pour intégrer de manière responsable l'IA dans les activités d'enseignement, d'apprentissage et de recherche, notamment avec une explication pédagogique des technologies d'IA générative, une discussion de leurs enjeux éthiques et politiques et de leurs perspectives d'encadrement. En outre, le rapport fournit des exemples d'utilisation de l'IA générative permettant d'en exploiter les avantages pour la pensée critique ainsi que la créativité dans l'éducation et la recherche, tout en en atténuant les risques. Il explique par exemple comment l'IA générative peut être utilisée de manière créative dans la conception des programmes, l'enseignement et les activités d'apprentissage tout en abordant les implications de plus long terme pour l'éducation et la recherche. Cf. le guide publié en septembre 2023 par l'Unesco sur l'IA dans l'éducation et la recherche : <https://www.unesco.org/fr/articles/orientations-pour-lintelligence-artificielle-generative-dans-l'education-et-la-recherche>

<sup>2</sup> Cf. en particulier l'article de Mathilde Neugnot-Cerioli et Olga Muss Laurenty, 2024, « The Future of Child Development in the AI Era. Cross-Disciplinary Perspectives Between AI and Child Development Experts » : <https://arxiv.org/abs/2405.19275> ainsi que le site de la Fondation Everyone.AI <https://everyone.ai>

En outre, l'automatisation des processus de création artistique est susceptible de **modifier en profondeur l'ensemble des industries culturelles**, en permettant aux studios de créer des contenus rapidement et à moindre coût.

### 1. Des régimes de propriété intellectuelle fragilisés

L'intelligence artificielle a d'ores et déjà un **impact sur les régimes de propriété intellectuelle**, aussi bien sur le droit d'auteur français que sur le *copyright* anglo-saxon, qui bien que convergents sur la forme depuis la convention de Berne de 1886 restent différents. En effet, il n'existe toujours pas de droit d'auteur au niveau international.

Pour mémoire, le droit d'auteur français considère une œuvre comme une **extension de la personnalité de son créateur**, conférant un droit de propriété incorporel fort, qui englobe des attributs d'ordre intellectuel, moral et patrimonial<sup>1</sup>. Le droit moral est ainsi **perpétuel, inaliénable et imprescriptible** même si le droit patrimonial, d'une durée variable, a fait tomber l'œuvre dans le domaine public.

En revanche, le *copyright* relève d'une logique strictement économique et accorde un droit moral restreint, qui se concentre principalement sur le **support matériel de l'œuvre** et les **intérêts financiers** du titulaire du *copyright* (qui peut être l'auteur ou pas)<sup>2</sup>. Des modalités de régime juridique différentes peuvent également exister d'un pays à un autre.

Entraîner un modèle d'intelligence artificielle nécessite de grandes quantités de données, collectées de manière automatique à travers de vastes **jeux de données**, qui **contiennent donc parfois tout ou partie d'œuvres soumises au *copyright* ou au droit d'auteur**.

Sans transparence de la part des développeurs de modèles de fondation, il est difficile de savoir exactement quelles œuvres sont présentes au sein des données d'entraînement du modèle. Il est donc **difficile pour un ayant droit de faire valoir ses droits** pour atteinte au *copyright* ou au droit d'auteur.

---

<sup>1</sup> Beaumarchais, fondateur de la première société d'auteurs en 1777, défendait le principe d'un tel droit d'auteur qui sera consacré par la loi pendant la Révolution française. On retrouve cette conception dans notre droit aujourd'hui. Cf. par exemple l'article L 111-1 du code de la propriété intellectuelle sur le site Légifrance : [https://www.legifrance.gouv.fr/codes/article\\_lc/LEGIARTI000042814694](https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000042814694)

<sup>2</sup> Cette approche est présente dès la fondation du droit américain des brevets avec le Patent Act de 1790. Un discours ou une chorégraphie n'est par exemple pas protégé tant qu'il n'est pas transcrit sur un support. Le droit américain prévoit l'enregistrement de protection du *copyright* auprès du U.S. Copyright Office : ce n'est pas obligatoire, mais en cas de litige à la suite d'une utilisation ou d'une copie, si l'œuvre n'est pas enregistrée, son auteur risque de ne bénéficier d'aucun dédommagement financier.

Cela est d'autant plus problématique que la directive 2019/790 du 17 avril 2019 de l'Union européenne<sup>1</sup> prévoit paradoxalement une possibilité pour les ayants droit de refuser (« *opt-out* ») que leurs œuvres soient utilisées dans les bases de données d'IA. Un droit de retrait qui ne peut donc rester que formel. Alexandra Bensamoun, professeure de droit à l'Université Paris-Saclay et spécialiste du droit d'auteur et de l'intelligence artificielle, a affirmé lors de son audition devant la commission de la culture, de l'éducation et de la communication du Sénat<sup>2</sup> que l'on sait aujourd'hui que des modèles d'IA ont utilisé dans leurs données d'entraînement des **données d'auteurs qui avaient pourtant d'ores et déjà fait valoir leur droit de retrait des bases de données**.

L'absence de respect des règles relatives à la propriété intellectuelle est également **reconnue par les représentants des grandes entreprises américaines de l'intelligence artificielle eux-mêmes**. Ils ne se cachent pas de violer le droit d'auteur ou le copyright, dès lors que les œuvres sont sur Internet. Ainsi, le directeur technique d'OpenAI, Mustafa Suleyman, a pu avouer : « *Je pense qu'en ce qui concerne le contenu qui se trouve déjà sur le Web ouvert, le contrat social de ce contenu depuis les années 1990 est l'utilisation équitable. Tout le monde peut le copier, le recréer, le reproduire. C'est ce que l'on appelle le "freeware" si l'on veut, et c'est ce que l'on a compris* »<sup>3</sup>.

Cette question relevant principalement du sujet des données utilisées, la Commission nationale de l'informatique et des libertés (Cnil) est directement concernée par ces questions. Lors de son audition par vos rapporteurs, Bertrand Pailhès, directeur des technologies et de l'innovation à la Cnil et ancien coordinateur national pour l'intelligence artificielle, a rappelé que le règlement général pour la protection des données (RGPD)<sup>4</sup> doit s'appliquer à l'entraînement des modèles d'intelligence artificielle. Aussi, en théorie, la **collecte de contenus ne peut se faire en provenance de sources manifestement illégales**. La collecte de données doit en outre répondre à une **finalité explicitement définie** et s'appuyer sur une base légale, souvent **l'intérêt légitime**. La Cnil a constitué des « fiches pratiques » qui permettent d'assurer le respect de la protection des données lors de leur traitement par des systèmes d'IA, ce qui inclut donc, sans s'y limiter, la protection de la propriété intellectuelle.

---

<sup>1</sup> Cf. la directive du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique : <http://data.europa.eu/eli/dir/2019/790/oj/fra>

<sup>2</sup> Le compte rendu de la réunion est disponible en ligne : <https://www.senat.fr/compte-rendu-commissions/20231218/cult.html>

<sup>3</sup> Sean Hollister, 2024, « Microsoft's AI Boss Thinks It's Perfectly Okay to Steal Content If It's on the Open Web », *The Verge* : <https://www.theverge.com/2024/6/28/24188391/microsoft-ai-suleyman-social-contract-freeware>

<sup>4</sup> Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données) (Texte présentant de l'intérêt pour l'EEE).

Les législations françaises et européennes déterminent ainsi un **cadre théorique assez clair** pour le respect de la propriété intellectuelle et fixent des limites aux cas d'entraînement des modèles d'intelligence artificielle. Elles permettent de **définir ce qui constitue des pratiques acceptables ou pas**. Cela est d'autant plus important qu'il semble que beaucoup d'entreprises développant des solutions d'intelligence artificielle semblent **tentées de contourner ces principes**. Ces discours se retrouvent en effet souvent dans le secteur du numérique, y compris du côté de figures emblématiques de la *French Tech*, comme Oussama Ammar, cofondateur de l'incubateur The Family avec Alice Zagury et Nicolas Colin, qui a souvent fait de la transgression des règles le cœur de ses conseils aux *start-up* en vue d'une innovation disruptive<sup>1</sup>.

Néanmoins, fixer des règles définissant ce cadre légal n'est pas en soi suffisant, il faut également **s'assurer de disposer des moyens permettant de contraindre les entreprises de respecter ce cadre**. Or, les sociétés font elles-mêmes valoir des **principes de confidentialité** qui contreviennent à cet objectif, ce qui pose un problème de transparence de l'information. Il est **quasi impossible pour les autorités ou pour un juge d'obtenir des informations sur les bases de données d'entraînement utilisées** sans y avoir directement accès<sup>2</sup> et il est donc **très difficile de prouver**, sur la seule base des résultats d'un modèle, que ce dernier contient dans sa base de données d'entraînement du contenu illicite. Et quand bien même le régulateur ou un juge aurait accès aux milliards de données utilisées, le **tri pour identifier les œuvres protégées resterait lui aussi très ardu**.

Les entreprises développant des modèles de fondation considèrent, de plus, que **les données collectées et leur traitement avant même l'entraînement** (nettoyage/curation) constituent une **source de valeur** pour elles. Ouvrir ces données constituerait pour elles une perte d'avantages comparatifs vis-à-vis de leurs concurrents. C'est cette position qu'ont tenue les responsables de MistralAI lors de leur audition devant vos rapporteurs. Il y a donc **un équilibre à trouver** entre, d'une part, **l'intérêt général et la préservation des intérêts des ayants droit**, le régulateur ou les titulaires des droits devant pouvoir avoir accès aux bases de données d'entraînement des modèles d'IA, et, d'autre part, **les intérêts des entreprises développant les modèles d'IA** pour qui ces données d'entraînement sont un élément de valeur en soi et une ressource indispensable pour l'entraînement des modèles.

---

<sup>1</sup> Cf. par exemple les conseils destinés aux entrepreneurs et aux jeunes (qu'Oussama Amar appelait la génération « tuto ») et distillés tout au long des années 2010 dans les conférences nommées « Les barbares attaquent » réalisées par The Family : <https://www.youtube.com/watch?v=6k7ClozOnjk>

<sup>2</sup> Il existe des techniques dites « d'extraction » qui permettent d'obtenir des informations sur les données utilisées dans l'entraînement d'un modèle, mais il s'agit souvent d'un cas de détournement des règles d'utilisation des modèles selon les entreprises. Ces techniques peuvent il est vrai être utilisées de façon malveillante.



## 2. Des risques contentieux

Dans ce contexte d'incertitudes, **les risques contentieux sont de plus en plus grands**, qu'il s'agisse de l'utilisation d'œuvres protégées pour entraîner les modèles, de la protection des œuvres générées par des systèmes d'IA ou, encore, de tout autre litige qui pourrait émerger. En l'absence de règles claires, il reviendra aux juges de trancher les litiges. **Le rôle de la jurisprudence sera donc central** et laisse les artistes, les entreprises et les utilisateurs dans un flou juridique anxiogène, avec des risques financiers qui ne sont pas négligeables.

C'est pourquoi **une clarification de ces enjeux et des régimes juridiques applicables est indispensable**. Le rapport d'information de la commission des lois de l'Assemblée nationale déposé en conclusion des travaux de sa mission d'information sur les défis de l'intelligence artificielle générative en matière de protection des données personnelles et d'utilisation du contenu généré fournit **des pistes utiles à la réflexion**<sup>1</sup>. Vos rapporteurs y renvoient et souhaitent qu'un débat ait lieu à ce sujet.

## 3. Le modèle économique de la création artistique par l'IA

Par définition, **le système de droit d'auteur se retrouve davantage fragilisé par l'apparition d'œuvres générées par l'IA que le système de copyright**. En effet, le premier rattache l'œuvre à son auteur avant de poser la question du titulaire du droit patrimonial. Or, la définition de l'auteur dans le cas d'une œuvre créée par un modèle d'IA est difficile. **Qui est l'auteur d'une œuvre générée par IA ?** L'auteur est-il le développeur du modèle de fondation, le distributeur de l'application d'IA générative, l'utilisateur qui a formulé une instruction ou, de manière plus complexe, l'ensemble des auteurs qui ont vu leurs œuvres être utilisées pour l'entraînement du modèle et parvenir à l'œuvre finale ? Ces questions appellent à **réfléchir à la notion même de droit d'auteur qui, s'il apparaît de moins en moins adapté, doit tout de même être sanctuarisé à l'heure de l'IA générative, fusse au prix d'importantes adaptations**.

Vos rapporteurs ont rencontré des artistes créateurs d'œuvres utilisant l'IA : le collectif d'artistes Obvious, représenté par Pierre Fautrel, ainsi que Christophe Labarde, organisateur de l'exposition « Irruption – Quand l'intelligence artificielle bouleverse la création » au Château de Turenne. Pour mémoire, le collectif Obvious est notamment à l'origine de l'œuvre générée par IA « La famille de Belamy », vendue aux enchères pour un prix de 432 000 dollars, ce qui a constitué un record.

---

<sup>1</sup> Rapport d'information de la commission des lois de l'Assemblée nationale déposé en conclusion des travaux de sa mission d'information sur les défis de l'intelligence artificielle générative en matière de protection des données personnelles et d'utilisation du contenu généré de Philippe Pradal et Stéphane Rambaud), rapport d'information n° 2207, 16<sup>e</sup> législature : [https://www.assemblee-nationale.fr/dyn/16/rapports/cion\\_lois/l16b2207\\_rapport-information](https://www.assemblee-nationale.fr/dyn/16/rapports/cion_lois/l16b2207_rapport-information)



Ces deux experts ont admis qu'il était **trop difficile de déterminer si une œuvre réalisée par l'IA était inspirée d'une œuvre déjà existante**, sauf à tomber dans la pure et simple copie, la contrefaçon étant déjà un délit puni par la loi<sup>1</sup>.

Une solution minimale pour réguler les créations par l'intelligence artificielle pourrait être l'application d'un **filigrane** sur les médias créés par des modèles d'intelligence artificielle (*watermarking*). Ainsi, il serait possible de distinguer les œuvres générées par des modèles d'IA et celles d'origine humaine. Une telle solution se heurte néanmoins à des problèmes pratiques : il faudrait que le filigrane ne soit **pas trop visible afin de ne pas dénaturer l'œuvre** ; la modification de l'œuvre ne devrait plus être possible après l'application du filigrane pour ne pas altérer ce dernier ou le faire disparaître ; il faudrait que le filigrane puisse être détecté sur différents supports, numériques et physiques. Restera en outre la question de savoir comment appliquer de tels filigranes **à différents types d'œuvres** : textes, images, sons, musiques, vidéos en tout genre générés par intelligence artificielle. Autant de défis qui feront l'identification concrète des œuvres créées par des systèmes d'intelligence artificielle un sujet complexe et multidimensionnel, au moins à ce stade de l'état des connaissances et des techniques.

#### 4. L'avenir de la création artistique

L'arrivée d'outils d'intelligence artificielle capables de générer du contenu artistique sur la base de données numériques, couplée à l'avènement du *Big Data*, sont des facteurs de **transformation profonde de la création artistique et de l'ensemble des industries culturelles** qui pourront **créer du contenu plus rapidement, plus efficacement et à moindre coût** grâce à l'IA générative, avec une **offre plus personnalisée par l'analyse des données des utilisateurs**.

Aux États-Unis, vos rapporteurs ont rencontré Matthieu Lorrain et Surya Tubach, représentants de Google Creativ, la filiale de Google en charge des industries culturelles, qui leur ont parlé des initiatives prises par le géant du numérique en matière de création artistique. Ils affirment que le futur consistera par exemple à produire des « contenus liquides » (*liquid content*), à savoir des **contenus adaptables aux préférences des utilisateurs grâce à des intelligences artificielles capables d'analyser toutes les données produites par les utilisateurs**.

---

<sup>1</sup> Article L.335-2 du Code de la propriété intellectuelle : « Toute édition d'écrits, de composition musicale, de dessin, de peinture ou de toute autre production, imprimée ou gravée en entier ou en partie, au mépris des lois et règlements relatifs à la propriété des auteurs, est une contrefaçon et toute contrefaçon est un délit. »

Face à un tel modèle économique et technique, il pourrait devenir difficile aux productions traditionnelles de rivaliser avec ces types de contenu spécifiquement conçus pour répondre aux attentes de chaque utilisateur. Il s'agirait alors d'un pas supplémentaire vers la marchandisation des industries culturelles, notamment audiovisuelles, en faisant notamment disparaître la vision d'auteur des films et des séries. La perspective unique et très personnelle que peut apporter un réalisateur talentueux pourrait ne plus exister par exemple. Il ne s'agirait plus que de créer des **contenus audiovisuels sur mesure**, basés sur les « préférences » supposées des utilisateurs.

En plus de brider la créativité artistique, ce type de contenu conduirait paradoxalement à **renforcer les bulles de filtres** et mènerait à une **plus grande uniformisation des films et des séries**, limitées à une déclinaison de tropes narratifs populaires moyens ajustés à l'aide d'**artefacts superficiels de personnalisation basés sur les préférences supposées du spectateur**. Une telle évolution cloisonnerait encore davantage chacun dans un univers culturel aux perspectives restreintes et aux contenus appauvris. Avant même le recours généralisé à l'IA, les contenus et les recommandations de plateformes audiovisuelles telles que Netflix vont déjà en partie dans ce sens. Cette société du spectacle par individualisation des contenus inquiète vos rapporteurs.

De telles transformations profondes de la création artistique et des industries culturelles entraîneront de manière très probable **une forte conflictualité sociale**, comme l'ont montré **les longues grèves de syndicats de scénaristes** aux États-Unis, soutenues par différentes corporations d'artistes, dont celle des acteurs. La grève de la *Writers Guild of America* en 2023 avec ses 11 500 scénaristes face à l'*Alliance of Motion Picture and Television Producers* a ainsi duré du 2 mai au 27 septembre 2023, ce qui est historique.

## C. DES BÉNÉFICES CONSIDÉRABLES POUR LA RECHERCHE

### 1. La fertilisation des autres disciplines scientifiques par l'IA

Bien qu'elle constitue une discipline scientifique en soi, l'intelligence artificielle ne saurait être traitée isolément des autres disciplines scientifiques, selon une approche « en silo ». Par les possibilités qu'elle offre, l'intelligence artificielle est en effet **un outil pour toutes les disciplines scientifiques**, c'est d'ailleurs pourquoi en 2024 les **prix Nobel de Physique et de Chimie** sont l'un et l'autre revenus à des chercheurs en IA<sup>1</sup>. C'est aussi le sens de l'expression « IA plus X », développée par l'exécutif dans la seconde phase du programme d'investissement pour l'IA.

---

<sup>1</sup> En 2024, le prix Nobel de Physique a été attribué à Geoffrey Hinton et John Hopfield pour leurs travaux sur les réseaux de neurones artificiels et le prix Nobel de Chimie a été décerné aux trois pionniers de l'exploration des protéines assistée par ordinateur, David Baker, John Jumper et Demis Hassabis, fondateur et directeur du laboratoire d'IA Google DeepMind.

En France, nos instituts de recherches, à l'image de l'Institut national de recherche en sciences et technologies du numérique (Inria) ou du Commissariat à l'énergie atomique et aux énergies alternatives (CEA), rencontrés par vos rapporteurs, mènent des **projets transversaux visant à développer l'intelligence artificielle autour de projets liant recherche en IA et autres domaines de recherche**. Ainsi, le CEA conduit des projets visant à développer des systèmes d'intelligence artificielle suffisamment fiables pour être déployés dans des chaînes de production industrielle, ce qui nécessite de croiser les compétences de chercheurs en intelligence artificielle et de chercheurs spécialisés dans différents domaines industriels.

L'intelligence artificielle est un outil qui peut être utilisé dans un ensemble de disciplines scientifiques, ce qui implique une réflexion sur l'usage des intelligences artificielles dans des contextes divers. L'utilisation d'une IA comme agent conversationnel diffère par exemple de l'utilisation d'une IA dans un contexte d'industrie lourde ou médicale.

Le fait que GPT-4 ait été capable de résoudre en 2023 un problème complexe de physique du **concours d'entrée à Polytechnique**<sup>1</sup> prouve que les IA peuvent être des ressources très utiles pour des projets de recherche scientifique, y compris dans des domaines avancés. Il est souvent dit que les LLM ont actuellement, malgré leurs défauts (comme les hallucinations ou leur absence de logique), les capacités de **bons doctorants**<sup>2</sup>.

Les modèles d'IA permettent d'accélérer les progrès scientifiques. En 2022, ils ont par exemple été utilisés pour aider à la **fusion de l'hydrogène**, améliorer l'efficacité de la **manipulation de matrices** ou, encore, générer de **nouveaux anticorps**. En matière de santé, la recherche peut bénéficier de gains de temps, du criblage des médicaments et de nouvelles opportunités vers la médecine personnalisée. Pour analyser des problèmes de géométrie, DeepMind a créé **Alphageometry** qui donne d'excellents résultats et se base sur une architecture hybride (IA symbolique avec des systèmes à base de règles conjuguée à une IA connexionniste basée sur des réseaux de neurones).

En décembre 2023, un autre modèle d'IA de DeepMind – **FunSearch** (appelé ainsi car il recherche des fonctions mathématiques, pas parce qu'il serait drôle) – est parvenu à **résoudre un problème mathématique jusqu'alors non résolu**. Cette réussite est d'autant plus marquante que la solution n'était absolument pas présente dans les données d'entraînement<sup>3</sup>.

---

<sup>1</sup> La question posée relative à la détection de rayonnements avec un bolomètre nécessitait, avant même de répondre, de comprendre un problème de physique avec un diagramme complexe écrit en français. Cf. le rapport technique 2023 sur GPT-4 réalisé par OpenAI : <https://cdn.openai.com/papers/gpt-4.pdf>

<sup>2</sup> Cf. le commentaire Dan Hendrycks à propos du graphique d'OpenAI sur les capacités de leur nouveau système en écriture, mathématiques, raisonnement logique et codage dans ce tweet : <https://x.com/DanHendrycks/status/1778588753721950514>

<sup>3</sup> Cf. Will Douglas Heaven, décembre 2023, « Google DeepMind used a large language model to solve an unsolved math problem », MIT Technology Review : <https://www.technologyreview.com/2023/12/14/1085318/google-deepmind-large-language-model-solve-unsolvable-math-problem-cap-set/>

Selon plusieurs chercheurs, cette expérience s'apparente au premier cas où un grand modèle de langage utilisé pour résoudre un problème mathématique et son succès plaide en faveur de l'IA<sup>1</sup>.

Avant cette résolution d'un problème mathématique, un autre modèle d'IA de DeepMind – AlphaFold – s'est illustré avec sa capacité à produire une analyse décisive du repliement des protéines.

## 2. Les cas emblématiques de l'analyse du repliement des protéines en 2018 et de la génomique en 2024

L'un des premiers domaines scientifiques dans lequel l'intelligence artificielle s'est révélée au grand public est celui de l'analyse du **repliement des protéines**. Les protéines sont des molécules de grande taille composées de polypeptides, eux-mêmes composés d'acides aminés. Il existe vingt acides aminés différents qui peuvent constituer des polypeptides lorsqu'ils s'assemblent en chaîne. En connaissant les acides aminés composant les chaînes polypeptidiques, il est possible de déterminer la « séquence primaire » d'une protéine.

La forme d'une protéine est complexe à déterminer puisqu'elle dépend d'interactions entre les différents éléments des chaînes polypeptidiques de la protéine. Le repliement des protéines est le résultat d'interactions chimiques au sein de la molécule (liaisons hydrogène, interactions hydrophobes, forces de van der Waals, etc.). Pour connaître la forme d'une protéine, il faut déterminer la configuration dans laquelle l'énergie libre de la molécule est la plus basse afin d'atteindre une conformation stable, fonctionnelle et thermodynamiquement favorable. Trouver une telle configuration n'est pas chose aisée alors que ce sont des centaines voire des milliers d'atomes qui sont en interaction en son sein. Par exemple la **formule chimique de l'insuline humaine est  $C_{257}H_{383}N_{65}O_{77}S_6$  soit 788 atomes**.

Utiliser des moyens computationnels d'IA pour parvenir à connaître la forme d'une protéine à partir de sa structure primaire permet de mieux comprendre les propriétés d'une protéine, ce qui revêt une **importance capitale en médecine et en biologie**. Le *Critical Assessment of protein Structure Prediction (CASP)* est un concours biennuel organisé depuis 1994 par le *Protein Structure Prediction Center (PSPC)*<sup>2</sup>, faisant se confronter des modèles sur des tâches de prédiction de structure protéique. Dans ce concours, les organisateurs collectent la structure en trois dimensions de séquences de protéines qui ont été déterminées mais qui n'ont pas fait l'objet d'une publication scientifique. On communique au participant la structure de séquences cibles et ils doivent alors trouver la structure en trois dimensions de la protéine associée à cette structure.

---

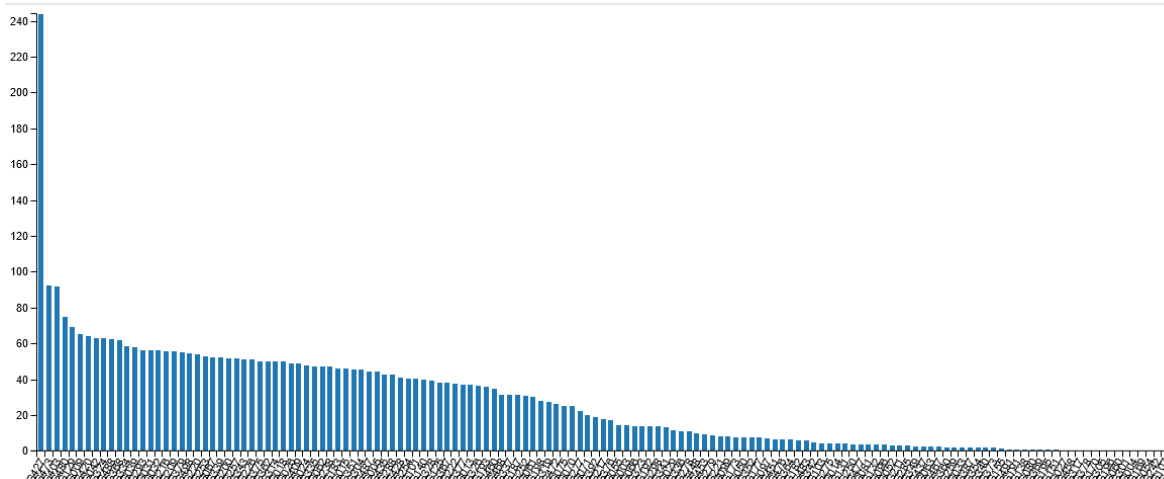
<sup>1</sup> Cf. Bernardino Romera-Paredes et al., 2024, « Mathematical discoveries from program search with large language models », Nature n° 625, 468–75 :

<https://doi-org.stanford.idm.oclc.org/10.1038/s41586-023-06924-6>

<sup>2</sup> Pour en savoir plus, le site du PSPC : <https://predictioncenter.org/>

En **2018**, la société **DeepMind** participe au CASP13 avec son logiciel d'intelligence artificielle **AlphaFold** et remporte le concours. Ils participent au concours deux ans plus tard avec AlphaFold2 et arrivent à nouveau en tête, encore plus largement<sup>1</sup>. C'est cette **large domination du CASP14 en 2020** qui va mettre en lumière les performances des modèles d'apprentissage profond pour le cas de la prédiction de la structure des protéines.

### Mesure de la performance d'Alphafold par rapport aux autres modèles



N.B. : La mesure de la performance d'AlphaFold2 est la barre bleue la plus à gauche, dominant largement tous les autres modèles participant au concours CASP14 en 2020.

Source : PSPC

Ce sont ces travaux qui ont conduit David Baker, John Jumper et Demis Hassabis, fondateur et directeur du laboratoire d'IA Google DeepMind, à se voir décerner en **2024** le **prix Nobel de chimie**.

Dans le rapport de l'Office de 2017, la perspective d'une prédiction optimisée du repliement des protéines par l'IA avait été dessinée en soulignant l'enjeu qu'un tel progrès représenterait pour la médecine<sup>2</sup>. Le rapport précisait que « *le processus physique, par lequel un polypeptide se replie dans sa structure tridimensionnelle caractéristique dans laquelle il est fonctionnel, est important en ce que de nombreuses maladies, en particulier les maladies neurodégénératives, sont considérées comme résultant d'une accumulation de protéines mal repliées* ».

En novembre 2024, l'analyse génomique connaît un **progrès important avec le lancement du premier modèle d'IA dédié, nommé Evo**. Formé par des données de 2,7 millions de génomes de procaryotes et de phages diversifiés sur le plan de l'évolution, sur le modèle des LLM, ce modèle est **capable d'interpréter et de générer de longues séquences génomiques à grande échelle**. Cela représente un grand potentiel pour **interpréter** les

<sup>1</sup> Résultats du CASP14 : [https://predictioncenter.org/casp14/zscores\\_final.cgi](https://predictioncenter.org/casp14/zscores_final.cgi)

<sup>2</sup> Cf. le rapport précité de l'OPECST, op. cit., pp. 77 et 78.

données de séquences biologiques (l'ADN, l'ARN, les protéines...) et réaliser des **prédictions** (comment de petits changements d'ADN affectent la forme physique d'un organisme, générer des séquences réalistes de la longueur du génome), voire **concevoir de nouveaux systèmes biologiques**, y compris avec la validation en laboratoire des technologies synthétiques CRISPR et des éléments transposables de type IS200/IS605.

**La revue *Science* a consacré sa couverture à ce nouveau modèle dans un numéro spécial du 15 novembre 2024** contenant l'article scientifique qui présente cette révolution de la compréhension complète du code génétique jusqu'aux génomes entiers, une avancée majeure dans la capacité à **concevoir la biologie et les sciences du vivant selon des modalités et des échelles multiples et complexes jusqu'alors impossibles**<sup>1</sup>.

### 3. Les jumeaux numériques et le perfectionnement des simulations

Un **jumeau numérique** (parfois désigné par son terme anglais « *digital twin* ») est une **représentation virtuelle d'un objet, bâtiment, processus ou système**<sup>2</sup>. Il s'agit d'un perfectionnement des simulations.

Cette pratique est utilisée depuis le début des années 2000 et se développe particulièrement depuis les années 2010. Elle permet d'avoir un **modèle informatique d'objets ou de processus existants, qui donne ainsi la possibilité de l'étudier dans un environnement numérique et de réaliser des simulations en agrégeant des données concernant l'objet**.

Ces jumeaux numériques peuvent être **créés et animés par de l'intelligence artificielle**, les données récoltées sur le système réel servent alors à entraîner le jumeau pour qu'il parvienne à imiter au mieux son comportement. De tels jumeaux numériques sont développés par des entreprises comme Dassault Systèmes ou des laboratoires comme le CEA-List (Laboratoire d'intégration des systèmes), spécialisé dans les systèmes numériques intelligents. Les **applications** sont **innombrables**.

---

<sup>1</sup> Eric Nguyen et al., 15 novembre 2024, « *Sequence modeling and design from molecular to genome scale with Evo* », in *Science*, volume 386, n° 67 : <https://www.science.org/doi/10.1126/science.ado9336>

<sup>2</sup> Mengnan Liu, Shuiliang Fang, Huiyue Dong, et Cunzhi Xu, 2021, « *Review of digital twin about concepts, technologies, and industrial applications* », *Journal of manufacturing systems* : [https://www.researchgate.net/profile/Mengnan-Liu/publication/342807853\\_Review\\_of\\_digital\\_twin\\_about\\_concepts\\_technologies\\_and\\_industrial\\_applications/links/5fa7f90e92851cc286a04b21/Review-of-digital-twin-about-concepts-technologies-and-industrial-applications.pdf](https://www.researchgate.net/profile/Mengnan-Liu/publication/342807853_Review_of_digital_twin_about_concepts_technologies_and_industrial_applications/links/5fa7f90e92851cc286a04b21/Review-of-digital-twin-about-concepts-technologies-and-industrial-applications.pdf).

#### **4. Adapter nos politiques de recherche aux perspectives ouvertes par l'IA**

Ces progrès grâce à l'IA nécessitent d'adapter nos politiques de recherche qui doivent **tirer le meilleur parti de ces technologies**. Non seulement il faut chercher à **déployer des outils d'IA dans tous les champs de recherche**, mais il faut que **la recherche fondamentale en IA elle-même se nourrisse des autres disciplines**.

Il faut donc envisager un **nouveau paradigme dans le rapport au savoir** et une **politique de recherche** plus que jamais **pluridisciplinaire** voire, mieux, **transdisciplinaire**. Cet objectif global pourra se décliner dans des dispositifs plus précis.



## TROISIÈME PARTIE LA GOUVERNANCE ET LA RÉGULATION DE L'INTELLIGENCE ARTIFICIELLE

Dès le début de l'année 2017, le rapport de l'Office affirmait que « *ces technologies doivent être maîtrisées, utiles et faire l'objet d'usages conformes à nos valeurs humanistes* »<sup>1</sup>. Sa réflexion avait largement porté sur les questions éthiques, de l'amont avec l'éthique de la recherche en intelligence artificielle jusqu'à l'aval avec l'éthique pratique des robots intelligents, en passant par la revue de l'ensemble des démarches engagées en la matière. Le rapport appelait à « *ne pas céder à la tentation de définir un cadre juridique trop contraignant, qui aurait pour inconvénient de figer des règles codifiant des préceptes moraux et, partant, de gêner et de ralentir l'innovation* ». Les rapporteurs, au terme de leurs investigations, n'avaient pas été convaincus de l'urgence d'une intervention législative ou réglementaire en matière d'intelligence artificielle, surtout en raison de son caractère très évolutif. Une proposition de loi leur paraissait inopportune.

Cependant, le rapport affirmait également qu'un « *régime du type de celui qui est applicable aux médicaments avant autorisation de mise sur le marché, avec une période de tests et d'observations, pourrait devenir obligatoire pour les systèmes autonomes d'intelligences artificielles, au stade où leur commercialisation massive sera envisagée* »<sup>2</sup>. Cette approche semble bien être celle retenue par l'Union européenne en 2024, à l'heure du déploiement massif de l'IA générative.

Qu'il s'agisse de dispositions contraignantes ou de simples guides éthiques, **l'objectif reste le plus souvent le même : maîtriser l'IA pour que se déploient les technologies les plus sûres et responsables possibles**, dans la confiance et le respect du droit et des libertés individuelles.

Selon Bertrand Braunschweig, coordinateur scientifique du programme Confiance.ai et ancien coordinateur pour la recherche du programme national d'intelligence artificielle, entendu par vos rapporteurs, le sujet de la confiance dans l'IA est un **aspect fondamental** pour les développeurs comme pour les pouvoirs publics ainsi qu'une **condition essentielle à un déploiement fluide et efficace des technologies**.

---

<sup>1</sup> Cette formulation est tirée du rapport de l'OPECST « Pour une intelligence artificielle maîtrisée, utile et démystifiée » en sa page 26. Ses références ont déjà été données.

<sup>2</sup> Extrait du rapport op. cit., page 144.

Le dernier rapport de la Cnil énonçant ses **recommandations pour le déploiement des systèmes d'IA** a été publié en avril 2024 à la suite d'une vaste consultation publique<sup>1</sup>. Il s'intéresse directement à ce sujet. Vos rapporteurs y renvoient donc pour les **réponses concrètes à apporter aux questions à la fois juridiques et techniques posées par le déploiement des systèmes d'IA**. Il s'agit d'évolutions importantes pour aller vers **une IA de confiance**.

Le fait d'**adopter des dispositifs juridiques encadrant l'IA ne va pas de soi**. Beaucoup, notamment dans le monde économique, croient d'abord à **l'autorégulation**. Le *Partnership on AI* créé en 2016<sup>2</sup> ou, plus récemment, le *Frontier Model Forum*, partenariat créé en 2023 par Microsoft, OpenAI, Google et Anthropic, se proposent de définir les conditions du développement d'IA sûres et responsables. Une telle option pose néanmoins le problème des pratiques réelles des entreprises qui, en l'absence de contrôles, risquent fort de se concentrer sur leurs objectifs fondamentaux, à savoir enregistrer des profits et gagner des parts de marché.

Certains prônent un **encadrement vertical** par la législation, ce qui fait courir le risque de politiques publiques lourdes et mal ciblées, **sans pour autant éviter les problèmes de « capture du régulateur »** par les entreprises en raison de leur quasi-monopole d'expertise technique ou en tout cas de leur meilleure connaissance des modèles d'IA qu'elles développent et/ou déploient.

Cette régulation peut en pratique **prendre des formes multiples** :

- se baser sur des **principes** ou sur des **niveaux de risque** ;
- être tournée vers les **technologies**, les **applications** ou les **usages** ;
- laisser un **accès libre au marché** avec de simples déclarations et des autoévaluations ou obliger à **communiquer des descriptions techniques** (au régulateur voire au grand public) ou, encore, fixer un **régime d'autorisation préalable**, avec le cas échéant la création d'une **agence de régulation**.

Entre ces deux pôles, se situent les solutions de **corégulation**, qui associent les entreprises à la définition des normes, à leur mise en œuvre, voire à leur contrôle.

Ce sont, de fait, les perspectives qui se dessinent dans le monde occidental, même si **l'Union européenne va un peu plus loin et se rapproche d'une régulation verticale** avec son règlement de 2024 sur l'IA, dit *AI Act*, dont les dispositions complexes seront présentées plus loin.

---

<sup>1</sup> Cf. la page de présentation du rapport qui contient le lien vers le document complet :

<https://www.cnil.fr/fr/ia-la-cnil-publie-ses-premieres-recommandations-sur-le-developpement-des-systemes-dintelligence>

<sup>2</sup> Fondée par Amazon, Facebook, Google, DeepMind, Microsoft et IBM, rejoints par Apple en 2017, cette association regroupe une centaine de structures, non seulement des entreprises mais aussi des associations et des organismes du monde de la recherche.

## I. LES DISPOSITIFS NATIONAUX OU RÉGIONAUX

### A. UNE POLITIQUE FRANÇAISE DE L'INTELLIGENCE ARTIFICIELLE EN DEMI-TEINTE

#### 1. La stratégie nationale en faveur de l'IA depuis 2017 : un retard à l'allumage

**Des stratégies nationales ambitieuses pour le secteur numérique ont déjà été définies par le passé.** Le « Plan Calcul », lancé en 1966, le rapport sur l'informatisation de la société publié en 1978, qui a inventé le concept de télématique et proposé le lancement du réseau Minitel<sup>1</sup>, le plan « Informatique pour tous » impulsé en 1985, le rapport sur les autoroutes de l'information en 1994 ou encore le programme d'action gouvernemental pour la société de l'information (PAGSI) et les espaces publics numériques (EPN) en 2000<sup>2</sup>, en sont différents exemples qui ont souvent laissé des souvenirs doux-amers en raison de leurs résultats qui n'étaient pas à la hauteur des espoirs suscités.

En France, le pouvoir exécutif a depuis 2017, sans directement réguler l'intelligence artificielle, annoncé des **stratégies nationales pour l'IA** et mis en place une série de mesures concernant **la gouvernance de l'IA**. Refaire l'histoire de cette succession d'annonces ou de mesures dans le domaine des technologies d'intelligence artificielle est indispensable avant de pouvoir dessiner des perspectives pour le futur.

Une **première stratégie nationale pour l'IA** a été voulue par le Président de la République, alors François Hollande, dès janvier 2017, en écho à la stratégie américaine dévoilée en octobre 2016 par le président américain Barack Obama.

---

<sup>1</sup> Ce rapport, remis au Président de la République Valéry Giscard d'Estaing en décembre 1977 par Simon Nora et Alain Minc, préconisait de manière audacieuse d'associer les télécommunications et l'informatique grâce à la connexion de terminaux informatiques permettant la visualisation et l'échange, à travers les réseaux de télécommunication, de données stockées dans des ordinateurs. Il a inventé le concept de télématique et a proposé le lancement du réseau Minitel, exactement seize ans après qu'un chercheur du Massachusetts Institute of Technology (MIT), Joseph Carl Robnett Licklider, eut rédigé les premiers textes décrivant les interactions sociales rendues possibles par l'intermédiaire d'un réseau d'ordinateurs alors baptisé « réseau galactique » qu'il développa dans les années 1970 pour le gouvernement américain sous le nom d'Arpanet (ou réseau de l'Arpa, du nom de la structure ayant précédé la Darpa) et ayant conduit à Internet après l'adoption du protocole TCP/IP en 1983.

<sup>2</sup> Ce programme reposait sur le Comité interministériel sur la société de l'information, Cf. son rapport préparé en amont du programme : Préparer l'entrée de la France dans la société de l'information, Paris, La Documentation française, 1998. Cf. également la déclaration du 10 juillet 2000 de Lionel Jospin, alors Premier ministre, sur les priorités du gouvernement concernant la société de l'information, notamment la volonté de combler le fossé numérique en offrant à tous la possibilité de s'initier à l'Internet : <https://www.vie-publique.fr/discours/137517-declaration-de-m-lionel-jospin-premier-ministre-sur-les-priorites-du>

La secrétaire d'État chargée du numérique et de l'innovation, Axelle Lemaire, avait ainsi été missionnée pour préparer les détails de ce plan baptisé « **France IA** », lancé dans l'incubateur Agoranov le 20 janvier 2017 et ayant conduit à la remise d'un rapport au Président de la République à la Cité des Sciences et de l'industrie le 21 mars 2017.

Ce plan visait notamment à mettre en place un comité de pilotage « France IA », à financer de nouveaux projets de recherche en IA à travers le programme pour les investissements d'avenir (PIA), à créer un centre interdisciplinaire sur l'IA, à réaliser une cartographie de l'écosystème de l'intelligence artificielle en France en vue de mobiliser les acteurs publics et privés, à réfléchir à des normes et des standards, et à suivre l'impact de l'IA sur l'économie et la société.

En parallèle de ce plan gouvernemental, dès le printemps 2016, la sénatrice Dominique Gillot et le député Claude de Ganay ont préparé le premier rapport de l'OPECST sur l'intelligence artificielle. Publié le 15 mars 2017, ce rapport présentait 15 propositions au nom de l'Office. Ses auteurs y regrettaient qu'en raison de l'actualité, le plan France IA **arrive trop tard pour pouvoir être réellement pris en compte dans les politiques publiques**<sup>1</sup>.

Le plan France IA du Gouvernement a **ensuite été suspendu puis a été purement et simplement enterré** : il n'a pu en effet être mis en œuvre dans le contexte des élections présidentielles de 2017 et de l'élection d'Emmanuel Macron à la tête de l'État.

Ce dernier a préféré **retravailler une nouvelle stratégie** pour l'intelligence artificielle durant son premier mandat, qu'il annoncerait lui-même : un an plus tard c'est avec la **remise d'un autre rapport<sup>2</sup> émanant du président de l'Office, notre ancien collègue député Cédric Villani**, alors en mission pour l'exécutif, que le Président de la République a pu proposer au Collège de France, le 29 mars 2018, en présence de la ministre allemande de la recherche et du commissaire européen à l'innovation, une « Stratégie nationale et européenne pour l'intelligence artificielle », lors d'un événement baptisé « AI for Humanity »<sup>3</sup>.

---

<sup>1</sup> Cf. le rapport de l'OPECST op. cit. page 142.

<sup>2</sup> Le rapport abordait successivement différentes facettes de l'IA : politique économique, recherche, emploi, éthique, cohésion sociale. Et cinq annexes insistaient sur des domaines d'intérêt particulier : éducation, santé, agriculture, transport, défense et sécurité. Rapport de Cédric Villani « Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne », 2018, <https://www.vie-publique.fr/rapport/37225-donner-un-sens-lintelligence-artificielle-pour-une-strategie-nation>

<sup>3</sup> L'expression AI for Humanity sera utilisée par EDF, Thalès et Total pour nommer en 2020 leur premier laboratoire industriel conjoint de recherche. Rapidement débaptisé, il s'intitule désormais Sinclair (pour Saclay INdustrial Collaborative Laboratory for Artificial Intelligence Research).

Rétrospectivement, cette approche semblait se **préoccuper insuffisamment de l'innovation technique en IA**, le rapport de notre ancien collègue était d'ailleurs **exclusivement tourné vers les technologies de Deep Learning en vogue à l'époque** et appelait essentiellement à **centrer les efforts de notre stratégie en matière d'intelligence artificielle sur leur déploiement dans certains domaines d'application comme les transports, la santé, la sécurité, l'environnement et la défense**. En matière de santé, il proposait par exemple d'interconnecter les fichiers déjà existants pour développer ce qui est devenu le *Health Data Hub*.

Avec un financement pluriannuel d'**1,5 milliard d'euros annoncé**, la stratégie nationale pour l'IA ambitionnait de faire de la France un des leaders mondiaux de l'intelligence artificielle. Or, elle a surtout consisté dans la labellisation de **quatre instituts interdisciplinaires d'intelligence artificielle (3IA)**, le **financement de chaires et de doctorats** ou encore l'investissement indispensable dans des **infrastructures de calcul** à travers des supercalculateurs gérés par le GENCI (pour Grand équipement national de calcul intensif) comme Jean Zay, inauguré en 2019, ou Adastra, inauguré en 2023, et dont les performances atteignent respectivement 36,85 pétaflops et 74 pétaflops (Jean Zay devrait toutefois atteindre 125,9 pétaflops cette année). Pour mémoire un pétaflop représente un million de milliards de calculs d'opérations en virgule flottante par seconde.

Il faut noter que ces deux supercalculateurs n'étant plus au niveau exascale des standards internationaux, c'est-à-dire dépassant **un exaflop par seconde** (un milliard de milliards de calculs par seconde), la France va héberger un supercalculateur européen (qui devait s'appeler Jules Verne mais portera finalement le nom de la chercheuse oubliée Alice Recoque) dont elle devra partager l'usage avec l'Union européenne<sup>1</sup>. À titre de comparaison, l'entreprise d'Elon Musk spécialisée en IA, appelée « xAI » et qui développe le système Grok, s'est engagée à acheter entre 3 et 4 milliards de dollars de GPU à Nvidia et s'est dotée d'un **supercalculateur** de 150 MW du nom de Colossus, construit en 19 jours<sup>2</sup>, développant théoriquement **3,4 exaflops**

---

<sup>1</sup> L'entreprise commune européenne pour le calcul à haute performance, ou European High-Performance Computing Joint Undertaking (EuroHPC JU), a sélectionné en 2022 le Forschungszentrum Jülich, pour exploiter en Allemagne Jupiter (pour Joint Undertaking Pioneer for Innovative and Transformative Exascale Research), le premier supercalculateur européen exascale. Il a ensuite sélectionné en 2023 le consortium franco-néerlandais Jules Verne pour un second supercalculateur exascale qui sera hébergé par le CEA, à Bruyères-le-Châtel, près d'Arpajon dans l'Essonne et devait prendre le nom du consortium. Enfin, en 2024 le Gouvernement et le GENCI annoncent que le supercalculateur prendra finalement le nom de la chercheuse en informatique Alice Recoque.

<sup>2</sup> Sa construction complète a pris 122 jours en incluant la fabrication des processeurs mais entre la livraison des processeurs à Memphis, Tennessee, et l'entraînement des modèles il n'a fallu que 19 jours. Cf. le communiqué de Nvidia : <https://nvidianews.nvidia.com/news/spectrum-x-ethernet-networking-xai-colossus?ncid=so-twit-249598> ainsi que l'interview du PDG de Nvidia Jensen Huang expliquant que « building xAI's computing supercluster in just 19 days, from concept to training, was a superhuman effort and the only person in the world who could have done it is Elon Musk » : <https://x.com/tsarnick/status/1845528355833319769>

par seconde, car composé de **100 000 processeurs Nvidia Hopper 100**<sup>1</sup>. Sa taille devrait doubler d'ici quelques mois pour atteindre **200 000 processeurs**<sup>2</sup>. Pour mémoire, Google, OpenAI, Microsoft, Meta et Nvidia étaient les seules entreprises au monde à dépasser les 50 000 GPU et xAI vient de repousser la frontière du concevable avec la perspective de cette structure de 200 000 processeurs GPU. Les puces utilisées, les Nvidia Hopper H100, ont un coût unitaire qui varie de 30 000 à 70 000 dollars, un coût du même ordre que la nouvelle génération de puces Nvidia, les Blackwell B200, dont vos rapporteurs ont pu voir le format réduit au siège de Nvidia, et qui seront commercialisées prochainement<sup>3</sup>. Les 200 000 processeurs de xAI ont une valeur comprise **entre 6 et 14 milliards de dollars** environ. Le supercalculateur Jean Zay, après son extension prévue d'ici la fin de l'année 2024, sera quant à lui doté de **1 456 GPU Nvidia H100**<sup>4</sup>.

La stratégie nationale pour l'intelligence artificielle, divisée en deux phases<sup>5</sup>, a également conduit à une **coordination qui n'a d'interministérielle que le nom** avec un coordinateur rattaché initialement à la direction interministérielle du numérique et du système d'information et de communication (DINSIC), puis à la direction générale des entreprises (DGE) du ministère de l'économie, sans autorité réelle sur les conditions de mise en œuvre de la stratégie.

---

<sup>1</sup> Cf. cet article du 9 septembre 2024 qui explique que « the biggest AI computer the world has ever seen (and it's not even close) boasts some of the most astonishing numbers one can fathom and helps us toy with the idea of how large the next generation of models will be » : <https://pub.towardsai.net/putting-the-worlds-largest-ai-supercomputer-into-perspective-60afde9bc653> et cet article du 4 septembre 2024 « xAI has apparently completed the world's fastest supercomputer » : <https://www.heise.de/en/news/xAI-has-apparently-completed-the-world-s-fastest-supercomputer-9857540.html>

<sup>2</sup> Cf. la déclaration d'Elon Musk le 28 octobre 2024 sur X : <https://x.com/elonmusk/status/1850991323010261230> et surtout le communiqué assez complet de Nvidia du même jour : <https://nvidianews.nvidia.com/news/spectrum-x-ethernet-networking-xai-colossus?>

<sup>3</sup> Le prix de 70 000 dollars est évalué par HSBC, cf. <https://www.lebigdata.fr/une-puce-ia-a-70-000-le-prix-des-nvidia-blackwell-fait-trembler-lindustrie> en dépit des déclarations du PDG de Nvidia qui parle, lui, d'un coût de 30 000 à 40 000 dollars pièce : [https://next.ink/brief\\_article/le-gpu-blackwell-b200-coutera-entre-30-000-et-40-000-dollars-piece/](https://next.ink/brief_article/le-gpu-blackwell-b200-coutera-entre-30-000-et-40-000-dollars-piece/)

<sup>4</sup> Il faut ajouter à ces 1 456 GPU H100 des processeurs d'ancienne génération, comme les 416 GPU A100 et les 1 832 GPU V100. Cf. <https://www.cnrs.fr/fr/presse/genci-et-le-cnrs-choisissent-eviden-pour-faire-du-supercalculateur-jean-zay-lune-des>

<sup>5</sup> La première phase (2018-2022) visait à doter la France de capacités de recherche compétitives. Cette première étape a été financée à hauteur de 1,5 milliard d'euros. Elle a notamment permis la création et le développement du réseau d'instituts interdisciplinaires d'intelligence artificielle (3IA), la mise en place de chaires d'excellence et de programmes doctoraux, ainsi que le supercalculateur Jean Zay. La deuxième phase (2021-2025) cherche à diffuser les technologies d'IA au sein de l'économie et à soutenir le développement et l'innovation dans des domaines prioritaires comme l'IA embarquée, l'IA de confiance, l'IA frugale et l'IA générative. Cette seconde phase devrait être dotée de 2,2 milliards d'euros, notamment dans le cadre de France 2030, et s'articule autour de trois piliers : le soutien à l'offre deep tech, la formation et l'attraction des talents, le rapprochement de l'offre et de la demande de solutions en IA, cf. <https://www.entreprises.gouv.fr/fr/numerique/enjeux/la-strategie-nationale-pour-l-ia>



L'**instabilité** du titulaire de cette fonction et les **vacances** répétées du poste sont également frappantes : Bertrand Pailhès nommé en juillet 2018, quitte ses fonctions un an plus tard en novembre 2019, il n'est remplacé par Renaud Vedel qu'en mars 2020, qui ne reste chargé de cette mission que jusqu'à l'été 2022, Guillaume Avrin ne rejoint la structure qu'en janvier 2023, à nouveau après une vacance de six mois. Cette instabilité et ces vacances révèlent **une fonction problématique et mal définie**.

Au total, le **pilotage de la stratégie nationale en IA reste toujours défaillant** : elle demeure d'ailleurs en réalité toujours sans pilote, évoluant au gré des annonces du Président de la République, à l'instar de son courrier du 25 mars 2024 faisant d'Anne Bouverot « *l'envoyée spéciale du Président de la République* » pour le futur sommet sur l'IA ou de son discours lors du rassemblement des plus grands talents français de l'IA à l'Élysée le 21 mai 2024, qui présentait les nouveaux cinq grands domaines de la stratégie nationale pour l'IA.

## **2. Le bilan critique de la stratégie et son évaluation par la Cour des comptes**

La première étape de la stratégie nationale pour l'IA mise en place par l'exécutif à partir de 2018 a déjà pu être **évaluée** et les effets produits par les mesures prises ont pu être comparés à leurs effets escomptés.

La **Cour des comptes** a ainsi publié, en avril 2023, un rapport<sup>1</sup> qui dresse le bilan de la stratégie mise en place par l'exécutif concernant l'intelligence artificielle. Elle évalue en particulier les mesures prises et leurs effets par rapport aux objectifs fixés par le Président de la République lors de la mise en place de ce plan : positionner la France parmi les cinq meilleurs pays en termes d'IA, et devenir les chefs de file européens dans le domaine.

L'**évaluation est plutôt mitigée**. En effet, **loin d'atteindre les objectifs affichés, comme celui de devenir un leader mondial**, il semble que la stratégie nationale pour l'IA ait surtout permis à la France de **ne pas décrocher davantage** dans la compétition au niveau mondial, au moins dans les quelques domaines investis par le plan.

La Cour des comptes affirme ainsi que la priorité donnée à la recherche en IA a permis à la France de **maintenir un niveau honorable en termes de publication d'articles scientifiques** et d'efficacité de la recherche.

---

<sup>1</sup> Cf. le rapport de la Cour des comptes, 2023, « La stratégie nationale de recherche en intelligence artificielle » : <https://www.ccomptes.fr/fr/publications/la-strategie-nationale-de-recherche-en-intelligence-artificielle>



Ce **résultat est cependant à relativiser** puisque lorsque ces classements sont rapportés au PIB, la position de la France chute au 44<sup>e</sup> rang : *« Si la France apparaît assez performante en matière de recherche en IA (10<sup>e</sup> rang mondial et 2<sup>e</sup> rang européen en 2021 en nombre de publications en IA sur un total de 47 pays comparés), sa recherche dans ce domaine apparaît peu efficiente au regard du produit intérieur brut de la France (44<sup>e</sup> rang mondial et 25<sup>e</sup> européen en 2021 pour le même critère rapporté au PIB)<sup>1</sup> ».*

La mise en place des **Instituts 3IA** est jugée plutôt efficace par le rapport, qui constate que ces instituts contribuent à l'augmentation du nombre de publications scientifiques et de coopérations internationales en matière d'IA. Ceci est permis grâce à des **collaborations interdisciplinaires** et à l'acquisition de **ressources technologiques** avancées. Un exemple est l'acquisition de nœud de clusters GPU par le 3IA de Nice Sophia-Antipolis pour renforcer ses capacités de calcul ; il est mis à disposition en priorité pour les chaires 3IA.

Le rapport note toutefois que **les résultats des Instituts 3IA ne sont pas à la hauteur des attentes initiales**. D'après la Cour des comptes, la cartographie des formations en IA, loin d'être homogène, est **complexe et peu lisible**. De plus, les instituts 3IA ne sont **pas autonomes** et apparaissent **précaires** : la pérennité des financements de l'État est cruciale pour maintenir la dynamique de ces instituts, or la visibilité des perspectives financières est très réduite.

En raison de ce contexte global, le rapport juge que la **formation des talents** en IA, même si elle a progressé, reste **encore très insuffisante** pour combler le déficit de compétences en intelligence artificielle dans le pays. Vos rapporteurs ajoutent que **rien n'a non plus été fait pour lutter contre la fuite des cerveaux**.

S'agissant de la gouvernance et des investissements en IA, la **complexité de l'écosystème d'acteurs et des outils de financement public** constitue un frein à l'efficacité des mesures voulues par l'exécutif, et ce, **malgré la présence du coordinateur national à l'intelligence artificielle** (aujourd'hui Guillaume Avrin). Cette coordination **ne joue pas un rôle assez important** et n'est pas assez financée selon la Cour.

Ce déficit de cohérence dans la gouvernance française de sa stratégie en IA crée des **difficultés pour attirer et concentrer les investissements**. Aussi, bien que les sommes investies représentent des montants importants, la **dispersion des moyens dans un saupoudrage peu rationnel** empêche ces investissements d'être pleinement efficaces.

---

<sup>1</sup> Op. cit. p. 47.

Aussi, la Cour des comptes préconise l'élaboration d'une **nouvelle politique publique globale en matière d'IA**, au moins pour assurer une bonne **gouvernance** et une **coordination de l'ensemble des dispositifs de financement**.

### 3. Les perspectives de relance de la politique nationale de l'IA

La dernière étape concernant la gouvernance française de l'intelligence artificielle a été la **remise au Gouvernement du rapport « IA : notre ambition pour la France » le 13 mars 2024<sup>1</sup>**.

Ce rapport a été rédigé par la Commission de l'intelligence artificielle dont les présidents étaient Anne Bouverot et Philippe Aghion, rencontrés par vos rapporteurs, tout comme l'ont été les deux rapporteurs généraux du rapport, Cyprien Canivenc et Arno Amabile.

Ce rapport important contient **25 recommandations pour une politique ambitieuse de la France en matière d'IA**, avec un investissement d'un peu plus de cinq milliards d'euros par an sur cinq ans. Ces dépenses évaluées à **27 milliards d'euros au total** sont récapitulées ci-après.

**Tableau récapitulatif des recommandations de la Commission de l'intelligence artificielle**

Recommandations		Coût estimé sur cinq ans
1	Créer les conditions d'une appropriation collective de l'IA et de ses enjeux afin de définir collectivement les conditions dans lesquelles elle s'insère dans notre société et nos vies quotidiennes	10 millions d'euros
2	Investir dans l'observation, les études et la recherche sur les impacts des systèmes d'IA sur la quantité et la qualité de l'emploi	5 millions d'euros
3	Faire du dialogue social et professionnel un outil de co-construction des usages et de régulation des risques des systèmes d'IA	-
4	Porter une stratégie de soutien à l'écosystème d'IA ouverte au niveau international en soutenant l'utilisation et le développement de systèmes d'IA ouverts et les capacités d'inspection et d'évaluation par des tiers	-
5	Faire de la France un pionnier de l'IA en renforçant la transparence environnementale, la recherche dans des modèles à faible impact, et l'utilisation de l'IA au service des transitions énergétique et environnementale	100 millions d'euros
6	Généraliser le déploiement de l'IA dans toutes les formations d'enseignement supérieur et acculturer les élèves dans l'enseignement secondaire pour rendre accessibles et attractives les formations spécialisées	1,2 milliard d'euros

<sup>1</sup> Cf. la présentation du rapport de la Commission de l'intelligence artificielle présidée par Anne Bouverot et Philippe Aghion : <https://www.info.gouv.fr/actualite/25-recommandations-pour-lia-en-france>

<b>Recommandations</b>		<b>Coût estimé sur cinq ans</b>
7	Investir dans la formation professionnelle continue des travailleurs et dans les dispositifs de formation autour de l'IA	200 millions d'euros
8	Former les professions créatives à l'IA, dès les premières années de l'enseignement supérieur et en continu	20 millions d'euros
9	Renforcer la capacité technique et l'infrastructure du numérique public afin de définir et de passer à l'échelle une réelle transformation des services publics grâce au numérique et à l'IA, pour les agents et au service des usagers	5,5 milliards d'euros
10	Faciliter la circulation des données et le partage de pratiques pour tirer les bénéfices de l'IA dans les soins, améliorer l'offre et le quotidien des soignants	3 milliards d'euros
11	Encourager l'utilisation individuelle, l'expérimentation à grande échelle et l'évaluation des outils d'IA pour renforcer le service public de l'éducation et améliorer le quotidien des équipes pédagogiques	1 milliard d'euros
12	Investir massivement dans les entreprises du numérique et la transformation des entreprises pour soutenir l'écosystème français de l'IA et en faire l'un des premiers mondiaux	3,6 milliards d'euros
13	Accélérer l'émergence d'une filière européenne de composants semi-conducteurs adaptés aux systèmes d'IA	7,7 milliards d'euros
14	Faire de la France et de l'Europe un pôle majeur de la puissance de calcul installée	1 milliard d'euros
15	Transformer notre approche de la donnée personnelle pour mieux innover	16 millions d'euros
16	Mettre en place une infrastructure technique favorisant la mise en relation entre les développeurs d'IA et les détenteurs de données culturelles patrimoniales	35 millions d'euros
17	Mettre en œuvre et évaluer les obligations de transparence prévues par le règlement européen sur l'IA en encourageant le développement de standards et d'une infrastructure adaptée	-
18	Attirer et retenir des talents de stature internationale avec des compétences scientifiques ou entrepreneuriales et managériales dans le domaine de l'IA	10 millions d'euros
19	Assumer le principe d'une « Exception IA » sous la forme d'une expérimentation dans la recherche publique pour en renforcer l'attractivité	1,025 milliard d'euros
20	Inciter, faciliter et amplifier le recours aux outils d'IA dans l'économie française en favorisant l'usage de solutions européennes	2,6 milliards d'euros
21	Faciliter l'appropriation et l'accélération des usages de l'IA dans la culture et les médias pour limiter la polarisation entre grands groupes et petits acteurs et lutter contre la désinformation	60 millions d'euros
22	Structurer une initiative diplomatique cohérente et concrète visant la fondation d'une gouvernance mondiale de l'IA	300 millions d'euros

Recommandations		Coût estimé sur cinq ans
23	Structurer dès maintenant un puissant écosystème national de gouvernance de l'IA	5 millions d'euros
24	Doter la France et l'Europe d'un écosystème d'évaluation public et privé des systèmes d'IA au plus proche des usages et des derniers développements technologiques	15 millions d'euros
25	Anticiper les concentrations de marché sur l'ensemble de la chaîne de valeur de l'intelligence artificielle	-
<b>TOTAL</b>		<b>27 milliards d'euros</b>

Les principaux leviers d'action actuels de l'exécutif sont ceux de la **deuxième phase de la stratégie nationale pour l'intelligence artificielle (2021-2025)**, ainsi que les **compléments apportés à cette stratégie par les mesures annoncées** par le Président de la République lors du « rassemblement des plus grands talents français de l'IA » à l'Élysée le 21 mai 2024.

Lors de son audition par vos rapporteurs, Guillaume Avrin, coordinateur national pour l'IA, a détaillé la structure du dispositif mis en place pour favoriser l'IA dans notre pays. Il a précisé que le plan de l'exécutif se découpait en **plusieurs phases**. La première phase, aujourd'hui terminée consistait d'abord à promouvoir une IA d'excellence notamment grâce à la création des Instituts 3IA. Le second volet, en cours de déploiement, censé couvrir la période 2021-2025, consiste à **diffuser l'IA** dans l'économie et la société à l'aide de trois moyens d'investissement : la **commande publique**, les **subventions** et **l'investissement en capital**.

Le développement de cette deuxième phase vise spécifiquement quatre domaines : l'IA **frugale**, soit l'IA compétitive tout en étant écologiquement responsable, l'IA **embarquée**, c'est-à-dire des systèmes pouvant être exécutés localement, l'IA de **confiance** et l'IA **générative**, des aspects thématiques bien plus mis en avant que lors de la première phase. Le souhait de l'exécutif est de **développer l'interdisciplinarité** pour l'IA en promouvant l'initiative « l'IA plus X », c'est-à-dire l'utilisation de l'IA comme outil dans divers secteurs. L'approche applicative semble donc très privilégiée par le gouvernement, ce qui a été confirmé par l'audition de la DGE par vos rapporteurs.

Lors du rassemblement des plus grands talents français de l'IA à l'Élysée le 21 mai 2024, le Président de la République, Emmanuel Macron, a indiqué qu'il souhaitait voir la stratégie en matière d'IA se déployer autour de **cinq grands domaines** : les **talents**, les **infrastructures**, les **usages**, **l'investissement** et la **gouvernance**. Une telle approche est plus satisfaisante qu'une concentration sur les seules applications.

Il a annoncé un plan d'investissement de **400 millions d'euros** pour financer **neuf pôles d'excellence en IA**, comprenant les quatre anciens Instituts 3IA lancés en 2019 (MIAI@Grenoble-Alpes, 3IA Côte d'Azur, PRAIRIE et ANITI) auxquels s'ajoutent désormais SequoIA à Rennes, un projet de l'université de Lorraine, *Hi Paris!* de l'Institut polytechnique de Paris, PostGenAI@Paris de la Sorbonne et DATA IA de l'Université Paris-Saclay, l'objectif étant de passer de 40 000 à 100 000 personnes formées à l'IA par an.

Un autre projet annoncé par le chef de l'État est « **Scribe** » d'une durée prévue de deux ans, inclus dans le plan France 2030 et visant à soutenir l'IA sectorielle. Ce plan vise à encourager le développement de modèles de fondation et des applications sectorielles de l'IA, la création de jeux de données d'alignement sectoriels et le développement d'outils d'évaluation et de sécurité.

Un **nouveau fonds d'investissement** devrait être mis en place, dont un quart sera financé par l'État afin de financer des domaines aujourd'hui moins en vue, voire oubliés, bien que pourtant indispensables à l'IA, comme la **filière des semi-conducteurs et des puces** ainsi que **l'informatique en nuage**. Cet aspect est à suivre de près.

Le Président de la République a dans le même temps annoncé vouloir créer des **fonds d'investissement similaires au niveau européen avec le même objectif de financement de ces secteurs stratégiques**. Il faut espérer qu'une telle démarche ne conduise pas à prendre encore davantage de retard dans la mise en œuvre d'un financement urgent et ciblé en direction de la filière française des semi-conducteurs et des puces ainsi que de l'informatique en nuage.

Le Président de la République a également prévu de doter le **Conseil national du numérique** de 10 millions d'euros supplémentaires pour lui permettre de réaliser sa mission d'acculturation des citoyens à l'IA, notamment à travers les « **Cafés IA** », proposés par la commission de l'intelligence artificielle.

Enfin, l'ouverture du centre d'évaluation en IA au sein du **Laboratoire national de métrologie et d'essai** (LNE) confirme le rôle important joué par cette structure dans le suivi et l'évaluation des modèles d'IA, en lien avec les exigences posées par l'*AI Act* de l'UE en 2024. Ce texte prévoit en effet un cadre assez précis en la matière.

## B. D'AUTRES DISPOSITIFS NATIONAUX DANS L'UNION EUROPÉENNE

Pour établir **une comparaison internationale des dispositifs et des situations nationales en matière d'intelligence artificielle**, vos rapporteurs ont envoyé des questionnaires aux services scientifiques de diverses ambassades présentant un intérêt particulier pour leur travail. Les réponses à ces questionnaires ont permis d'appréhender différents dispositifs nationaux mis en place en Europe. Vos rapporteurs ont enrichi leur réflexion par des déplacements en Belgique et aux Pays-Bas. Enfin, d'autres sources, dont l'étude comparée réalisée par Florence G'sell, auditionnée par vos rapporteurs à l'Université de Stanford<sup>1</sup>, ont été mobilisées.

### 1. L'Allemagne : le pays le plus proche du nôtre

En termes de stratégie pour l'intelligence artificielle, l'Allemagne est le pays dont **le profil se rapproche le plus de celui de la France**. Ainsi, l'Allemagne comme la France font toutes les deux partie, au niveau mondial, certes loin derrière des États-Unis et de la Chine, du club restreint des **pays leaders dans les technologies** d'intelligence artificielle, avec des entreprises de pointe en IA comme Mistral et Aleph-Alpha. Les deux pays se sont d'ailleurs positionnés de façon similaire **contre la régulation des technologies d'IA en elles-mêmes** lors des discussions de l'*AI Act* européen (le texte initial n'envisageait que les risques liés aux usages, et comme il sera vu, des modifications lors de la procédure législative ont conduit à élargir le cadre aux modèles en eux-mêmes, en fonction de leur puissance).

L'Allemagne a mis en place en 2018 une **stratégie nommée « AI made in Germany »** visant au développement d'un réseau fédéral de centres de compétences et de *hubs* de transferts des connaissances des laboratoires aux industries, la création de cent chaires d'IA, le lancement d'un programme de financement pour les universités, les instituts de recherche et les entreprises. Ce plan a également mis en place les trois « *Konrad Zuse school of excellence in AI* », **écoles d'excellence** dans le domaine de l'IA, et créé un cursus IA dans 81 universités du pays. Le gouvernement fédéral a promis d'augmenter les dépenses en IA **de trois à cinq milliards d'euros pour la période 2019-2025** dans le cadre du plan de relance contre la crise liée à la covid-19 (en France, les ordres de grandeur pourraient être du même ordre bien que la commission sur l'IA ait proposé un budget renforcé, de 27 milliards d'euros sur cinq ans). En plus de ce plan, des **ministères** ont également mis en place des **dispositifs plus sectoriels** pour le développement de l'IA au niveau fédéral.

---

<sup>1</sup> Cf. Florence G'Sell, 25 octobre 2024, « *Regulating under Uncertainty: Governance Options for Generative AI* », Stanford Cyber Policy Center, Université de Stanford : [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4918704](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4918704)

Le **ministère de l'éducation et de la recherche** (*Bundesministerium für Bildung und Forschung* ou BMBF) a mis au point l'initiative IA pour les PME ou « *AI for SMEs* » qui a créé huit centres de compétences nommés « *Mittelstand 4.0* » (PME 4.0 en français). Le ministère est également à l'origine d'un plan d'action pour l'IA comportant 70 mesures visant, entre autres, à développer l'IA générative pour le système éducatif, créer 50 chaires en IA, développer les **infrastructures de supercalcul** et mettre en place des centres de services IA pour les PME, pour un budget total de 483,3 millions d'euros pour 2024. Ce ministère est à l'origine de la loi *Forschungsdatengesetz* (FDG) sur les données de la recherche, qui permet aux chercheurs publics et privés d'accéder aux données publiques, de créer un centre de données allemand avec une obligation de cataloguer les données de recherche grâce à des métadonnées pour améliorer leur accessibilité.

Le **ministère de l'économie** et du climat (*Bundesministerium für Wirtschaft und Klimaschutz*, BMWK, ex-BMWi) a lancé plusieurs programmes pour le développement de l'IA dans l'économie. C'est le cas du programme EXIST qui vise à créer des incubateurs de *start-up* en IA comme le *Künstliche Intelligenz Entrepreneurship Zentrum* (KIEZ) de Berlin ou l'AI+MUNICH. Un second projet mis au point par le BMWK est le programme WIPANO de gestion de la propriété intellectuelle, doté d'un budget de 23 millions d'euros de 2016 à 2019 et regroupant les agences de brevet et de valorisation en une alliance technologique nommée *Technologieallianz e.V.*

Le **ministère du travail et des affaires sociales** (*Bundesministerium für Arbeit und Soziales* ou BMAS) est à l'origine d'espaces d'expérimentations de l'IA en entreprise.

L'Allemagne étant un état fédéral, il faut souligner l'importance des **actions régionales** qui sont conduites. En effet, les *Länder* mettent également en place des mesures sur leurs territoires respectifs. La Bavière a, par exemple, publié un plan dénommé « *High-tech agenda plus* », investissant la somme considérable de 360 millions d'euros dans l'IA, pour financer en particulier 112 chaires dédiées à la recherche et à l'enseignement de l'IA.

## 2. L'Italie : une stratégie de soutien et de vigilance

L'Italie a adopté une **stratégie nationale sur l'IA en 2021** autour de six objectifs : renforcer la recherche fondamentale en IA ; réduire la fragmentation de la recherche en IA ; développer et adopter une IA anthropocentrique et fiable ; augmenter le développement des technologies basées sur l'IA ; développer des services basés sur l'IA dans le secteur public ; former, retenir et attirer des chercheurs en IA. Le 12 mars 2024, la présidente du Conseil des ministres Giorgia Meloni a annoncé un plan de financement d'un **milliard d'euros** pour le développement de l'intelligence artificielle en Italie.



La stratégie italienne pour l'IA est en cours d'actualisation : le département pour la transformation numérique de la Présidence du Conseil des ministres et l'AGID (Agence pour l'Italie Numérique) ont publié **en mai 2024 les grandes lignes de la nouvelle stratégie**. Sur le volet recherche, le document prévoit :

- de consolider de l'écosystème italien de la recherche ;
- retenir et attirer les talents ;
- développer des LLM italiens ;
- mettre en œuvre des projets interdisciplinaires pour le bien-être social ;
- financer la recherche fondamentale pour l'IA de prochaine génération ;
- renforcer les collaborations internationales.

La recherche autour de l'IA en Italie est donc devenue une priorité politique et **la recherche publique y est plus développée que la recherche privée**. Elle est encadrée par divers organismes, universitaires et ministériels. La gouvernance se partage entre le secrétaire d'État à l'innovation technologique rattaché à la présidence du Conseil des ministres et auprès duquel est placé le département pour la transformation numérique, le ministère de l'université et de la recherche, qui encadre les activités de recherche à travers le Programme national de Recherche et le ministère de l'industrie et du « *Made in Italy* » (MIMIT). L'IA est l'un des domaines de recherche couverts par le PNR 2021-2027 et est articulé en six sous-domaines : l'IA pour l'IA, l'IA humano-centrée, l'IA pour la santé, l'IA pour la société, l'IA pour l'environnement et les infrastructures critiques et l'IA pour la production industrielle.

Le **blocage temporaire de ChatGPT de mars à avril 2023** souligne la vigilance des institutions italiennes face aux risques de fuites de données. Cette décision a été prise par la Cnil italienne (le Garant pour la protection des données personnelles), afin de respecter le RGPD (règlement européen sur la protection des données) : ce n'est **pas une initiative du gouvernement** à proprement parler. Le manque de transparence de la part d'OpenAI était l'élément mis en cause, ainsi que le risque élevé de fuite de données sensibles, l'absence de base légale de stockage des données, l'existence de fausses informations sur des personnes physiques et morales, et la non-vérification de l'âge des utilisateurs. À la suite de ce blocage s'est tenue une rencontre entre l'entreprise (en présence du PDG Sam Altman) et le Garant italien. L'autorité italienne a formulé des conditions pour la remise en marche de ChatGPT en Italie, auxquelles OpenAI a répondu en mettant en place diverses nouvelles mesures. Le Garant a fait savoir qu'il poursuivrait ses « *activités d'enquête sur OpenAI également sous l'égide de la task force ad hoc mise en place par le Comité européen pour la protection des données personnelles* ».

En ce qui concerne la réaction de l'opinion publique à ce blocage, les milieux technologiques ont parlé d'une décision inutile (les moyens de la contourner sont simples) et anachronique, mais la communauté scientifique italienne de l'IA, qui aurait influencé la décision, revendique un contrôle majeur de l'application de ces technologies dans le sens d'une éthique centrée sur le respect de l'être humain.

Barbara Caputo, professeur d'IA au Politecnico de Turin, souligne l'impact environnemental considérable de ces IA qui utilisent d'énormes quantités de données, donc de puissance de calcul et par conséquent d'énergie, alors qu'il est possible de développer une IA frugale – comme celle utilisée dans les satellites, sur lesquelles les équipes de recherche italiennes se sont fortement spécialisées. Cette décision italienne peut donc aussi être lue comme une forme d'**affirmation des compétences nationales**, qui va à rebours de l'accord de coopération bilatérale que le précédent gouvernement avait signé avec les États-Unis sur l'IA. L'Italie témoigne d'une volonté de compter sur la scène européenne et internationale en matière de déploiement de l'IA.

En dépit de ce contexte de **vigilance marquée des autorités italiennes** vis-à-vis des IA américaines, Nvidia, leader mondial de la conception des puces, a choisi l'Italie pour installer sa recherche européenne en IA.

### 3. L'Espagne : un cadre complet avec un riche volet culturel

La **présidence espagnole de l'UE en 2023** a conduit l'Espagne à traiter au plus haut niveau de nombreux sujets numériques, en particulier les négociations autour de l'*AI Act*. Mais l'Espagne avait commencé à s'investir sur le sujet de l'IA avant cela. Le gouvernement espagnol a présenté une **stratégie nationale pour l'IA en mars 2019**, initialement sans budget alloué, en raison de la paralysie budgétaire du pays. Par la suite, la 16<sup>e</sup> composante du Plan national de relance et de résilience (PNRR) espagnol a été dédiée en 2021 à cette stratégie nationale pour l'intelligence artificielle, en tant qu'investissement principal, mobilisant au total 500 millions d'euros (soit 0,7 % du plan de relance). Le manque d'investissements publics et privés en IA en Espagne est souvent souligné. Par ailleurs, l'harmonisation des données entre communautés autonomes est rendue difficile par le très fort degré de décentralisation : c'est par exemple le cas pour la santé, compétence des gouvernements régionaux (données non uniformisées entre les régions).

Une autre mesure du PNRR vise le **renforcement des capacités de supercalcul**, et plus particulièrement la facilitation de l'accès des PME et des entreprises à ces capacités. Le cadre organisationnel de cet accès n'étant pas détaillé, la mesure risque d'avoir un impact réduit.

Une dernière mesure du plan concernant l'influence linguistique de la langue espagnole touche aussi l'IA.

La stratégie espagnole en matière d'IA poursuit les objectifs suivants :

- positionner l'Espagne comme pays d'excellence scientifique et d'innovation en IA interdisciplinaire ;
- devenir leader en développement d'outils, technologies et applications pour l'utilisation de la langue espagnole dans l'IA ;
- promouvoir la création d'emploi qualifié, en favorisant la formation et l'éducation et stimulant le talent espagnol et attirant le talent global ;
- intégrer l'IA comme facteur pour améliorer la productivité des entreprises espagnoles et l'efficacité de l'administration, ainsi qu'en tant que moteur de croissance économique durable et inclusive ;
- instaurer un environnement de confiance par rapport à l'IA, tant en ce qui concerne son développement technologique que son cadre réglementaire et son impact social ;
- promouvoir le débat global sur l'humanisme technologique, en créant et en participant à des forums et activités de divulgation pour le développement d'un cadre éthique qui puisse garantir les droits individuels et collectifs des citoyens ;
- renforcer l'IA comme vecteur transversal pour faire face aux défis sociaux, dont l'égalité femme-homme, la fracture numérique, la transition écologique et la structuration territoriale.

Le Président du gouvernement Pedro Sánchez a présenté en juillet 2020 l'Agenda Digital 2025, qui prévoit l'instauration d'un **conseil consultatif de l'IA** et la création d'un Bureau de l'État des Données (*Oficina del Dato*) avec à sa tête un *Chief Data Officer*. Le conseil consultatif sur l'intelligence artificielle est composé d'experts espagnols de renommée internationale, y compris de représentants des domaines scientifique, économique et éducatif. Il fournit des conseils et des recommandations indépendants sur les mesures à adopter pour garantir l'utilisation sûre et éthique de l'intelligence artificielle. La ministre des affaires économiques et de la transformation numérique préside le Conseil et la secrétaire d'État à la numérisation et à l'intelligence artificielle en est la vice-présidente.

Plus récemment, l'Espagne a annoncé en juin 2022 un projet pilote pour la mise en place avec la Commission européenne du **premier bac à sable réglementaire de l'UE sur l'intelligence artificielle** (*sandbox*). Huit secteurs d'application qualifiés « à haut risque » ont été retenus pour les projets : (I) biométrie, (II) gestion, exploitation et fonctionnement de la circulation routière, de l'approvisionnement en eau, en gaz, en chauffage et en électricité, ainsi que gestion et exploitation des infrastructures numériques critiques, (III) éducation et formation professionnelle, (IV) recrutement, (V) accès aux services publics et privés, (VI) poursuite des crimes, (VII) migration (VIII) justice. Le budget total du projet pilote, financé par des fonds du PNRR, s'élève à environ 4,3 millions d'euros.

L'objectif de cette collaboration est de mettre en relation les autorités publiques avec les entreprises qui développent des intelligences artificielles afin de définir conjointement, à travers des tests techniques du banc d'essai de l'*AI Act*, les meilleures pratiques en vue de la mise en œuvre de la réglementation européenne sur l'IA. Le lancement du projet en présence du commissaire européen chargé du marché intérieur et des services, Thierry Breton, a conduit ce dernier à « *féliciter l'Espagne d'être le premier pays à avoir décidé d'investir une partie de son Plan de relance et de résilience pour lancer cet ambitieux premier projet pilote de sphère de sécurité en matière d'IA en Europe* ». Les résultats des premiers tests ont été compilés dans un guide de bonnes pratiques au deuxième semestre 2023 sous présidence espagnole de l'UE.

Le gouvernement espagnol a surtout décidé en août 2023 de créer une **agence nationale de supervision de l'IA (AESIA)**, dont le siège est à La Corogne en Galice. L'Espagne est alors devenue, avec les Pays-Bas, le **premier pays européen à se doter d'un tel organisme**, avant même l'entrée en vigueur de l'*AI Act*. Le règlement européen sur l'IA prévoit en effet l'obligation pour les États membres de se doter d'une autorité de supervision dans ce domaine. Ses missions comprennent le contrôle du respect de la réglementation européenne ; la sensibilisation, la diffusion et la promotion du développement et de l'utilisation responsable, durable et fiable de l'IA ; la définition de mécanismes de conseil et d'assistance ; la collaboration et la coordination avec d'autres autorités (nationales et supranationales) de supervision de l'IA ; et la promotion d'environnements de test de systèmes d'IA en conditions réelles pour renforcer la protection de l'utilisateur.

Par ailleurs, la Première vice-présidente et ministre des affaires économiques et de la transformation numérique a suggéré la création d'une **agence internationale de l'IA dans le cadre des Nations unies, proposant l'Espagne comme pays d'accueil du siège de cette institution**.

Le gouvernement espagnol a lancé en 2022 un plan national autour de la « nouvelle économie de la langue », lié à l'enjeu de **l'influence linguistique de la langue espagnole dans le domaine technologique**, dominé par l'anglais. La ministre des affaires économiques et de la transformation numérique a rappelé au cours du IX<sup>e</sup> congrès international de la langue espagnole l'objectif de « *préservation de notre patrimoine linguistique dans la sphère numérique* ». Elle a souligné que « *nous devons faire en sorte que les nouvelles technologies ne pensent pas seulement en chinois ou en anglais, mais aussi en espagnol. Notre langue doit monter dans le train de la révolution numérique* ».

Sur un total d'un milliard d'euros consacré à la nouvelle économie de la langue (un second milliard d'euros d'investissements privés est attendu), une partie sera consacrée à des investissements liés à l'IA. En effet, ce projet a notamment pour objectif de **placer l'espagnol** (591 millions d'hispanophones, soit 7,5 % de la population mondiale) et les autres « langues co-officielles » de l'Espagne (catalan, basque, galicien), **au cœur de la transformation numérique et de la promotion de la chaîne de valeur de la nouvelle économie de la connaissance et de l'intelligence artificielle**.

Le projet « nouvelle économie de la langue » s’articule autour de cinq axes, concernant principalement l’IA, autour de la constitution de bases de données textuelles dans les langues espagnoles, de la construction d’un LLM en espagnol et d’investissements pour améliorer la compréhension et l’expression en langues espagnoles pour les intelligences artificielles. L’encadré ci-après détaille ce plan.

### **Un projet stratégique autour de la nouvelle économie de la langue**

#### I. Présence et action d’organisations internationales de prestige autour de la langue espagnole

**L’Espagne dispose de trois grandes institutions spécialisées autour de la langue espagnole.** L’**Académie royale espagnole (RAE)** est une institution culturelle privée financée avec des fonds publics dont le rôle est de normaliser la langue espagnole. L’**Institut Cervantes (IC)** est une institution culturelle dépendant du ministère des affaires étrangères ayant un important réseau international (87 centres dans 44 pays) pour l’apprentissage de l’espagnol. La **Bibliothèque nationale d’Espagne (BNE)** produit des données textuelles de haute qualité, normalisées et interopérables selon le cadre européen pour la réutilisation des informations du secteur public. Ces institutions participent à des projets importants, au croisement de la langue et des industries innovantes.

#### II. Centres d’excellence de recherche, universités et centres d’apprentissage de l’espagnol

**Le Plan national de technologies du langage pour 2022-2025 prévoit de renforcer l’écosystème de centres d’excellence spécialisés en traitement de langage naturel et linguistique informatique.** La société espagnole pour le traitement du langage naturel (SEPLN) a près de 40 ans d’existence et rassemble plusieurs organismes publics, centres scientifiques et technologiques autour du sujet. En complément, **le réseau universitaire et de centres d’apprentissage de l’espagnol assure la promotion de l’attractivité du talent et du tourisme linguistique** (900 000 étudiants par an uniquement à travers l’Université de Salamanque et d’Alcalá de Henares ; par ailleurs, l’espagnol est dès à présent inclus comme 2<sup>e</sup> langue d’épreuve de fin d’études en Chine, offrant ainsi des opportunités de développement exponentiel dans ce secteur).

Plus largement, la langue s’inscrit dans la **Stratégie espagnole pour la science, la technologie et l’innovation (2021-2027)**. Ce domaine apparaît dans 2 des 6 grands axes : d’une part, dans les enjeux liés à **la culture, la créativité et l’inclusion** ; d’autre part, dans les défis de **l’intelligence artificielle et de la robotique**. La stratégie cible ainsi respectivement les domaines de l’acquisition du langage, l’impact des nouveaux moyens de communication sur l’évolution de la langue, ainsi que les technologies du langage et la compréhension profonde de la langue.

#### III. Administration publique numérique et entreprises espagnoles contribuant à la NEL

**L’administration publique espagnole est l’une des plus avancées en termes de numérisation**, avec de nombreuses applications pour l’accès aux services publics et démarches administratives. Le secteur privé espagnol comprend des **entreprises et groupes de référence dans l’édition (Planeta, Penguin Random House, Santillana...), les plateformes de production audiovisuelle (Mediaset, Atresmedia, Prisa...) et les télécommunications (Telefónica), parmi d’autres secteurs stratégiques pour la nouvelle économie de la langue.**

#### IV. Les actions regroupées en cinq axes

##### **1- Création de corpus larges en espagnol et langues co-officielles (97 millions d'euros)**

Le corpus, constitué de données orales, écrites, chantées et en langue des signes, a pour objectif d'atteindre un volume suffisamment important de données en espagnol pour générer un modèle et permettre des applications d'IA. Les collaborations avec des pays ibéro-américains seront développées pour sa constitution. Des projets seront financés (10 millions d'euros) pour la constitution de corpus spécifiques pour les langues co-officielles (catalan, basque et galicien). Cette base de connaissances sera mise à disposition des entreprises, des chercheurs et des administrations publiques. Par ailleurs, est créé sur ces crédits, un Observatoire de l'espagnol, organisme public spécialisé dans la promotion universelle de l'enseignement, l'étude, la certification et l'usage de l'espagnol. Enfin, l'Académie royale espagnole poursuit son projet *Corpus del Español del siglo XXI* (CORPES XXI) avec la constitution d'une base de connaissances de plus de 327 000 documents et 350 millions de formes orthographiques, provenant de textes écrits et de transcriptions orales. Ce projet fonctionne selon une série de paramètres avec des données traitées avec une codification conçue pour ce corpus pour la récupération des données à partir de chaque paramètre.

##### **2- Des LLM en espagnol (334 millions d'euros)**

**La promotion de l'IA en espagnol passe par la création de modèles de langage de haute valeur pour l'industrie**, une des principales actions constituant l'évolution du modèle de langage MarIA, tout en créant des tests de référence d'évaluation de compréhension de langage pour les tâches de traitement (équivalent des tests GLUE/SUPERGLUE pour l'anglais) et des certifications du bon usage de l'espagnol dans les outils technologiques et l'IA. Des appels d'offres pour le financement de projets d'innovation pour intégrer l'IA dans les chaînes de valeurs industrielles, ainsi que des aides aux entreprises ayant un modèle économique basé sur la langue espagnole, sont prévus. Le projet MarIA mené par le Barcelona Supercomputing Center avec la BNE depuis 2021 est un système massif d'IA expert en compréhension et écriture en espagnol, avec un entraînement de plus de 135 000 milliards de mots. En raison du volume et de la capacité de MarIA, la langue espagnole est la troisième langue ayant des systèmes massifs d'accès ouvert, après l'anglais et le mandarin. Le système est en accès ouvert pour l'utilisation libre de développeurs d'applications.

##### **3- La Science en espagnol (128 millions d'euros)**

Face à une sous-représentation de l'espagnol dans la science (2 % de la production scientifique mondiale est en espagnol, dont 60 % en provenance d'Espagne), diverses institutions, dont la Fondation espagnole pour la science et la technologie, contribueront à la divulgation nationale et internationale de la science en espagnol auprès du grand public. De plus, des actions seront menées pour le maintien et la conservation du patrimoine technique et scientifique hispanophone (dictionnaires et textes biomédicaux entre autres).

##### **4- Apprentissage de l'espagnol et en espagnol dans le monde (474 millions d'euros)**

L'apprentissage de l'espagnol et en langue espagnole dans le monde sera notamment soutenu à travers une plateforme technologique de certification de l'espagnol comme langue étrangère (80 millions d'euros). Le développement de l'apprentissage de la langue à l'international passera aussi par l'Amérique latine dans un écosystème d'apprentissage numérique de l'espagnol, par la digitalisation de l'Institut Cervantes et par des accords éducatifs bilatéraux.

##### **5- Industries culturelles (67 millions d'euros)**

Ce dernier axe intègre notamment des mesures déjà prises dans le cadre du plan « Espagne hub audiovisuel », avec **la mise en place d'appels d'offres d'aides à la production audiovisuelle et du secteur des jeux vidéo**, secteurs en essor en Espagne, accompagnée de la promotion du secteur à l'international, parmi d'autres actions.

*Source : Service économique régional et Service de coopération et d'action culturelle*

Il peut être souligné, en conclusion, qu'au-delà de l'économie, l'Espagne se préoccupe des enjeux culturels de l'IA et souhaite son développement éthique et ordonné, comme en témoigne son conseil consultatif de l'IA, son observatoire de l'IA, son label IA, sa déclaration de droits numériques, son plan de protection pour les collectifs vulnérables, etc. La défense de la langue et de la culture espagnoles est au cœur de la politique conduite par le gouvernement espagnol.

#### **4. Les Pays-Bas : une coalition public-privé efficace et une régulation précoce**

L'intelligence artificielle bénéficie de la volonté du gouvernement néerlandais de reprendre la main sur le développement de l'économie. Des dispositifs publics ciblés ont ainsi été mis en place depuis 2018-2019 : un **plan d'action national pour l'IA** (feuille de route pour l'État et l'industrie), une **instance de dialogue NL AI Coalition** et des **dotations budgétaires** dont celle du fonds national de croissance (plus de 276 millions d'euros par an). Une des forces de ces outils réside dans la responsabilisation des acteurs, leurs recommandations étant décisives dans la mise en œuvre des politiques nationales (axes de recherche à financer, programmes de subvention, soutiens à l'export, etc.). Mais si l'État néerlandais soutient fortement l'IA, il a été contraint d'instaurer en 2022 un cadre strict de contrôle des risques associés à ces technologies.

Le gouvernement néerlandais a d'abord déployé un cadre propice au développement de l'IA. Demandé par l'organisation patronale VNO-NCW et le secteur de la R&D depuis 2018, le **plan d'action national pour l'IA** (*Strategisch Actieplan voor Artificiële Intelligentie - SAPAI*) a été publié en octobre 2019. Ce plan a alloué 45 millions d'euros de financements publics annuels à la R&D en IA et promeut l'application et la commercialisation des avancées technologiques en matière d'IA.

Conscient de la valeur stratégique de l'IA, le gouvernement soutient son développement par des **partenariats avec le secteur privé** visant plusieurs objectifs : intensification de la recherche, financement de la capacité d'innovation, accélération des applications industrielles. Les instances de dialogue public-privé permettent d'identifier des marchés de niche dans lesquels l'écosystème néerlandais présente un potentiel d'excellence.

En octobre 2019, en même temps que la publication de la stratégie nationale sur l'intelligence artificielle réclamée depuis 2018 par divers secteurs économiques, le ministère de l'économie et du climat (EZK) et le patronat (VNO-NCW, MKB-Nederland), avec l'aide de l'association fédératrice du secteur numérique Dutch Digital Delta, d'entreprises (Seedlink, Philips, Ahold Delhaize et IBM), de centres de R&D comme l'institut de recherche appliquée TNO, ont créé **l'association NL AI Coalition**, rencontrée par vos rapporteurs à La Haye.



Organisée en 18 groupes de travail thématiques et 7 centres régionaux d'IA, l'association compte aujourd'hui 475 organisations membres et vise à favoriser le développement de l'IA et son appropriation par les acteurs économiques. Les groupes de travail, composés de chercheurs, d'entrepreneurs et de fonctionnaires, formulent des recommandations pour le gouvernement ainsi que pour des secteurs spécifiques : agriculture et nutrition ; environnement bâti ; culture et médias ; défense ; éducation ; énergie et développement durable ; services financiers ; santé ; mobilité, transport et logistique ; portuaire et maritime ; services publics ; sécurité, paix et justice ; industrie technique. Les centres régionaux d'IA jouent un rôle dans la mise en relation et l'implication des entreprises locales, des institutions de connaissance et des autres organisations travaillant avec cette technologie. Il s'agit d'accélérer les développements technologiques, l'innovation, l'intégration sociale et le développement économique.

En s'appuyant sur la « triple hélice » (gouvernement, entreprises, recherche), NL AI Coalition formule des recommandations à l'adresse de l'industrie et de l'État. La qualité du dialogue au sein de l'écosystème de l'IA stimule la **contribution des entreprises à la recherche** : 15 % des publications scientifiques associent au moins un auteur issu du secteur privé (seuls les États-Unis devançant ce taux à 19 %). La NL AIC a également pour objectif de favoriser le développement d'une **expertise dans des marchés de niche** pour renforcer l'avantage concurrentiel des Pays-Bas à l'international. L'Institut Rathenau met en évidence que 26,3 % des publications scientifiques néerlandaises en IA portent sur la **prise de décision** (gestion, planification, etc.) notamment pour la conduite autonome et la robotique, spécialité dans laquelle l'écosystème veut se démarquer.

En outre, à travers son **Fonds national de croissance**, lancé en 2021 et doté de 20 milliards d'euros pour renforcer le potentiel de croissance de l'économie, l'État a réservé **276 millions d'euros au programme AiNed**, une initiative de **l'association NL AI Coalition** pour accélérer l'innovation et renforcer la compétitivité du secteur à l'international.

Cet ensemble de mesures a eu pour effet de faire passer les Pays-Bas de la 14<sup>e</sup> à la 5<sup>e</sup> place de l'index *AI Government Readiness* de l'Institut de recherche Oxford Insights entre 2019 et 2021, devant la France (11<sup>e</sup>).

À la suite d'un mésusage des algorithmes, le gouvernement a par ailleurs aménagé **des garde-fous réglementaires** pour l'IA : les mésaventures liées à un recours fréquent à l'IA par les services publics ont en effet conduit à un scandale rendant nécessaire le renforcement du cadre d'utilisation de l'IA.

**L'affaire des allocations familiales** (cf. l'encadré), qui a contraint le gouvernement Rutte III à la démission fin 2020, illustre la sensibilité de l'utilisation d'algorithmes par les pouvoirs publics. Dans cette affaire, les autorités publiques ont été mises en cause pour l'utilisation d'algorithmes visant à identifier les ménages les plus susceptibles de frauder en se fondant sur des critères ethniques.

### L'affaire des fraudes aux allocations familiales

En 2004, à la demande du parlement néerlandais, le gouvernement Balkenende II (2003-2006) a mis en place une politique plus stricte de contrôle et de recouvrement des allocations familiales. Cette politique a été poursuivie par les gouvernements Rutte II (2012-2017) et Rutte III (2017-2022). Afin de renforcer les contrôles anti-fraude, des algorithmes ont été utilisés pour cibler des individus et des ménages à risque. Le ciblage était parfois fondé sur les origines ou la nationalité, à la suite de révélations de la presse sur des fraudeurs originaires d'Europe de l'Est. Plus de 26 000 parents ont été accusés à tort de fraude et tenus de rembourser en moyenne plus de 30 000 euros à l'administration fiscale. Certaines de ces personnes ont vu leur vie bouleversée : pertes d'emploi, expropriations de leur logement, divorces, placements de leurs enfants, surendettements, problèmes psychologiques... Cela a notamment poussé le gouvernement Rutte III à remettre sa démission en janvier 2021.

Le scandale des allocations (*toeslagenaffaire*) a été mis au grand jour par un parlementaire en 2017. Outre le fait que des parlementaires de la coalition, qui auraient eu connaissance des faits (ces ménages injustement pénalisés), auraient choisi, pour des raisons politiques, de ne pas mettre en difficulté le gouvernement, les commissions d'enquête parlementaires ont révélé que les méthodes de travail de l'administration fiscale étaient illégales, discriminatoires et inappropriées. Dans certains cas, il y avait un parti pris institutionnel et une violation des principes fondamentaux de l'État de droit. La Cour des comptes néerlandaise a par ailleurs montré (étude de mai 2022) que les algorithmes mis en œuvre par l'administration présentaient des risques en matière d'atteinte à la vie privée, de discrimination ou de fuite de données personnelles.

*Source : Service économique régional*

En réaction, l'accord de coalition du gouvernement Rutte IV (janvier 2022), a conduit à la création en janvier 2023 d'une **instance de supervision des algorithmes et de l'IA** rattachée à l'Autorité de Protection des Données (APD). Elle est dotée d'un budget annuel d'un million d'euros, qui atteindra 3,6 millions d'euros d'ici 2026. Cette nouvelle instance de supervision est intervenue pour la première fois fin avril 2023 auprès du ministère néerlandais des affaires étrangères. Ce dernier doit justifier l'utilisation d'algorithmes de présélection des demandes de visas nécessitant une instruction « approfondie » en se fondant sur des critères possiblement ethniques.

Ce souci du gouvernement d'encadrer les risques de l'IA suite au traumatisme de l'affaire des allocations familiales a conduit les Pays-Bas à mettre en place, à peu près en même temps que l'Espagne, le **premier cadre réglementaire en Europe** comprenant la création d'une instance de supervision. Cette initiative a devancé l'*AI Act* de l'Union européenne et a pu l'influencer dans son approche.

## 5. L'Estonie : un État numérique préoccupé par la sécurité de l'IA

En termes de technologies, l'Estonie est un **pays particulièrement avancé**. Son administration est déjà utilisatrice de la technologie des *blockchains* ou chaînes de blocs<sup>1</sup>. Les citoyens estoniens sont habitués aux technologies numériques avec lesquelles ils sont en contact pour de nombreuses démarches administratives. De ce fait, l'intelligence artificielle n'y est pas perçue comme un risque majeur. La protection des données y est également une problématique moins sensible : les citoyens considèrent que la croissance économique doit primer la stricte protection des données personnelles. Cela se répercute dans les choix réalisés par les décideurs politiques dont le principal objectif est d'**utiliser l'IA pour faire en sorte que les services administratifs soient assurés de la façon la plus fluide et aisée possible**.

Ces objectifs sont détaillés dans un **livre blanc pour l'IA** qui détermine des recommandations pour le pays en matière de développement de l'IA et de gestion des données. La volonté des pouvoirs publics est de former 80 % de la population à la bonne gouvernance des données. Ils souhaitent parallèlement créer un **corpus de ressources en langue estonienne**, langue rare et difficile à apprendre, ce qui permettrait d'entraîner des LLM dans cette langue.

Pour réaliser ces objectifs, le gouvernement estonien a défini une **stratégie nationale pour l'IA en 2022**. Ce plan prévoit un investissement de 85 millions d'euros dans les secteurs porteurs pour le développement du numérique comme l'e-santé, les technologies vertes, les technologies éducatives, etc. La stratégie est également tournée vers la coopération entre la recherche et les entreprises privées, dans des centres de recherche comme Cybernetica, développeur du système administratif numérique estonien (*X-Road*).

L'Estonie a été **victime de la première cyberattaque d'envergure nationale de la part de la Russie**, ayant paralysé l'entièreté du pays en 2007. De ce fait, les Estoniens sont particulièrement attentifs aux questions de cybersécurité, y compris la sécurité des systèmes d'IA. Des ambassades de données (*data embassies*) ont été mises en place au Luxembourg pour sauvegarder les données estoniennes en cas d'attaques. Il est aussi à noter que l'Estonie organise tous les ans le *Tallin digital summit*, sommet dédié à la gouvernance du numérique.

---

<sup>1</sup> Dans le rapport de l'OPECST sur la technologie des *blockchains*, des développements sont consacrés à l'usage de cette technologie en Estonie. Cf. Valéria Faure-Muntian, Claude de Ganay et Ronan Le Gleut, 2018, *Comprendre les blockchains : fonctionnement et enjeux de ces nouvelles technologies*, disponible sur le site du Sénat : <https://www.senat.fr/notice-rapport/2017/r17-584-notice.html> et de l'Assemblée nationale : [https://www.assemblee-nationale.fr/dyn/15/rapports/ots/115b1092\\_rapport-information](https://www.assemblee-nationale.fr/dyn/15/rapports/ots/115b1092_rapport-information)

## 6. La Finlande : une stratégie tournée vers l'appropriation de l'IA et l'éducation

La **stratégie finlandaise pour l'IA est lancée dès 2017** avec la **nomination d'un comité de pilotage** chargé de préparer une proposition de programme, avec notamment l'idée de **stimuler la recherche et l'éducation dans le domaine**. Comme le montre un article paru en août 2024<sup>1</sup>, la Finlande vise l'excellence en matière de formation et de coopération active entre les acteurs afin de créer un environnement de recherche et d'innovation dynamique et attractif. Pour atteindre ces objectifs, le gouvernement finlandais a mis en place un plan d'action en trois volets :

- la mise en place d'une communauté d'experts, d'acteurs et de personnalités académiques, scientifiques et économiques, sous l'égide du ministère de l'économie et de l'emploi, chargé de proposer des recommandations pour le développement de l'IA en Finlande ;

- un programme de recherche en IA ;

- une enveloppe de 200 millions d'euros sur la période 2018-2021, inscrite au budget de *Business Finland*<sup>2</sup>.

Le rapport du comité de pilotage paru en 2019<sup>3</sup> identifie **trois piliers** (un secteur public efficace, une société proactive et un secteur des affaires compétitif) et préconise **onze actions** pour faire entrer la Finlande dans l'ère de l'IA et en faire un leader mondial dans l'intelligence artificielle :

1. Améliorer la compétitivité des entreprises par l'utilisation de l'IA ;
2. Utiliser efficacement les données dans tous les secteurs ;
3. Faire en sorte que l'IA puisse être adoptée plus rapidement et plus facilement ;
4. Garantir une expertise de haut niveau et attirer les meilleurs experts ;
5. Prendre des décisions et des investissements audacieux ;

---

<sup>1</sup> Cf. le focus pays de la revue ActuIA d'août 2024, « Finlande : le pari de l'éducation et de la coopération » : <https://www.actuia.com/actualite/focus-pays-finlande-le-pari-de-leducation-et-de-la-cooperation/>

<sup>2</sup> Business Finland est l'organisation gouvernementale finlandaise pour le financement de l'innovation et la promotion du commerce, des voyages et des investissements. Cette agence publique est au cœur de la stratégie de développement finlandaise. Cette organisation gouvernementale est chargée notamment du financement de l'innovation et des start-up dans le domaine de l'IA. Le programme AI Business dirigé par Outi Keski-Äijö est l'une des initiatives lancées par le plan de 2017, qui a aidé plus de 300 entreprises d'IA, ce qui représente l'essentiel de ces 200 millions d'euros de financement. Ce programme vise à soutenir les start-up et à augmenter l'attractivité du pays pour la recherche et le développement de l'IA.

<sup>3</sup> Rapport « Leading the way into the era of artificial intelligence: Final report of Finland's Artificial Intelligence Programme 2019 » : <https://tem.fi/en/publication?pubid=URN:ISBN:978-952-327-437-2>

6. Mettre en place les meilleurs services publics du monde ;
7. Établir de nouveaux modèles de collaboration ;
8. Faire de la Finlande un précurseur à l'ère de l'intelligence artificielle ;
9. Se préparer à ce que l'intelligence artificielle change la nature du travail ;
10. Orienter le développement de l'intelligence artificielle dans une direction fondée sur la confiance et centrée sur l'homme ;
11. Se préparer aux défis de la sécurité.

Pionnière en matière de technologies numériques et cherchant à se situer à la pointe de l'innovation, notamment grâce à des **initiatives publiques et des politiques ambitieuses**, la Finlande a su **mobiliser sa population, son administration<sup>1</sup> et ses entreprises<sup>2</sup>** autour du **développement** et de la **diffusion des technologies d'IA**, *via* l'accès aux données publiques ou l'attention particulière portée au système éducatif.

Cette stratégie avant-gardiste s'inscrit dans **l'intérêt déjà ancien de la Finlande pour les nouvelles technologies** : **Nokia** a été durant de nombreuses années l'image de marque du pays, sans compter son tissu dynamique d'experts à la croisée du monde académique et de la sphère entrepreneuriale. L'article d'ActuIA donne l'exemple de **Teuvo Kohonen**, académicien, chercheur et professeur émérite à l'Université technologique d'Helsinki. Spécialiste des réseaux neuronaux artificiels, il a travaillé sur l'algorithme du *Learning Vector Quantization*, basé sur la quantification vectorielle, ou encore sur la théorie fondamentale sur la mémoire. Il a également présenté la carte autoadaptative dite « carte de Kohonen » dès les années 1980, qui a marqué l'histoire de la recherche sur les réseaux de neurones et la reconnaissance de formes.

Depuis 2018, le secteur public et le secteur privé finlandais se sont mis d'accord sur un **cadre éthique pour l'intelligence artificielle**, au terme d'un partenariat mixte auquel ont participé plus de 200 experts du gouvernement, des entreprises et des universités ou instituts de recherche. Le rapport intitulé « Politique d'information éthique à l'ère de l'intelligence artificielle » représente une **base à l'aune de laquelle sont jugées les nouvelles lois et réglementations**.

---

<sup>1</sup> En 2021 et 2022, sur l'indice de préparation à l'IA des Oxford Insights, la Finlande était classée 4<sup>e</sup> au niveau mondial derrière les États-Unis et le Royaume-Uni mais devançant l'Allemagne (8<sup>e</sup>) et la France (11<sup>e</sup>). Il s'agit de savoir, à travers 42 indicateurs et 10 dimensions, dans quelle mesure les gouvernements tirent parti des avantages de l'IA dans leurs opérations et la prestation de services publics. Cf. le « Oxford Insights AI Readiness Index » <https://www.oxfordinsights.com/government-ai-readiness-index2021>

<sup>2</sup> En novembre 2020, la Finlande a lancé le programme « Intelligence artificielle 4.0 » pour promouvoir le développement de l'IA dans les entreprises, en mettant l'accent sur les PME. Cf. le programme pour l'intelligence artificielle « AI Business program » de Business Finland : <https://www.businessfinland.fi/en/for-finnish-customers/services/programs/ended-programs/ai-business>

Il plaide notamment en faveur du principe d'égalité afin d'éviter tout biais ou discrimination, d'une meilleure protection des données, de la fiabilité de leur traitement, des pratiques plus transparentes et des lignes directrices dans le développement des algorithmes et des architectures. Il réaffirme le fait que l'utilisation des systèmes d'intelligence artificielle ne dédouane pas les individus de leurs responsabilités. Il montre aussi la volonté du gouvernement finlandais de nourrir **un débat national continu sur ces questions** et de **participer aux discussions internationales** sur la régulation de l'IA.

En 2020, Business Finland a lancé un **projet pour une intelligence artificielle responsable** appelé AIGA (pour *Artificial Intelligence Governance and Auditing*). Le projet traite à la fois des dimensions juridiques, éthiques et techniques de la prise de décision algorithmique et est financé à hauteur de 4,25 millions d'euros au sein d'un **consortium public-privé**.

La transformation numérique et le développement des technologies d'IA ont permis la **mise en place d'initiatives et de programmes destinés aux services publics** comme AuroraAI, lancé par le ministère finlandais des finances, et qui vise à **aider les citoyens et les entreprises** en leur proposant des services axés sur leurs besoins. Cet intérêt pour l'IA répond au potentiel économique important du domaine. Le pays serait la deuxième économie mondiale ayant le plus à gagner du développement de l'IA, derrière les États-Unis. Plusieurs études indiquaient en effet, dès 2017, que l'IA pourrait permettre à la Finlande de doubler son taux de croissance économique d'ici 2035 (études d'Accenture et de Frontier Economics par exemple). En 2020, les municipalités d'Helsinki et d'Amsterdam ont lancé une initiative pour des IA ouvertes, afin de savoir comment les algorithmes sont utilisés dans les services publics afin de s'assurer que l'IA fonctionne selon les principes de responsabilité, de transparence et de sécurité et d'améliorer le service rendu et l'expérience des citoyens.

L'Académie de Finlande a créé en 2019 le **Centre finlandais pour l'intelligence artificielle** (FCAI) qui rassemble des experts académiques, industriels et venant du secteur public, travaillant sur l'IA. Avec un budget de 250 millions d'euros pour 2019-2026, ce *flagship programme* de la Finlande est devenu l'un des pôles d'innovation numérique de la Commission européenne (AI DIH). Le FCAI s'est associé avec le centre finlandais d'expertise en technologie de l'information (Center for Science-IT) et Nvidia pour créer le **centre technologique Nvidia AI** (NVAITC), qui contribue à accélérer la recherche et l'adoption de l'intelligence artificielle en Finlande en faisant bénéficier les chercheurs de l'expertise de Nvidia dans l'utilisation des processeurs graphiques (GPU) et de logiciels d'IA de pointe. Le NVAITC propose une **puissance de calcul** considérable aux chercheurs et a entrepris **neuf projets** dont les sujets sont, par exemple, les processus gaussiens, la vision par ordinateur, la modélisation générative ou le traitement du langage naturel.

Le pays cherche depuis plus de 50 ans à faciliter l'accès aux **supercalculateurs** et dispose ainsi depuis 1971 de son Center for Science-IT, basé à Aspoo et à Kajaani, qui se présente comme l'un des plus grands acteurs mondiaux dans le domaine du calcul haute performance. Cette entreprise publique à but non lucratif abrite le système national de calcul et de gestion des données de la Finlande. Elle a été choisie pour accueillir le **supercalculateur LUMI**, l'un des trois supercalculateurs pré-exascale de l'initiative EuroHPC. Ce projet de 200 millions d'euros est financé à 50 % par la Commission européenne et à 50 % par les dix pays participants. Le supercalculateur installé en 2021 dispose d'une puissance de calcul de 552 pétaflops (millions de milliards d'opérations en virgule flottante par seconde). Le CSC, qui fournit déjà aux *start-up* finlandaises des ressources informatiques gratuites pour leurs projets de recherche grâce à la subvention informatique de *Business Finland*, réserve 20 % de la capacité de calcul de LUMI aux industriels et aux PME-PMI.

En plus du FCAI, le Tampere AI Hub et l'Académie de l'IA de l'université de Turku, ainsi que des initiatives régionales et d'autres accélérateurs ont également pour objectif de **transférer efficacement les compétences aux start-up et aux autres entreprises** afin de stimuler la commercialisation de l'IA et d'accélérer son déploiement.

La Finlande compte d'ores et déjà **plus de 300 start-up en IA** dans différents domaines commerciaux, ce qui fait d'Helsinki **l'un des écosystèmes d'IA les plus importants d'Europe** avec Londres, Paris et Berlin. Un rapport sur l'écosystème mondial des *start-up* publié en 2020 classe même le Grand Helsinki au **quatrième rang mondial**<sup>1</sup>. Les liens avec la recherche académique, les organismes de recherche et les acteurs publics sont particulièrement renforcés pour que les jeunes pousses puissent accéder à toutes les clés pour s'inscrire dans les marchés et créer de nouveaux secteurs porteurs. La région d'Helsinki a notamment été reconnue comme l'un des plus importants écosystèmes pour le démarrage en IA en Europe. De larges bases de données sont mises à disposition des entreprises afin de susciter une plus grande et plus rapide adoption de l'IA dans le pays. Elles peuvent capitaliser sur les traditions de recherche, en reconnaissance des formes, en traitement automatique du langage (TAL) ou en vision industrielle par exemple, et sur les coopérations entre secteurs. La société de radiodiffusion nationale finlandaise (YLE) a lancé l'an dernier une campagne de collecte du finnois parlé dans tout le pays afin que les algorithmes puissent apprendre à comprendre et à reconnaître les spécificités de ce langage.

---

<sup>1</sup> Cf. le rapport 2020 sur l'écosystème mondial des start-up (GSER) de Startup Genome : <https://startupgenome.com/article/rankings-top-100-emerging>



En parallèle de ces enjeux strictement économiques, il s'agit aussi de **diffuser la connaissance des technologies** et de **permettre une appropriation de l'IA par les citoyens finlandais**, voire par les citoyens d'autres pays. Plusieurs universités finlandaises proposent un enseignement de haut niveau sur l'IA. La Finlande est déjà « *à la pointe de l'enseignement de l'IA* » au niveau mondial, selon un rapport réalisé sur les compétences par Coursera en 2021<sup>1</sup>.

Bien entendu, **sensibiliser la population aux enjeux de l'intelligence artificielle** répond aussi à l'objectif de la Finlande pour développer la recherche et le développement ainsi que le déploiement d'activités et de solutions sur un secteur économique en plein essor et au potentiel élevé.

C'est un sujet auquel l'Académie de Finlande s'intéresse à travers son programme ICT 2023 pour la R&D et l'innovation, en vue tout particulièrement de renforcer les connaissances et les applications en *Machine Learning*, Internet industriel, technologies et services de santé innovants centrés sur l'utilisateur. Des centres de recherches, comme l'Institut d'informatique d'Helsinki (HIIT), se sont rapidement développés pour accueillir les chercheurs et les entreprises.

Ces dernières années, plus de 6 300 étudiants suivaient chaque année au moins un cours d'IA dans le cadre de leur formation. Les grandes universités finlandaises proposent près de 250 cours d'IA et plus de 40 formations de niveau master, 19 programmes de niveau licence et trois programmes de doctorat, auxquels s'ajoutent les 26 programmes de formation dispensés par les grandes écoles spécialisées et le Centre de recherche technique VTT de Finlande. Le pays se place logiquement en deuxième position au regard du nombre d'experts en IA par habitant, parmi tous les pays européens (*LinkedIn Economic Graph* 2019).

Selon l'OCDE, les **principales universités** en nombre de publications de recherche portant sur l'IA, sont l'Université de Helsinki, Aalto Université, l'Université de Tampere, Tampere University of Technology, Helsinki University of Technology, l'Université de Oulu, l'Université de Turku, l'Université de Finlande orientale et l'Université de Jyväskylä.

Les universités finlandaises misent notamment sur des **opportunités d'apprentissage accessibles à tous** et sur l'attrait des citoyens pour le numérique, notamment grâce à des **cours publics gratuits en ligne**, comme le plus fameux d'entre eux, Elements of AI.

---

<sup>1</sup> Cf. le « *Global Skills Report* » de Coursera pour 2021 : <https://www.coursera.org/skills-reports/global>

### Le programme Elements of AI

Lancé début 2018 par l'université d'Helsinki, Elements of AI est une série de MOOC conçue en collaboration avec la société Reaktor. Elle a été classée **n° 1 mondial des MOOC en IA** par le portail de cours en ligne Class Central et par Forbes et a remporté le grand prix *Inclusive Innovation Challenge* du MIT. Ces cours en ligne gratuits sont disponibles dans la plupart des langues de l'UE. Ils peuvent être suivis au rythme de chacun et combinent théorie et exercices pratiques.

Son objectif était de démystifier l'IA et d'éduquer un pour cent de la population finlandaise (environ cinquante mille personnes) afin qu'elle se familiarise avec les concepts fondamentaux et les logiques sous-jacentes de ces technologies. L'objectif a été dépassé puisqu'outre les 100 000 personnes qui se sont inscrites aux cours en Finlande, Elements of AI aurait au total formé, depuis son lancement, plus de 750 000 personnes dans le monde et permis de diplômer des étudiants de plus de 170 pays.

Le premier volet, *Introduction to AI*, permet de se familiariser avec le *Machine Learning*, les réseaux de neurones, la résolution de problèmes grâce à l'IA ou encore les aspects philosophiques de l'IA. Plus de 1 % de la population finlandaise a déjà été formée aux bases de l'IA grâce à ce cours en ligne gratuit. À l'occasion de la présidence finlandaise de l'Union européenne en 2019, le MOOC a été **traduit dans de nombreuses langues** pour permettre aux citoyens européens de se former eux aussi aux bases de l'IA. Pour sa version française, le partenaire de cette initiative a été Sorbonne-Université.

En 2023, Elements of AI a mis en ligne son nouveau MOOC, Building AI, qui permet de découvrir les **algorithmes** servant à créer des méthodes d'IA. Certaines compétences de base en programmation Python sont recommandées pour tirer le meilleur parti du cours.

**La moitié des étudiants étant des femmes**, soit plus du double de la moyenne des cours d'informatique, ces cours permettent de **réduire la disparité entre les sexes** qui prévaut dans ce secteur.

Le MOOC Ethics of AI lancé par l'université d'Helsinki fin 2020 s'intéresse à **l'éthique de l'intelligence artificielle** et propose des textes, des exercices et un grand nombre de cas réels illustrant différentes problématiques d'un point de vue éthique.

*Source : Focus pays de la revue ActuIA op. cit.*

### C. LA GOUVERNANCE EUROPÉENNE DE L'INTELLIGENCE ARTIFICIELLE

Les réflexions menées au sein de l'UE en termes de régulation de l'IA depuis quatre ans sont **très proches de celles conduites par l'OCDE**. Mais l'encadrement européen des systèmes d'IA ne se limite pas au régime spécifique les concernant. Le droit de l'Union européenne prévoit en effet **différentes dispositions concernant les outils numériques qui peuvent impacter les systèmes d'intelligence artificielle**. Depuis le 13 juin 2024, **un règlement est totalement consacré à l'encadrement de l'IA** ; il s'ajoute à de nombreuses autres dispositions.

Les textes sont très nombreux, parfois d'application sectorielle, il est donc difficile de tous les récapituler ici. Pour mémoire, les plus récents et les plus transversaux viennent compléter le **règlement général sur la protection des données** (RGPD) du 27 avril 2016 et s'insèrent dans la « Stratégie numérique pour l'Europe pour la décennie 2020-2030 » : à l'instar du **règlement sur les marchés numériques** du 14 septembre 2022 (dit « DMA », pour *Digital Markets Act*, empêchant notamment les géants du numérique de privilégier leurs services sur leurs plateformes en laissant chacun choisir librement son moteur de recherche, son navigateur ou sa messagerie) ; du **règlement sur les services numériques** du 19 octobre 2022 (dit « DSA », pour *Digital Services Act*, qui responsabilise les plateformes en rendant illégal en ligne ce qui est illégal hors ligne, comme les contenus illicites, et dont l'esprit se retrouve dans la loi du 21 mai 2024 visant à sécuriser et à réguler l'espace numérique (dite loi « SREN ») en interdisant – en ligne – les arnaques, la haine, la désinformation et la publicité ciblée sur les mineurs ou encore en protégeant les mineurs de la pornographie ; du règlement sur la gouvernance des données du 30 mai 2022 ; du règlement sur la cybersécurité du 17 avril 2019, de la directive sur la cybersécurité du 14 décembre 2022 et du règlement sur la cybersécurité du 13 décembre 2023 ; du règlement du 12 mars 2024 sur la cyberrésilience, etc.

Dans ce contexte foisonnant, il est important de noter qu'en 2024, au terme de plusieurs années de travail préparatoire, un cadre visant spécifiquement les systèmes d'intelligence artificielle a été adopté à travers le **règlement du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle**<sup>1</sup>, communément dénommé *AI Act*. Il fait l'objet d'une application progressive d'ici à 2026 avec des règles harmonisées applicables à la mise sur le marché, à la mise en service et à l'utilisation de systèmes d'IA.

Présenté par l'Union européenne comme « *premier texte législatif de ce type au monde* », **cette réglementation, qui se veut pionnière, a donc vocation à devenir un standard mondial concernant la mise sur le marché, la mise en service et l'utilisation de ces systèmes** dans le but de garantir que l'IA soit une technologie « *axée sur l'humain* » et que les systèmes d'IA soient « *sûrs, éthiques et dignes de confiance* ».

Comme l'assure l'exposé des motifs du projet de règlement, **l'AI Act vient compléter le droit de l'Union et ne le remplace pas**, en particulier en ce qui concerne les droits fondamentaux ; la protection de la vie privée, des données, des consommateurs et des travailleurs ; l'emploi ; la sécurité des produits. Les droits et les recours existants pour les personnes sur lesquelles les systèmes d'IA sont susceptibles d'avoir des incidences négatives demeurent inchangés et pleinement applicables. L'intelligence artificielle ne suspend pas le droit en vigueur.

---

<sup>1</sup> Journal officiel de l'Union européenne (JOUE) le 12 juillet 2024 : [https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=OJ:L\\_202401689](https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=OJ:L_202401689)

## 1. Le travail préparatoire conduit par les institutions européennes entre 2018 et 2020

La construction des bases de la gouvernance européenne en matière d'intelligence artificielle s'est d'abord appuyée sur **trois principales contributions** : une **communication** de la Commission européenne en avril 2018, les différentes **recommandations d'un groupe d'experts** placé auprès de la Commission européenne en 2019 puis un **Livre blanc** de la Commission en 2020.

### a) La communication de la Commission européenne d'avril 2018

La Commission européenne a tout d'abord publié en avril 2018 une **communication pour l'intelligence artificielle en Europe**<sup>1</sup>. Elle appelait à la mise en place d'un **cadre éthique et juridique conforme aux valeurs** de l'Union européenne et à la Charte des droits fondamentaux, mais **sans préconiser la création d'une législation contraignante** spécifique à l'intelligence artificielle. Au lieu de cela, la communication attirait l'attention sur les cadres juridiques existants (la protection des données personnelles avec le RGPD de 2016, les règles relatives à la sécurité des produits et les régimes de responsabilité civile ordinaires).

Sur la base de cette première communication, un **travail de coordination a été conduit avec les États membres et la Norvège** en vue d'aboutir rapidement à un plan coordonné pour l'IA en Europe. Le 7 décembre 2018, la Commission européenne a publié une **communication sur les objectifs et les initiatives d'un plan coordonné dans le domaine de l'intelligence artificielle**, plan actualisé en 2021<sup>2</sup>.

### b) Le groupe d'experts de haut niveau sur l'IA

Dans la perspective de ses futurs travaux sur l'IA, la Commission européenne avait entretemps nommé en juin 2018 un **groupe d'experts de haut niveau sur l'IA** (*high-level experts group on artificial intelligence* ou *HLEG AI* en anglais). Ce groupe était composé d'experts des milieux industriels et scientifiques et présidé par Pekka Ala-Pietilä, président du conseil d'administration de Huhtamaki, Sanoma et Netcompany et ex-président de Nokia ainsi que docteur honoraire en technologie à l'université de Tampere en Finlande. Le groupe devait élaborer des conclusions permettant de guider l'action de la Commission ainsi que des colégislateurs en matière d'IA. Pour ce faire, il a d'abord publié en décembre 2018 une définition de la notion d'IA, qui lui a permis de circonscrire le périmètre de ses travaux.

---

<sup>1</sup> Communication de la Commission européenne, au Parlement européen, au Conseil, au comité économique et social européen et au comité des régions « L'intelligence Artificielle Pour l'Europe », avril 2018 : <https://digital-strategy.ec.europa.eu/en/library/communication-artificial-intelligence-europe>

<sup>2</sup> Communication sur le plan coordonné dans le domaine de l'intelligence artificielle actualisée en 2021 : <https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review> et première version de 2018 : <https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=COM:2018:795:FIN>

Cette définition reprend et étend la définition de la précédente communication de la Commission<sup>1</sup> :

*« L'intelligence artificielle (IA) désigne les systèmes qui font preuve d'un comportement intelligent en analysant leur environnement et en prenant des mesures – avec un certain degré d'autonomie – pour atteindre des objectifs spécifiques. Les systèmes dotés d'IA peuvent être purement logiciels, agissant dans le monde virtuel (assistants vocaux, logiciels d'analyse d'images, moteurs de recherche ou systèmes de reconnaissance vocale et faciale, par exemple) mais l'IA peut aussi être intégrée dans des dispositifs matériels (robots évolués, voitures autonomes, drones ou applications de l'internet des objets, par exemple). »*

Sur la base de cette définition, le HLEG AI a mis au point plusieurs livrables thématiques pour permettre à la Commission européenne de travailler sur la base de définitions et d'objectifs clairement établis.

Le **premier livrable**, publié en avril 2019<sup>2</sup>, fixe les **lignes directrices en matière d'éthique de l'IA** qui permettraient d'arriver à une « IA digne de confiance »<sup>3</sup>. Le groupe considère que pour qu'une technologie d'IA soit considérée comme « digne de confiance », il lui faut réunir au moins trois de ces éléments :

- être licite (respecter les législations et réglementations existantes) ;
- être éthique, et assurer l'adhésion à des principes et valeurs éthiques ;
- être équitable, diverse et non discriminatoire ;
- être transparente ;
- permettre un contrôle humain ;
- être robuste techniquement et socialement et sécurisée ;
- être robuste pour ne pas causer de préjudices involontaires ;
- respecter la vie privée et la gouvernance des données ;
- participer au bien-être social et environnemental ;
- savoir qui est responsable sur toute la chaîne de valeur.

**Ces exigences sont proches de celles préconisées par l'OCDE pour la mise en place d'une IA éthique.**

---

<sup>1</sup> Cf. la communication d'avril 2018 de la Commission précitée.

<sup>2</sup> « Lignes directrices en matière d'éthique pour une IA digne de confiance », 2019 : <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

<sup>3</sup> La notion d'IA digne de confiance est souvent utilisée, particulièrement par les organisations internationales qui peuvent traduire ce terme ou conserver le terme anglais de Trustworthy AI

Le second livrable du *HLEG AI* est une **liste de recommandations en matière de politiques publiques et d'investissements**, avec le même objectif de garantir une « IA de confiance ». Le document formule 33 recommandations pour orienter l'IA vers la durabilité, la croissance, la compétitivité et l'action<sup>1</sup>. Les investissements doivent s'appuyer sur quatre bases : le secteur privé, le secteur public, la société civile et la recherche. Le groupe identifie **huit leviers** que les pouvoirs publics peuvent utiliser pour développer le potentiel de l'IA : l'éducation et la formation, la gouvernance et la régulation, le financement et l'investissement, le secteur privé, la société civile, le secteur public, la recherche ainsi que les données et l'infrastructure. Ce document fait apparaître l'approche pluridisciplinaire que retient l'Union européenne : plutôt que de réguler un secteur en particulier, le *HLEG AI* estime que l'investissement doit être transversal et intégré.

Après la publication de ces deux premiers livrables, le groupe a mis en place **un outil pratique d'évaluation**<sup>2</sup> qui permet de traduire les lignes directrices en techniques d'autoévaluation.

Enfin, le HLEG AI a réalisé un dernier livrable concernant une **approche sectorielle de politiques publiques et d'investissement** pour l'IA<sup>3</sup>. Ce dernier livrable était particulièrement destiné à l'**Alliance européenne pour l'IA** (*European AI Alliance*), une coalition créée par la Commission européenne en 2018 dans le but « d'ouvrir le dialogue sur l'intelligence artificielle ». Cette alliance est composée de 6 000 parties prenantes de la société civile européenne (citoyens, représentants des consommateurs et des entreprises, syndicats, établissements de recherche, autorités et experts).

Après la publication de ce dernier livrable, le mandat de l'AI HLEG est arrivé à son terme en juillet 2020 et seule l'Alliance a continué ses activités.

### *c) Le Livre blanc de la Commission européenne sur l'IA*

La phase de consultation des experts et de la société civile organisée par les institutions européennes a été conclue par la publication par la Commission européenne d'un **Livre blanc sur l'IA** de 30 pages sous-titré « *Une approche européenne axée sur l'excellence et la confiance* »<sup>4</sup> en février 2020. Il a été soumis à consultation publique jusqu'au 19 mai 2020 et discuté lors de la deuxième assemblée de l'Alliance européenne pour l'IA en octobre 2020.

---

<sup>1</sup> « Policy and investment recommendations for trustworthy AI », 2019, <https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence>

<sup>2</sup> « Évaluation finale pour une IA digne de confiance (ALTAI) », 2019 : <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>

<sup>3</sup> « Considérations sectorielles sur les recommandations en matière de politique et d'investissement », 2019, <https://futurium.ec.europa.eu/en/european-ai-alliance/document/ai-hleg-sectoral-considerations-policy-and-investment-recommendations-trustworthy-ai>

<sup>4</sup> Livre blanc de la Commission européenne, 2020, « Intelligence artificielle. Une approche européenne axée sur l'excellence et la confiance » : [https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust\\_en](https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en)



Ce rapport souligne les **bénéfices** apportés par l'IA, mais aussi les **risques** de ces technologies. Il appelle donc à un **cadre juridique spécifique pour l'IA** en Europe, en vue de créer un « **écosystème de confiance** » unique en son genre, garantissant le respect des principes du droit de l'UE, notamment ceux qui protègent les droits fondamentaux et les droits des consommateurs, en particulier vis-à-vis des systèmes d'IA à haut risque. Avec ce cadre juridique, qui vise aussi à harmoniser les efforts aux niveaux européen, national et régional, par un partenariat entre les secteurs privé et public, l'UE pourra construire un « **écosystème d'excellence** » tout au long de la chaîne de valeur de l'IA.

Le livre blanc mentionne aussi quelques futures **obligations pour les structures développant des systèmes d'IA à haut risque** : s'assurer que les jeux de données limitent les risques et évitent les discriminations ; documenter les données utilisées pour entraîner les algorithmes, dont les techniques et méthodologies de conception et d'entraînement ; développer des systèmes robustes dès le stade de la conception puis pendant leur cycle de vie ; une supervision humaine permettant une intervention en temps réel et laissant la possibilité de désactiver le système.

Le Comité économique et social européen (CESE) a rendu un rapport sur ce livre blanc<sup>1</sup>, dont la rapporteure était Catelijne Muller, membre du Comité mais aussi présidente de l'ONG ALL AI et membre de l'AI HLEG de la Commission européenne. Il **regrettait la focalisation du document sur la seule intelligence artificielle axée sur les données** (l'IA connexionniste) et réclamait une **nouvelle génération de systèmes d'IA fondés sur la connaissance et le raisonnement**. Cette réflexion fait écho aux préoccupations de vos rapporteurs. Le CESE demandait notamment, outre un renforcement des investissements, une **pluridisciplinarité** dans la recherche, une **approche socio-technique des technologies**, une association de **toutes les parties prenantes** à la discussion et une **éducation du grand public**.

## **2. De la proposition de règlement du 21 avril 2021 à la juxtaposition de deux dispositifs à la suite des amendements adoptés**

La législation européenne sur l'intelligence artificielle (ou *AI Act*) a pour origine une **proposition de règlement du Parlement européen et du Conseil transmis par la Commission européenne le 21 avril 2021**<sup>2</sup>. Après un dialogue entre les deux législateurs de l'Union et la Commission (appelé trilogue), un consensus autour d'un texte est trouvé le 9 décembre 2023.

---

<sup>1</sup> Rapport du Comité économique et social européen en 2020 : <https://www.eesc.europa.eu/fr/our-work/opinions-information-reports/opinions/livre-blanc-sur-lintelligence-artificielle>

<sup>2</sup> Proposition de règlement du Parlement européen et du Conseil du 21 avril 2021 établissant des règles harmonisées concernant l'intelligence artificielle et modifiant certains actes législatifs de l'Union : <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>



Ce compromis est adopté par le Parlement européen le 13 mars 2024, puis par le Conseil de l'Union européenne le 21 mai 2024<sup>1</sup>. Ce projet est devenu le règlement du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle.

Une **évolution importante** doit être notée **entre le texte initial et la version finale du règlement** à la suite des amendements adoptés lors de la discussion du projet de la Commission européenne par le Conseil et le Parlement européen. Le dialogue entre les institutions européennes est un processus habituel au sein du processus législatif de l'Union européenne où des compromis doivent être trouvés pour obtenir une position acceptable par la Commission, le Conseil et le Parlement.

Toutefois, le dialogue entre les institutions européennes a dans ce cas précis mené à **un texte final assez différent du texte initialement proposé par la Commission** européenne. Alors que le premier texte visait une **régulation de l'IA orientée vers les usages de la technologie plutôt que vers la technologie elle-même, selon des niveaux de risques**, le Conseil et le Parlement européen ont ajouté des dispositions différentes de cette approche initiale : en effet, il a été décidé que certains « modèles à usage général », en l'occurrence surtout les **modèles de fondation les plus puissants**, fassent l'objet d'un **encadrement plus strict** par l'*AI Act*, parfois du simple fait de la puissance de calcul nécessaire à l'entraînement du modèle. Ainsi, il est possible de séparer au sein des dispositions du texte final les dispositions initialement prévues par la Commission européenne et les ajouts du Parlement européen et du Conseil.

*a) Le volet issu du projet initial d'AI Act en 2021 : une régulation des usages selon leurs risques plutôt qu'une régulation de la technologie elle-même*

La première partie du texte présente le projet initial de règlement, à savoir une **régulation de l'IA basée sur le niveau de risque engendré par son utilisation et non par la technologie elle-même**. Ainsi, la technologie n'est pas considérée comme dangereuse ou risquée en elle-même, mais pourrait l'être dans certains contextes d'utilisation. Les obligations pèsent sur les fournisseurs, les déployeurs, les fabricants et les mandataires de fournisseurs de systèmes d'IA<sup>2</sup>.

---

<sup>1</sup> Rectificatif à la position du Parlement européen arrêté en première lecture le 13 mars 2024 en vue de l'adoption du règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle : [https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01\\_FR.pdf](https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_FR.pdf)

<sup>2</sup> Cf. le champ d'application de l'*AI Act* défini dans son article 2 et les définitions des notions dans son article 3 : [https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=OJ:L\\_202401689](https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=OJ:L_202401689)

L'article 5 liste un certain nombre de **pratiques totalement interdites** en matière d'intelligence artificielle car estimées causer un « *risque inacceptable* ». Il s'agit notamment de l'évaluation des personnes physiques associée à un **système de crédit social** ou *social ranking*, de la **reconnaissance faciale**, ou encore de la **reconnaissance des émotions sur le lieu de travail**.

Pour les autres usages des systèmes d'IA, il existe une gradation du niveau de risque lié à leur utilisation, qui implique des règles plus contraignantes à mesure que le niveau de risque augmente.

Les usages licites présentant niveau de « *risque élevé* » font l'objet des dispositions les plus rigoureuses et sont décrits à l'article 6 du texte. Il s'agit des IA utilisées dans les **domaines de biométrie, d'infrastructures critiques, d'éducation et formation professionnelle, d'emploi, de gestion des travailleurs et d'accès à l'emploi indépendant, d'accès et jouissance de services privés essentiels et de services et prestations publiques essentiels, des services répressifs, de gestion de la migration, de l'asile et des contrôles aux frontières ainsi que d'administration de la justice et de processus démocratiques**. L'article 7 du règlement donne le droit à la Commission de modifier la liste de ces critères.

Les systèmes d'IA dont les usages ne présentent qu'un « *risque limité* » sont soumis à des obligations de transparence plus légères : les développeurs et les déployeurs doivent s'assurer que les utilisateurs finaux sont conscients qu'ils interagissent avec une IA (article 50)<sup>1</sup>.

Il est intéressant de noter que le centre de recherche sur les modèles de fondation (*Center for Research on Foundation Models* ou CRFM) de l'Université de Stanford a publié une analyse évaluant la conformité des différents modèles de fondation existants avec ce premier volet du projet de règlement<sup>2</sup>.

*b) Le volet ajouté par les co-législateurs : la régulation des modèles de fondation assortie d'un régime spécifique pour les modèles les plus puissants, dits « à risque systémique »*

Le Parlement européen a décidé d'ajouter au texte initial de la Commission une réglementation qui **ne vise pas seulement les risques liés aux usages de l'IA mais la technologie et ses modèles en eux-mêmes**, ce qui se rapproche de la réflexion conduite au sein de l'université de Stanford ainsi que de la réglementation américaine de 2023. Le Conseil et le Parlement européen ont ainsi introduit la notion de « modèle d'IA à usage général » (*general-purpose artificial intelligence model* ou GPAIM).

---

<sup>1</sup> Cf. <https://artificialintelligenceact.eu/fr/high-level-summary/>

<sup>2</sup> Cf. Rishi Bommasani et al., 2023, « Do Foundation Model Providers Comply with the Draft EU AI Act ? », *Center for Research on Foundation Models, Stanford University* : <https://crfm.stanford.edu/2023/06/15/eu-ai-act.html>

Cette notion est définie au point (63) de l'article 3 du texte :

*« Modèle d'IA, y compris lorsqu'il est entraîné à l'aide d'une grande quantité de données en utilisant l'autosupervision à l'échelle, qui présente une grande généralité et est capable d'exécuter avec compétence un large éventail de tâches distinctes, quelle que soit la manière dont le modèle est mis sur le marché, et qui peut être intégré dans divers systèmes ou applications en aval, à l'exception des modèles d'IA utilisés pour des activités de recherche, de développement ou de prototypage avant d'être mis sur le marché ».*

Dès lors qu'un modèle d'intelligence artificielle est considéré comme étant à usage général, il doit suivre **les règles minimales** visées à l'article 50 du règlement.

Au-delà, le régime juridique applicable dépend de son éventuelle catégorisation en « IA à usage général présentant un risque systémique ». L'article 51 prévoit en effet qu'un GPAIM est ainsi catégorisé si l'une des deux conditions suivantes est remplie :

- il a « des capacités d'impact élevées, évaluées sur la base d'outils et de méthodologies techniques appropriés, y compris des indicateurs et des critères de référence », ce qui implique pour le fournisseur du GPAIM de procéder aux évaluations nécessaires et de le notifier à la Commission ;

- sur la base d'une décision de la Commission, d'office ou à la suite d'une alerte qualifiée du groupe scientifique, il a des capacités ou un impact équivalents à ceux visés au point [précédent], compte tenu des critères énoncés à l'annexe XIII ».

Les sept critères de l'annexe XIII<sup>1</sup> permettant d'évaluer l'impact des modèles d'intelligence artificielle sont les suivants :

- le nombre de paramètres du modèle ;
- la qualité et la taille des données d'entraînement ;
- la quantité de calcul utilisée pour l'apprentissage ;
- les modalités d'entrée et de sortie du modèle ;
- l'évaluation des capacités du modèle ;
- l'impact du modèle sur le marché intérieur ;
- le nombre d'utilisateurs finaux.

L'article 51 établit également une présomption d'impact élevé, donc de risque systémique, lorsque **le volume cumulé de calcul** utilisé pour l'apprentissage du modèle **est supérieur à 10<sup>25</sup> opérations à virgule flottante par seconde (FLOPS)**.

Le **droit américain** avec l'*Executive Order* de 2023 a quant à lui prévu des **obligations de déclaration plus souples**, avec un **seuil de puissance de calcul dix fois plus grand, soit 10<sup>26</sup>**.

---

<sup>1</sup> Cf. <https://artificialintelligenceact.eu/fr/annex/13>

### Calculer les FLOPS utilisés pour l'entraînement d'un modèle

Le site Medium<sup>1</sup> présente une heuristique pour calculer approximativement le nombre de FLOPS utilisés pour l'entraînement d'un modèle sur la base d'une publication scientifique d'OpenAI<sup>2</sup>.

La formule utilisée est la suivante :

$$FLOPS = 6 \times N \times D$$

avec N correspondant au nombre de paramètres et D au nombre de *tokens* utilisés pour l'entraînement.

Par exemple, le modèle *Llama 3 8B* de Meta a huit milliards de paramètres et a été entraîné avec quinze milliards de *tokens*. En appliquant cette formule, une approximation du nombre de FLOPS utilisés pour l'entraînement du modèle peut faire l'objet d'un calcul :

$$6 \times 8 \text{ milliards} \times 15 \text{ milliards} = 7,2 \times 10^{23} \text{ FLOPS environ}$$

Sur la base de cette heuristique, le modèle *LLama 3 8B* se situe juste au-dessous du seuil et ne correspond donc apparemment pas à une IA à usage général présentant un risque systémique au sens de l'*AI Act* et encore moins au sens de l'*Executive order* américain. Pour mémoire, les modèles de pointe d'OpenAI ou de Google dépassent depuis 2023 le seuil de  $10^{25}$  FLOPS, avec GPT-4 d'une part et Gemini d'autre part.

Sur la base de cette condition automatique et des sept critères permettant de juger du niveau d'impact d'un modèle, **la Commission européenne peut**, d'office ou à la suite d'une alerte qualifiée du groupe scientifique, **faire entrer un modèle d'IA dans la catégorie des IA à usage général présentant un risque systémique**.

Les modèles à usage général présentant des risques systémiques sont soumis à des **obligations particulières**. Ces obligations sont au nombre de quatre et sont définies à l'article 55 du règlement. Ainsi, les fournisseurs de GPAIM présentant un risque systémique doivent :

- procéder à **l'évaluation du modèle** conformément à des protocoles et à des outils normalisés reflétant l'état de l'art, y compris effectuer et documenter des **tests contradictoires** du modèle en vue d'identifier et d'atténuer les risques systémiques ;

- évaluer et atténuer les éventuels **risques systémiques** au niveau de l'Union, y compris leurs sources, qui peuvent résulter du développement, de la mise sur le marché ou de l'utilisation de modèles d'IA à usage général présentant un risque systémique ;

<sup>1</sup> Stevens, Ingrid. 2024. « Regulating AI: The Limits of FLOPs as a Metric ». Medium. <https://medium.com/@ingridwickstevens/regulating-ai-the-limits-of-flops-as-a-metric-41e3b12d5d0c>

<sup>2</sup> Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, et al. 2020. « Scaling Laws for Neural Language Models ». <http://arxiv.org/abs/2001.08361>

- garder trace, documenter et rapporter, sans délai injustifié, au bureau de l'IA et, le cas échéant, aux autorités nationales compétentes, les **informations pertinentes concernant les incidents graves** et les mesures correctives possibles pour y remédier ;

- assurer un niveau adéquat de **protection de cybersécurité** pour les modèles d'IA à usage général présentant un risque systémique et pour l'infrastructure physique du modèle.

Dans la pratique, ces obligations seront assurées par l'élaboration et le respect de « codes de conduite » définis à l'article 56 du règlement. Le respect de ces obligations sera également assuré par des normes auxquelles les fournisseurs d'IA présentant des risques systémiques devront se conformer.

Les **critiques concernant l'AI Act** portent principalement sur cette **seconde partie du texte** relative aux IA à usage général, qui a provoqué un **blocage de la part de la France et de l'Allemagne**<sup>1</sup>.

En effet, la France et l'Allemagne, qui accueillent des entreprises créant des modèles de fondation comme MistralAI en France ou Aleph Alpha en Allemagne, craignaient que cette nouvelle portée de la réglementation freine l'innovation et donc la compétitivité de leurs entreprises.

Lors de leurs auditions, vos rapporteurs ont pu remarquer que les critiques de l'AI Act se maintiennent autour de cette partie du texte. Gilles Babinet, coprésident du Conseil national du numérique (CNNum) a par exemple pointé le caractère *ex ante* de la réglementation, alors qu'à l'heure actuelle on prédit mal les effets des systèmes d'IA sur le marché économique. Il considère qu'une telle réglementation précoce a tendance à **favoriser les « gros » acteurs américains face aux entreprises européennes naissantes**. Joëlle Tolédano, coprésidente du CNNum et économiste, a rejoint l'avis de Gilles Babinet en pointant du doigt le **déséquilibre entre les coûts** nécessaires au respect des normes visant à réguler les modèles d'IA à risque systémique et **le chiffre d'affaires des nouvelles entreprises** de l'IA comme Mistral, moindre que celui des géants américains.

Yoshua Bengio a dénoncé la référence aux **FLOPS** utilisés pour l'entraînement du modèle comme mesure de calcul alors que **les modèles deviendront de plus en plus performants**, et de ce fait pourront utiliser beaucoup moins de puissance de calcul pour leur entraînement tout en parvenant à des résultats bien plus impressionnants.

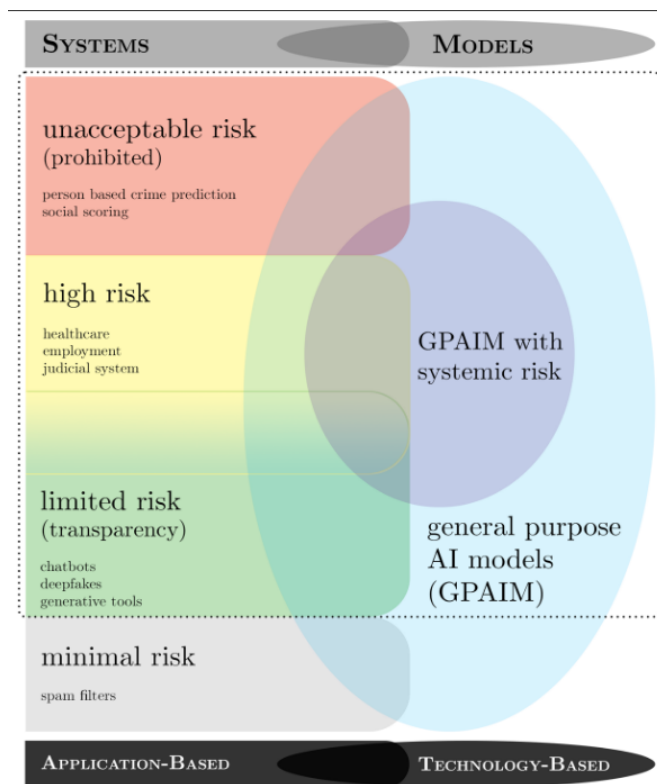
---

<sup>1</sup> Cf. Luca Bertuzzi, 2023, « AI Act : négociations bloquées à cause de divergences sur les modèles de fondation », in Euractiv : <https://www.euractiv.fr/section/intelligence-artificielle/news/ai-act-negociations-bloquees-a-cause-de-divergences-sur-les-modeles-de-fondation/>

Cette partie du règlement sur l'intelligence artificielle ne soulève toutefois pas que des critiques. Plusieurs personnes auditionnées ont souligné l'intérêt de soumettre les entreprises d'intelligence artificielle à des normes, y compris en termes de régulation des modèles, comme les chercheurs en informatique, IA et robotique Laurence Devillers, Serge Abiteboul ou Raja Chatila.

Florence G'sell, rencontrée à l'Université de Stanford, a résumé l'*AI Act* dans un graphique qui représente bien la **dualité de son dispositif – une dualité non exclusive qui conduit donc à des chevauchements possibles** en pratique – rendant encore plus complexe la mise en œuvre du règlement.

### Le double dispositif de l'*AI Act* : systèmes vs modèles d'IA



Source : Rapport de Florence G'sell, Université de Stanford, op. cit.

c) *Les autres aspects de l'AI Act : une polysynodie institutionnelle, une portée extraterritoriale, un calendrier très complexe et une normalisation désinvestie*

La **polysynodie** caractérise le dispositif institutionnel qui accompagne l'*AI Act*. Ce dernier confie la mise en œuvre des diverses dispositions du règlement à plusieurs structures dont la majorité aura une fonction de conseil plutôt qu'à la seule Commission européenne, qui aurait pu s'en acquitter elle-même en s'appuyant sur sa DG Connect.

La première sera une structure exécutive, par autonomisation au sein de la Commission européenne d'une de ses directions qui devient le **Bureau européen de l'IA** ou *EU AI Office*, composé de cinq unités et de deux conseillers. Ce nouveau centre d'expertise, dont la moitié des agents correspond à de nouveaux recrutements, a la capacité de procéder à des évaluations des modèles d'IA à usage général, de demander des informations et des mesures aux fournisseurs de modèles (y compris les codes sources des modèles et les API, mais le dialogue en amont sera privilégié) et d'appliquer le cas échéant des sanctions. La Commission européenne doit également mettre en place :

- un **Conseil européen de l'IA** ou *European Artificial Intelligence Board*, composé de représentants des États membres, qui doit conseiller la Commission européenne ainsi que le Bureau européen de l'IA et les autorités de régulation des États membres ;

- un **Forum européen consultatif sur l'IA** ou *European Artificial Intelligence Advisory Forum* qui doit conseiller le Conseil européen de l'IA, la Commission européenne et le Bureau européen de l'IA ;

- un **Panel scientifique européen d'experts en IA** ou *European Artificial Intelligence Scientific Panel of Independent Experts* qui doit conseiller le Bureau européen de l'IA spécialement sur les modèles d'IA à usage général et les risques systémiques afin de pouvoir classer les différents modèles et systèmes d'IA. C'est lui qui aidera à définir les outils, les méthodologies et les tests pertinents. Il pourra aussi éclairer les autorités de surveillance des marchés et les autorités de régulation des États membres. Enfin, il alertera le Bureau européen de l'IA en cas de modèles d'IA à usage général à risque systémique selon la réglementation de l'UE.

S'agissant de la **portée extraterritoriale du règlement**, il faut souligner que l'*AI Act* s'applique non seulement à tout acteur qui fournit, distribue ou déploie des **systèmes d'IA utilisés dans l'UE**, qu'ils soient conçus dans l'UE ou dans un pays tiers, mais couvre aussi le champ des systèmes d'IA conçus et utilisés hors UE si les **résultats** générés par le système sont destinés à être utilisés dans l'UE.

Le calendrier de mise en œuvre des dispositions du règlement frappe par sa **très grande complexité** : de 2024 à 2031, différentes étapes de déploiement de l'*AI Act* rythment chaque année. **Parler d'usine à gaz est à cet égard un euphémisme.**



### **Le calendrier très complexe de mise en œuvre de l'AI Act**

Le règlement du **13 juin 2024** établissant des règles harmonisées concernant l'intelligence artificielle est publié au Journal officiel de l'Union européenne le **12 juillet 2024**. Il s'agit de sa notification officielle. Le **1<sup>er</sup> août 2024** est la date de son entrée en vigueur, donc de son application mais en fait, ses dispositions ne s'appliqueront que plus tard et progressivement. Le **2 novembre 2024** est la date théorique d'expiration du délai au terme duquel les États membres identifient et rendent publique la liste des autorités nationales responsables de la protection des droits fondamentaux et en informent la Commission et les autres États membres.

Les interdictions relatives à certains systèmes d'IA (prévues aux chapitres I et II) commencent à s'appliquer le **2 février 2025**. Le **2 mai 2025** les codes de bonnes pratiques de la Commission doivent être prêts. Les règles relatives aux organismes notifiés (chapitre III, section 4), aux modèles d'IA à usage général (chapitre V), à la gouvernance (chapitre VII), à la confidentialité (article 78) et aux sanctions (articles 99 et 100) commencent à s'appliquer le **2 août 2025**. À la même date, si le code de pratique ne peut être finalisé ou si l'*AI Office* le juge inadéquat, la Commission peut fournir des règles communes pour la mise en œuvre des obligations des fournisseurs de modèles d'IA à usage général par le biais d'actes d'exécution. Toujours le **2 août 2025**, les États membres désignent les autorités nationales compétentes (autorités de notification et autorités de surveillance du marché), les communiquent à la Commission et mettent leurs coordonnées à la disposition du public ; fixent les règles relatives aux sanctions et aux amendes, les notifient à la Commission et veillent à ce qu'elles soient correctement mises en œuvre ; et font rapport à la Commission sur l'état des ressources financières et humaines des autorités nationales compétentes. Les fournisseurs de modèles d'IA à usage général mis sur le marché ou mis en service avant cette date doivent à partir du **2 août 2025** commencer à se conformer aux exigences du règlement, avec l'horizon d'une conformité d'ici le **2 août 2027**.

Avant le **2 février 2026**, la Commission doit fournir des lignes directrices précisant la mise en œuvre pratique de l'article 6, y compris le plan de surveillance après la mise sur le marché. Le **2 août 2026**, les autres dispositions du règlement sont applicables, à l'exception de l'article 6, paragraphe 1. Le règlement s'applique aux exploitants de systèmes d'IA à haut risque (autres que les systèmes visés à l'article 111, paragraphe 1), mis sur le marché ou mis en service avant cette date. Toutefois, il ne s'applique qu'aux systèmes dont la conception est modifiée de manière significative à partir de cette date. À la même date, les États membres veillent à ce que leurs autorités compétentes aient mis en place au moins un bac à sable réglementaire opérationnel en matière d'IA au niveau national.

Le **2 août 2027**, les fournisseurs de modèles d'IA à usage général mis sur le marché avant le **2 août 2025** doivent avoir pris les mesures nécessaires pour se conformer aux obligations prévues par le règlement avant cette date. Les systèmes d'IA composants de systèmes d'information à grande échelle qui ont été mis sur le marché ou mis en service avant cette date doivent être mis en conformité avec le règlement d'ici au 31 décembre 2030.

Le **2 août 2028**, la Commission évalue le fonctionnement de l'*AI Office* et l'impact et l'efficacité des codes de conduite volontaires (puis tous les trois ans). La Commission évalue et fait rapport au Parlement européen et au Conseil sur la nécessité de modifier les rubriques de l'annexe III, le système de supervision et de gouvernance et la liste des systèmes d'IA nécessitant des mesures de transparence supplémentaires à l'article 50. La Commission présente enfin un rapport sur l'état d'avancement des « résultats de la normalisation » qui couvrent le thème du développement économe en énergie de modèles d'IA à usage général. Ce rapport doit être soumis au Parlement européen et au Conseil, et rendu public.

Le **1<sup>er</sup> décembre 2028** (soit 9 mois avant le 1<sup>er</sup> août 2029), la Commission doit établir un rapport sur la délégation de pouvoir décrite à l'article 97.

Le **1<sup>er</sup> août 2029**, le pouvoir de la Commission d'adopter les actes délégués prévus par huit articles expire, à moins que cette période ne soit prolongée : la délégation de pouvoir sera alors, par défaut, prolongée pour des périodes récurrentes de 5 ans, à moins que le Parlement européen ou le Conseil ne s'oppose à cette prolongation trois mois ou plus avant la fin de chaque période.

Le **2 août 2029** (puis tous les quatre ans), la Commission présente au Parlement européen et au Conseil un rapport sur l'évaluation et le réexamen du règlement.

Les fournisseurs et déployeurs de systèmes d'IA à haut risque destinés à être utilisés par les autorités publiques doivent avoir pris avant le **2 août 2030** les mesures nécessaires pour se conformer aux exigences et aux obligations du règlement.

Le **31 décembre 2030** est la date limite pour la mise en conformité avec le règlement des systèmes d'IA composants de systèmes d'information à grande échelle qui ont été mis sur le marché ou mis en service avant le **2 août 2027**.

Enfin, le **2 août 2031**, la Commission procède à une évaluation de l'application du règlement et fait rapport au Parlement européen, au Conseil et au Comité économique et social européen.

Source : d'après le site <https://artificialintelligenceact.eu/fr/implementation-timeline/>

Le texte du règlement donne une place importante à la normalisation technique des produits utilisant l'IA. Cette normalisation doit **garantir des standards de qualité** et donc **d'assurer une confiance envers ces produits** pour lesquels les citoyens semblent, pour l'heure, relativement méfiants. Ces dix normes harmonisées auront aussi pour rôle de vérifier si un système d'IA correspond *a priori* aux obligations prévues au règlement : un système qui respecte ces normes sera en effet **supposé conforme à l'AI Act**.

Plusieurs spécialistes entendus par vos rapporteurs ont exprimé leur intérêt pour le processus de normalisation technique introduit par la seconde partie de l'*AI Act*. Yann Ferguson, sociologue directeur scientifique du LaborIA d'Inria ainsi que Patrick Bezombes, conseiller pour la stratégie et la gouvernance de l'IA à l'Association française de normalisation (Afnor) et représentant de la France au CEN-CENELEC ont défendu ce point de vue.

Selon eux, il faut considérer **deux catégories d'entreprises** : d'une part, les entreprises industrielles « classiques », Patrick Bezombes parle d'entreprises de la « *safety* », d'autre part, les nouvelles entreprises du numérique. Ces deux types d'entreprises ne voient pas de la même façon le développement de leurs produits. Les entreprises classiques de l'industrie comme Airbus, Alstom ou Framatome sont habituées à une normalisation stricte de leurs *process* de production, normalisation nécessaire dans leur propre intérêt pour garantir la sécurité essentielle de leurs produits et leur réputation. Ainsi, lorsque leurs produits arrivent sur le marché, le risque de défaillance est minime : la moindre panne d'un avion, d'un train ou d'un réacteur nucléaire serait catastrophique pour leur image et la confiance des consommateurs.

À l'inverse, les nouvelles entreprises du numérique, notamment les MAAAM, sont moins confrontées à ce type de problèmes, et s'autorisent des marges d'erreur élevées, d'autant plus qu'elles peuvent se permettre la **mise sur le marché de produits présentant des défauts qui pourront être corrigés par la suite, grâce à une mise à jour logicielle**. Dans ces entreprises, la culture de respect des normes est moins développée que dans les entreprises classiques, surtout celles du secteur industriel, où elle est présente dans toutes les étapes de la conception à la fabrication des produits.

De ce point de vue, le règlement sur l'intelligence artificielle peut être perçu comme **un moyen d'encourager les entreprises du numérique à changer de culture et à respecter des normes strictes lorsqu'elles commercialisent un produit**. Patrick Bezombes a particulièrement rappelé l'importance de ces normes qui permettent de créer de la confiance envers un produit, ce qui est essentiel pour la présence sur un marché.

Il relativise l'idée selon laquelle cette normalisation pénaliserait les petites entreprises par rapport aux géants du numérique. En effet, une défaillance majeure d'un produit sur le marché provoque une crise de la confiance sur l'ensemble du marché, tant à l'égard des grandes entreprises que des plus petites. Le cas des véhicules autonomes l'illustre. Aussi, en limitant le risque de défaillances sur les produits utilisant de l'intelligence artificielle, **la normalisation permet d'éviter ces crises de confiance**. La confiance générée par cette normalisation, créatrice de stabilité, **bénéficie à l'ensemble des acteurs du marché**.

Pour mieux comprendre comment les normes relatives à l'IA sont en train d'être définies, il faut comprendre qui est chargé de cette normalisation.

En Europe, le Comité européen de normalisation (CEN) et le Comité européen de normalisation en électronique et en électrotechnique (CENELEC) sont chargés de préparer les normes relatives à l'IA prévues par l'*AI Act*. Un comité commun ou *Joint Technical Committee* (JTC CEN-CENELEC) a été mis en place pour l'IA, le JTC 21.

Il doit produire des **livrables normatifs et des lignes directrices** à destination des autres comités techniques concernés par l'intelligence artificielle. Ce JTC doit également examiner l'adoption éventuelle de normes internationales pertinentes ainsi que les normes d'autres organisations compétentes, comme celles du JTC ISO-IEC 1 et ses sous-comités, tels que le SC 42.

Patrick Bezombes, responsable IA à l'Afnor représente la France au JTC 21 du CEN-CENELEC, dont il est le vice-président. L'architecture complexe des organismes de normalisation est résumée dans le tableau suivant.

**Tableau des agences de normalisation dans le monde, en Europe et en France**

	Normalisation des technologies de l'information	Normalisation de l'électrotechnique	Normalisation des télécommunications	Autres organisations sectorielles
<b>Monde</b>	Joint technical committees (JTC) de l'ISO et de l'IEC. Son sous-comité 42 (SC 42) est chargé de la <b>normalisation de l'IA</b>		Union internationale des télécommunications (UIT ou ITU), organe chargé de la normalisation des sujets de l'IT et de la communication en réseau	Institute of Electrical and Electronics Engineers (IEEE) pour l'informatique  Society of Automotive Engineers (SAE) pour les voitures autonomes
	International organization for standardization (ISO), chargée initialement des seuls sujets d'IT mais aujourd'hui chargée d'une normalisation plus large	International Electrotechnical Commission ou Commission électrotechnique internationale (IEC ou CEI), chargée de la normalisation dans le secteur de l'électrotechnique		
<b>Europe</b>	CEN-CENELEC Management Centre (CCMC), secrétariat commun au CEN et au CENELEC Joint technical committees (JTC) dont le JTC CEN-CENELEC 21 sur l'IA		European Telecommunications Standards Institute (ETSI), miroir de l'UIT au niveau européen  (marginalisé depuis la crise sur les normes du système Galileo, n'intervient pas sur la normalisation prévue par l'AI Act)	Aucun organisme reconnu officiellement par l'Union européenne
	Comité européen de normalisation (CEN), miroir de l'ISO au niveau européen	Comité européen de normalisation en électronique et en électrotechnique (CENELEC), miroir de l'IEC au niveau européen		
<b>France</b>	Association française de normalisation (Afnor), représentant la France à l'ISO et au CEN-CENELEC en matière d'IA			Union technique de l'électricité (UTE)  Centre technique industriel (CTI)

Légende : Les cases grisées sont des instances de normalisation communes, des JTC ISO-IEC ou CEN-CENELEC

La normalisation réalisée dans le cadre du CEN-CENELEC, notamment dans le JTC 21, loin d'être purement technique, reflète aussi des **choix politiques et économiques**. L'implication de la France et la qualité de sa participation au processus de normalisation dépendent des acteurs qui prennent part à ce processus pour notre pays ainsi que de l'investissement des pouvoirs publics dans ce travail.

Patrick Bezombes de l'Afnor déplore par exemple que le Comité français d'accréditation (Cofrac), chargé de certifier les entreprises respectant les normes, ne s'implique pas davantage dans les discussions relatives à la normalisation prévue par l'*AI Act*.

Il invite également à se méfier des positions tenues par des pays qui n'ont pas fait le choix d'envoyer des représentants « indépendants » au sein du JTC, soit parce qu'ils n'ont pas d'entreprises nationales de taille importante dans le domaine de l'IA, soit parce que les entreprises étrangères, souvent américaines ou parfois chinoises, ont investi ce champ de la normalisation.

En effet, derrière la représentation des 26 autres États membres dans les comités de normalisation, Patrick Bezombes déplore qu'on retrouve souvent des **représentants de grands groupes du numérique extra-européens** qui défendent leurs intérêts et tentent d'influencer les discussions en faveur d'une standardisation plus légère, de moins bonne qualité et plus conforme à leurs intérêts et à leurs attentes envers l'UE.

Il a fourni à vos rapporteurs une liste éclairante des **liens d'intérêt qui unissent plusieurs des membres du JTC 21 du CEN-CENELEC sur la normalisation de l'intelligence artificielle à des grandes entreprises étrangères du secteur du numérique**.

Au total, le **chantier stratégique de la normalisation est désinvesti** par les gouvernements des États membres et par la Commission européenne alors qu'il devrait être surveillé de très près par les pouvoirs publics, les administrations nationales et les institutions européennes.

### 3. Une gouvernance européenne de l'IA à compléter

#### *a) Mobiliser les entreprises et élaborer de la Soft Law : l'AI Pact et les bonnes pratiques*

En amont de l'opposabilité des dispositions de l'*AI Act*, la Commission européenne a proposé en septembre 2024 un **Pacte pour l'IA** ou *AI Pact* reposant sur les engagements volontaires des entreprises afin de créer une communauté collaborative qui facilite la mise en œuvre proactive de certaines mesures ainsi que la communication<sup>1</sup>.

---

<sup>1</sup> La citation originale : « the Commission is promoting the AI Pact, to help stakeholders prepare for the implementation of the AI Act. The AI Pact, is structured around two pillars : gathering and exchanging with the AI Pact network, facilitating and communicating company pledges », cf. <https://digital-strategy.ec.europa.eu/en/policies/ai-pact>

Selon la page d'un site officiel du gouvernement, ce « *pacte européen sur l'IA ressemble à une vaste blague* »<sup>1</sup>. L'UE qui chercherait « *par tous les moyens à rallier les entreprises à sa grande cause, à savoir l'application de son règlement sur l'intelligence artificielle* » utiliserait ce pacte d'engagements volontaires « *dont personne ne vérifiera s'ils sont respectés pour broser ses signataires dans le sens du poil* ». OpenAI serait même un « *parfait exemple de l'hypocrisie* » des grandes entreprises du numérique puisque « *la société la plus en vogue du marché se targue de soutenir les priorités fondamentales du pacte* » alors que son propre comité d'éthique, de sûreté et de sécurité est composé de membres de son conseil d'administration<sup>2</sup>.

Vos rapporteurs soulignent que **plusieurs entreprises ont tout de même refusé de rejoindre ce pacte** pourtant proposé sous la forme de simples engagements volontaires, marquant par-là leur opposition à la démarche de la Commission européenne, à l'instar d'Apple et de Meta<sup>3</sup>.

Parallèlement à ce Pacte pour l'IA, l'UE débute ses travaux sur la **rédaction de codes de bonnes pratiques** qui doivent aboutir au plus tard le 2 mai 2025. La Commission européenne et son Bureau de l'IA ont ainsi lancé le 30 septembre 2024 ces travaux, qui devraient s'appuyer sur un panel scientifique d'experts indépendants, avec notamment pour objectif de **préciser les règles de classification et le cadre applicable aux modèles d'IA à usage général en distinguant parmi eux ceux présentant un risque systémique** ; de proposer des **outils d'évaluation des risques et des mesures d'atténuation** ; d'établir des **modalités de transparence et de respect des droits d'auteur**.

b) *Deux projets de directive complémentaires à l'AI Act : l'intelligence artificielle comme produit sur le marché unique*

En parallèle de l'adoption de l'AI Act, les institutions européennes discutent de **deux projets de directives pour définir des règles en matière de responsabilité civile extracontractuelle** pour l'IA. Ces textes, qui ont pour objet de protéger les utilisateurs, sont complémentaires de l'AI Act – surtout axé sur la prévention des risques – et y renvoient directement<sup>4</sup>.

Alors que les règles nationales en matière de responsabilité pour faute obligent la victime à prouver l'existence d'un acte préjudiciable ou d'une omission de la part de la personne qui a causé le dommage, le fonctionnement des systèmes d'IA rend difficile une telle identification.

---

<sup>1</sup> Cf. la page suivante : <https://www.bercynumerique.finances.gouv.fr/le-pacte-europeen-sur-lia-signe-par-une-centaine-dentreprises-ressemble-une-vaste-blague>

<sup>2</sup> Cf. cet article du 29 mai 2024 « Le Board d'OpenAI s'approprie les questions de sécurité et balaie les critiques d'anciens salariés » : <https://www.usine-digitale.fr/article/le-board-d-openai-s-approprie-les-questions-de-securite-et-balaie-les-critiques-d-anciens-salaries.N2213540>

<sup>3</sup> Cf. cet article du 26 septembre 2024 « Apple and Meta snub AI safety pact. Amazon, Google, Microsoft and OpenAI among signatories of EU pact » : <https://www.independent.co.uk/tech/apple-and-meta-snub-ai-safety-pact-b2619542.html>

<sup>4</sup> La prévention et la protection à l'égard des risques forment deux volets classiques des politiques de la maîtrise des risques.

Ceci est particulièrement préoccupant pour les TPE et PME qui ne peuvent pas se permettre d'avancer des frais dans une procédure pour obtenir des réparations. Ainsi, il est proposé d'**adapter la politique de responsabilité du producteur du fait des produits défectueux** en visant une **harmonisation des régimes de responsabilité pour l'IA** au sein de tous les pays de l'UE.

La directive 2022/0302<sup>1</sup> est une révision de la directive 85/374/CEE. Elle considère les systèmes d'IA et les biens dotés d'IA comme **des « produits » sur le marché**. De ce fait, les personnes lésées par un système d'IA défectueux peuvent obtenir des réparations **sans qu'elles n'aient à prouver la faute du fabricant**. Le demandeur doit prouver la défectuosité du produit, le dommage subi et le lien de causalité entre la défectuosité et le dommage.

La directive 2022/0303<sup>2</sup> introduit de **nouveaux concepts dans le cas de responsabilité civile en réparation de dommages causés par une faute extracontractuelle**. La logique de cette directive est proche de celle citée ci-dessus, elle vise aussi à faciliter l'obtention de réparations. D'une part, un demandeur lésé peut **demandeur la divulgation d'éléments de preuve de la part du défendeur** qui opère un système d'IA, d'autre part, si cette demande n'est pas satisfaite, il y a **présomption de non-respect du devoir de vigilance pour le défendeur**.

Le demandeur qui a été lésé peut en effet saisir les juridictions nationales pour qu'elles ordonnent la divulgation d'éléments de preuve pertinents concernant un système d'IA à haut risque soupçonné d'avoir causé un dommage. Pour cela, le demandeur doit présenter des éléments de preuve suffisants pour étayer la plausibilité d'une action en réparation. Un défendeur qui ne divulguerait pas ces informations aux instances compétentes serait présumé avoir manqué à son devoir de vigilance pertinent, ce qui constitue une faute, à moins qu'il ne parvienne à renverser cette présomption.

Par cette **présomption réfragable**, les juridictions nationales présumant le lien de causalité entre la faute du défendeur et le résultat produit par un système d'IA ou son incapacité à produire un résultat. Pour cela, les trois conditions prévues à l'article 4 de la directive doivent être remplies :

- le demandeur ou la juridiction a présumé la faute du défendeur de manquement à son devoir de vigilance ;
- l'influence de la faute sur le résultat du système d'IA ou son incapacité à produire un résultat est « raisonnablement probable » ;
- le demandeur a pu prouver que le dommage a pour origine le résultat du système d'IA ou son incapacité à produire un résultat.

---

<sup>1</sup> Proposition de directive du Parlement européen et du Conseil relative à La responsabilité du fait des produits défectueux, 2022, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0495>

<sup>2</sup> Proposition de directive du Parlement européen et du Conseil relative à l'adaptation des règles en matière de responsabilité civile extracontractuelle au domaine de l'intelligence artificielle (directive sur la responsabilité en matière d'IA), 2022, <https://eur-lex.europa.eu/legal-content/FR/ALL/?uri=CELEX:52022PC0496>



Dans le cas des **IA à risque élevé**, le manquement du défendeur à son devoir de vigilance est encore **plus facile** à démontrer pour le demandeur : il suffit que ces systèmes ne respectent pas les obligations fixées par la réglementation sur l'IA. Ces systèmes doivent être transparents, pouvoir être contrôlés, avoir un niveau approprié d'exactitude et pouvoir être rappelés ou retirés.

*c) Le soutien européen à la recherche et à l'innovation en IA au-delà du dispositif EuroHPC*

EuroHPC est **l'entreprise commune pour le calcul à haute performance européen** (European high-performance computing joint undertaking ou EuroHPC). Cette entreprise est un **partenariat public-privé** créé par un règlement de l'Union européenne du 28 septembre 2018<sup>1</sup>, permettant la mise en commun de ressources à l'échelle européenne. L'objectif de l'entreprise est défini à l'article 3 du règlement : « *La mission de l'entreprise commune est de créer, de déployer, d'étendre et de conserver dans l'Union une infrastructure intégrée de supercalcul et de données de classe mondiale, ainsi que de mettre en place et de soutenir un écosystème hautement compétitif et innovant en matière de calcul à haute performance* ».

Pour financer ces infrastructures, l'Union européenne investit **486 millions d'euros et les États membres s'engagent à contribuer chacun à hauteur de 10 millions d'euros au minimum**. Les entreprises privées, quant à elles, doivent financer à due proportion l'EuroHPC. Le texte distingue deux types de supercalculateurs selon leur puissance de calcul : les pétaflopique (de 1 pétaflop, soit  $10^{15}$  FLOPS, à 100 pétaflops soit  $10^{17}$  FLOPS), d'une part, et les pré-exaflopique (de 100 pétaflops à un exaflop, soit  $10^{18}$  FLOPS), d'autre part. L'entreprise permet aux pays de financer 50 % des coûts d'acquisition et 50 % des coûts d'exploitation des supercalculateurs pré-exaflopiques et 35 % des coûts d'acquisition des supercalculateurs pétaflopiques.

En contrepartie de ces financements, **l'Union européenne dispose de 50 % du temps d'accès aux supercalculateurs**, proportionnellement à sa contribution financière, qu'elle distribue entre chaque État membre participant au consortium européen.

Le soutien européen à la recherche et à l'innovation en IA s'étend également à d'autres initiatives. Le 10 septembre 2024, la Commission européenne a lancé un appel à propositions pour la mise en place de **fabriques d'IA** ou *AI Factories*. Un réseau européen d'écosystèmes d'IA conjuguant puissance de calcul, accès aux données et talents des développeurs devrait ainsi éclore en 2026. La Commission européenne compte sur l'accélération du développement d'**espaces européens communs des données**, mis à la disposition de la communauté de l'IA.

---

<sup>1</sup> Règlement (UE) 2018/1488 du Conseil du 28 septembre 2018 établissant l'entreprise commune pour le calcul à haute performance européen, 2018, <http://data.europa.eu/eli/reg/2018/1488/oj/fra>

L'Union européenne a également annoncé au début de l'année 2024 un **soutien financier à l'IA générative**, accordé par la Commission dans le cadre d'Horizon Europe et du programme pour une Europe numérique. Ce paquet mobiliserait environ 4 milliards d'euros d'investissements publics et privés d'ici à 2027.

Dans la continuité du consortium « AI4EU » lancé en 2019, l'**initiative « GenAI4EU »** devrait soutenir le développement de nouveaux cas d'utilisation et d'applications émergentes liés à l'IA générative dans 14 écosystèmes industriels européens, ainsi que dans le secteur public. Les domaines d'application comprennent la robotique, la santé, les biotechnologies, l'industrie manufacturière, la mobilité, le climat et les mondes virtuels.

L'Union européenne poursuivra par ailleurs la **promotion des investissements publics et privés dans les *start-up*** dans le domaine de l'IA, à l'aide de soutien en capital-risque ou en fonds propres, y compris au moyen de nouvelles initiatives dans le cadre du programme d'accélération du CEI et d'InvestEU.

En outre, deux consortiums pour une infrastructure numérique européenne (EDIC) sont mis en place : d'une part, une **Alliance pour les technologies du langage** (ALT-EDIC qui doit promouvoir la diversité linguistique et la richesse culturelle de l'Europe et soutenir l'élaboration de grands modèles de langage européens par une infrastructure commune remédiant à la pénurie de données langagières européennes pour l'entraînement des IA), d'autre part l'EDIC «**CitiVERSE**», qui utilisera des outils d'IA de pointe pour les *Smart Cities* afin de mettre au point des **jumeaux numériques**, aidant les villes à simuler et à optimiser certaines de leurs politiques, notamment la **gestion des transports et des déchets**.

## D. PANORAMA D'AUTRES RÉGULATIONS NATIONALES DANS LE RESTE DU MONDE

### 1. Aux États-Unis, une régulation inachevée

a) *Au niveau fédéral : de la stratégie de 2016 à l'Executive Order présidentiel d'octobre 2023*

Depuis la **stratégie fédérale américaine pour l'IA** et sa vingtaine de recommandations dévoilée en **octobre 2016** par le **Président Barack Obama**<sup>1</sup>, **aucune loi encadrant l'IA n'a été votée aux États-Unis**. Il est vrai qu'en octobre 2016, le Président américain n'était pas favorable à l'adoption d'un cadre législatif contraignant : « *si vous parlez à Larry Page, le cofondateur de Google ou aux autres, leur réaction en général, et on peut les comprendre, c'est "la dernière chose que nous voulons c'est que des bureaucrates viennent nous ralentir pendant que nous chassons la licorne"* ».

En dépit de ce discours, **plusieurs projets de « bills » ont été déposés au cours de ces dernières années**, comme l'ont expliqué les sénateurs et les représentants américains rencontrés par vos rapporteurs au Congrès à Washington : la sénatrice Marsha Blackburn du Tennessee, le sénateur Mike Rounds du Dakota du Sud, co-président du groupe sénatorial sur l'IA, le sénateur Todd Young de l'Indiana et la représentante Anna Eshoo de Californie.

Tous ne sont **pas convaincus de l'intérêt de légiférer au niveau fédéral**. Seule Anna Eshoo se prononce pour une législation bipartisans fédérale complète en matière d'IA. Marsha Blackburn, bien qu'opposée à une régulation verticale de l'IA et aux dispositions de l'AI Act de l'UE, souhaite des **législations fédérales sectorielles**, notamment un cadre américain en faveur de la protection des données personnelles. Elle a par exemple déposé un projet de texte bipartisan avec Chris Coons, Amy Klobuchar et Thom Tillis en faveur d'un **No Fake Act, protégeant notamment les artistes**. De même Mike Rounds souhaite que le Congrès identifie une **solution législative fédérale pour protéger et rémunérer les titulaires de droits de propriété intellectuelle**, comme les détenteurs de droits d'auteur (copyrights) et de brevets.

---

<sup>1</sup> Cf. le rapport de la Maison-Blanche, Executive Office of the President, National Science and Technology Council (NSTC) & Committee on Technology, octobre 2016, « *Preparing for the Future of Artificial Intelligence* » :

[https://obamawhitehouse.archives.gov/sites/default/files/whitehouse\\_files/microsites/ostp/NSTC/preparing\\_for\\_the\\_future\\_of\\_ai.pdf](https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf) ; le plan stratégique national pour l'IA rendu par le NSTC joint au premier rapport :

[https://obamawhitehouse.archives.gov/sites/default/files/whitehouse\\_files/microsites/ostp/NSTC/national\\_ai\\_rd\\_strategic\\_plan.pdf](https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf) et l'édition de novembre 2016 de Wired, magazine dont le Président américain a été exceptionnellement le rédacteur en chef tout au long du mois d'octobre 2016 et qui comprend une longue interview exclusive par le directeur du Media Lab du MIT qui restera dans l'histoire, avec ce titre si évocateur, « Barack Obama, les réseaux de neurones, les voitures autonomes et l'avenir du monde » (titre original : « Barack Obama, Neural Nets, Self-Driving Cars, and the Future of the World ») : <https://www.wired.com/2016/10/president-obama-mit-join-ito-interview/>

De manière plus large, Todd Young travaille conjointement avec les sénateurs Schumer, Heinrich et Rounds à une **feuille de route pour orienter, dans une logique bipartisane, les travaux du Sénat américain** en matière d'IA. En 2023, Todd Young a introduit une proposition de loi pour l'Algorithmic Accountability afin d'**encadrer l'utilisation de prises de décision automatisées** dans des situations à incidence forte pour les secteurs du logement, des finances, de l'emploi et de l'éducation. Ce texte prévoit aussi la création d'un bureau de la technologie au sein de la Federal Trade Commission qui serait chargé de veiller à l'application et à la mise en œuvre de la loi.

Il n'existe donc pas encore de lois encadrant l'IA aux États-Unis mais le 30 octobre 2023, le Président des États-Unis a cependant signé un *Executive order*, décret présidentiel à portée générale, dénommé « *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*<sup>1</sup> ». Le texte vise le développement et l'utilisation de systèmes d'intelligence artificielle sûrs et fiables. Il s'agit d'un document assez long (le plus long de tous les décrets présidentiels jamais publiés) qui a principalement pour objet d'**encadrer les modèles de fondation** présentant des risques pour la sécurité, l'économie ou la santé. Il est surtout contraignant pour les administrations et agences fédérales.

Il demande aux entreprises développant ces modèles de les **notifier au gouvernement fédéral** et de **partager les résultats de leurs tests**. Ce premier axe conforte la dynamique d'**engagements volontaires** en faveur d'une IA « sûre, sécurisée et digne de confiance » obtenus en amont de l'adoption du décret par l'exécutif américain de la part de **15 entreprises** (sept entreprises en juillet 2023, soit Amazon, Anthropic, Google, Inflection, Meta, Microsoft et OpenAI rejointes par huit autres en septembre 2023, soit Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI et Stability).

Le texte prévoit aussi des **obligations de déclaration pour les modèles dont les entraînements sont jugés à risque « systémique »**, en retenant un **seuil plus élevé que l'AI Act européen** (et donc plus souple en termes de contraintes pour les entreprises que le règlement européen) puisque le seuil de puissance de calcul est fixé à **10<sup>26</sup> flop/s au lieu de 10<sup>25</sup> flop/s pour le seuil européen**.

Le décret confie, par ailleurs, au ministère du commerce la responsabilité de préparer des **lignes directrices sur l'authentification des contenus et les filigranes** ou *watermarking* afin de reconnaître les contenus générés par une IA.

---

<sup>1</sup> Cf. « *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* » : <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>

Ce ministère abrite désormais, au sein de son *National Institute of Standards and Technology* (NIST), l'**Institut américain pour la sécurité de l'IA** (*US AI Safety Institute* ou AISI), dont les responsables ont été entendus, chargé d'« élaborer des lignes directrices, évaluer les modèles et poursuivre des recherches fondamentales (...) pour faire face aux risques et saisir les opportunités de l'IA ».

L'AISI a été fondé en novembre 2023, dès le lendemain de la signature du décret présidentiel sur le développement et l'utilisation d'une intelligence artificielle sûre et fiable. En février 2024, la conseillère à la politique économique de Joe Biden, Elizabeth Kelly, a été nommée à la tête de l'Institut.

En février 2024, le gouvernement américain a aussi créé de façon complémentaire le *US AI Safety Institute Consortium* (AISIC), regroupant plus de 200 organisations telles que les géants du secteur, Google, Anthropic ou Microsoft. En mars 2024, un budget de 10 millions de dollars lui a été alloué. Les observateurs ont noté la faiblesse de ce financement, surtout si l'on considère la puissance des grandes sociétés de la Tech et de l'IA aux États-Unis. Le NIST lui-même, qui accueille l'AISI, est également connu pour son manque chronique de financement.

D'autres aspects de l'*Executive order* peuvent être mentionnés. Il appelle à une plus grande **protection des données personnelles et des droits des consommateurs** et exige de toute l'**administration fédérale et des agences gouvernementales** d'être exemplaires dans leur utilisation de l'IA.

La dernière partie du décret appelle à « des actions supplémentaires » et affirme la volonté de l'exécutif de travailler avec le Congrès « en vue d'une législation bipartisane ».

Le NIST, dans le cadre duquel travaille l'AISI, largement à l'origine du texte de l'*Executive order*, avait déjà publié en 2023, avant le décret, un ensemble de lignes directrices non contraignantes intitulé *US AI Risk Management Framework*, afin de gérer les risques liés à l'IA et à accroître la confiance dans les systèmes d'IA en formalisant une série de bonnes pratiques au stade de leur conception, de leur développement, de leur entraînement et de leur utilisation. Il peut également être relevé qu'en octobre 2022, l'*Office of Science and Technology Policy* de la Maison-Blanche a publié un « *Blueprint for an AI bill of rights* » qui fait figure de projet de déclaration des droits pour la protection des citoyens américains vis-à-vis de l'IA<sup>1</sup>. Il contient ainsi cinq principes : « des systèmes sûrs et efficaces », « une protection contre la discrimination algorithmique », « la confidentialité des données », « la notification et l'explication des décisions », « les alternatives humaines, la considération et le retour en arrière ».

---

<sup>1</sup> White House Office of Science and Technology Policy (OSTP), octobre 2022, « *Blueprint for an AI bill of rights* » : <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>

Pour conclure, en dehors de ces dispositifs réglementaires et de ces lignes directrices, les États-Unis n'ont pas encore de législation fédérale sur l'IA, en dépit de plusieurs propositions en ce sens. Pour autant, différents États cherchent, à une échelle plus locale, à **aller plus vite et plus loin que les initiatives fédérales**, notamment pour lutter contre les *deepfakes* pendant les élections.

*b) Au niveau des États : plusieurs projets à commencer par celui de la Californie*

Le Colorado, l'Utah, l'Illinois, le Massachusetts, l'Ohio et la Californie ont élaboré des textes pour réglementer le développement et l'utilisation de l'IA, notamment l'IA générative et les systèmes à risque élevé.

Le **projet le plus intéressant est celui de la Californie sur la régulation des modèles d'IA générative**, connu sous le nom de *California AI Bill* adopté par le Sénat et l'Assemblée de l'État, mais auquel le gouverneur Gavin Newsom a opposé son **véto le 29 septembre 2024**. Le texte, qui s'appliquait à tous les développeurs de systèmes d'IA offrant des services en Californie, indépendamment du lieu de leur siège social, n'a donc pas pu être promulgué.

Contrairement à l'AI Act de l'UE et aux lois chinoises, canadiennes et brésiliennes, ses dispositions visaient uniquement **les modèles d'IA à risque systémique, définis selon un coût d'entraînement ou un seuil de puissance de calcul** et non par les risques dans l'utilisation des systèmes. Le champ concernait ainsi les modèles d'IA ayant nécessité un coût de plus de **100 millions de dollars d'entraînement de leur seul modèle de fondation** ou de plus de **10 millions de dollars de réglage fin** (*fine-tuning*) ou encore une **puissance de calcul supérieure à  $10^{26}$  FLOPS**.

Les développeurs de tels modèles auraient alors eu à respecter de **nombreuses exigences en termes de transparence et de tests de sécurité**. En cas de non-conformité, ils se seraient exposés à des sanctions financières élevées, pouvant aller jusqu'à 10 % du coût de la puissance de calcul utilisée pour entraîner le modèle. Un modèle ayant nécessité un milliard de dollars aurait par exemple pu exposer son développeur à une amende de 100 millions de dollars. Le projet prévoyait aussi un dispositif de *Kill Switch* pour désactiver une IA en cas de problèmes.

Le véto du gouverneur le 29 septembre 2024 fait suite à une **lettre des grands groupes de l'IA, à commencer par Microsoft, Google ou Meta**, qui dénonçaient des exigences de sécurité trop vagues quant aux tests obligatoires à réaliser pour les développeurs de modèles. Les risques financiers encourus par ces entreprises en raison de l'engagement de leur responsabilité en cas de dommages causés par leurs systèmes d'IA faisaient aussi partie des motifs de leurs réserves.



Malgré ce veto qui marque la fin de l'année 2024, l'État de Californie a d'ores et déjà adopté pas moins de **17 autres nouvelles lois spécifiques ou sectorielles** dans la période récente, encadrant divers aspects des technologies d'IA, comme la désinformation, la lutte contre les *deepfakes*, la protection des données personnelles, la protection des consommateurs, la protection des enfants ou la lutte contre la pédophilie.

## 2. En Chine, un développement rapide et centralisé de l'IA et de sa régulation

### a) La politique chinoise en faveur d'une IA maîtrisée depuis 2017

La Chine a identifié l'intelligence artificielle comme un domaine stratégique dès 2015, avec des plans tels que « Internet+ » et « Made in China 2025 ». Après le plan d'action pour l'Internet Plus dans lequel l'IA occupe une place (2016-2018), **un plan ambitieux de développement pour l'IA de nouvelle génération est adopté en juillet 2017** visant à faire de la Chine le leader mondial de l'IA d'ici 2030. Les investissements publics sont fixés entre 10 et 15 milliards de dollars par an, et un cadre juridique se construit peu à peu. En 2019, une Commission nationale pour la gouvernance de l'intelligence artificielle est mise en place et a proposé huit principes de gouvernance en faveur du développement d'une IA responsable, concernant tant la protection de la vie privée ou la sécurité que le contrôle et la transparence des algorithmes<sup>1</sup>. La même année, le Consensus de Pékin a établi des principes éthiques pour l'IA, visant surtout l'éducation.

En 2021, la Commission nationale pour la gouvernance de l'intelligence artificielle a publié des normes éthiques pour l'intelligence artificielle de nouvelle génération<sup>2</sup>. Ces principes ont été suivis d'un cadre établi en 2021 par la Cyberspace Administration of China (CAC)<sup>3</sup> ainsi que d'une liste de normes définies par le National Information Security Standardization Technical Committee (dite « TC260 »)<sup>4</sup>.

---

<sup>1</sup> Cf. <https://digichina.stanford.edu/work/translation-chineseexpert-group-offers-governance-principles-for-responsible-ai/>

<sup>2</sup> The « Ethical Norms for New Generation Artificial Intelligence » promulgated by the National Professional Committee for the Governance of the New Generation of Artificial Intelligence, Sept. 25, 2021, [https://www.most.gov.cn/kjbgz/202109/t20210926\\_177063.html](https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html) en anglais, <https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/>

<sup>3</sup> « Guiding Opinions on Strengthening the Comprehensive Governance of Algorithms in Internet Information Service », Cyberspace Administration of China, September 2021, [http://www.cac.gov.cn/2021-09/29/c\\_1634507915623047.htm](http://www.cac.gov.cn/2021-09/29/c_1634507915623047.htm) et cf. aussi en anglais <https://digichina.stanford.edu/work/translation-guiding-opinions-on-strengthening-overall-governance-of-internet-information-service-algorithms/>

<sup>4</sup> « Basic Safety Requirements for Generative Artificial Intelligence Services promulgated by the National Technical Committee 260 on Cybersecurity of Standardization Administration », février 2024, <https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf> et cf. aussi en anglais <https://cset.georgetown.edu/publication/china-safetyrequirements-for-generative-ai-final/>



Un décret interministériel sur les algorithmes et leurs recommandations a été publié en décembre 2021, dans la continuité des travaux du CAC, avec une entrée en vigueur le 1<sup>er</sup> mars 2022<sup>1</sup>. Outre l'enregistrement des systèmes, la fourniture d'informations et l'évaluation de ceux-ci, le décret prévoit l'interdiction de certains usages<sup>2</sup> et, plus largement, un cadre de classification des algorithmes basé sur des facteurs tels que l'influence sur l'opinion publique et les mobilisations sociales ou leur impact sur les comportements des utilisateurs. Une loi sur la protection des informations personnelles a parallèlement été promulguée le 20 août 2021<sup>3</sup>.

Parmi les plans d'orientation existants, il faut aussi mentionner le cadre global du plan quinquennal 2021-2025. Ce dernier reconnaît l'IA comme essentielle pour la sécurité nationale et le développement du pays, avec des applications dans la cybersécurité, la défense, et des secteurs civils comme les villes intelligentes et l'agriculture intelligente. Les subventions du ministère de la science et des technologies (MOST) et de la Commission nationale du développement et de la réforme (NDRC) soutiennent notamment ces initiatives.

En 2023 et 2024, la jurisprudence chinoise a prévu que les contenus originaux générés par l'IA pouvaient conduire à reconnaître l'existence de droits d'auteur et à protéger, par exemple, les images créées (les *prompts* sont retenus comme créateurs d'une œuvre originale)<sup>4</sup>, en revanche les œuvres élaborées avec l'assistance d'IA générative qui ressemblent trop à des œuvres existantes ou, pire, qui les copient purement et simplement font l'objet d'une répression, les fournisseurs des services d'IA pouvant même voir leur responsabilité engagée<sup>5</sup>.

---

<sup>1</sup> Décret « Administrative Provisions on Algorithm Recommendation for Internet Information Services » appelé aussi « Algorithm Recommendation Provisions » :

<https://digichina.stanford.edu/work/translation-guiding-opinions-on-strengthening-overall-governance-of-internet-information-service-algorithms/>

<sup>2</sup> Il s'agit, par exemple, des modèles qui induisent des comportements immoraux ou illégaux, de ceux qui diffusent des fausses informations ou qui manipulent l'opinion ou, encore, de ceux qui peuvent affecter la santé physique ou mentale des mineurs.

<sup>3</sup> Cf. <https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/>

<sup>4</sup> Cf. <https://copyrightblog.kluweriplaw.com/2024/02/02/beijing-internet-court-grants-copyright-to-ai-generated-image-for-the-first-time/> et <https://www.natlawreview.com/article/computer-love-beijing-court-finds-ai-generated-image-copyrightable-split-united/>

<sup>5</sup> Cf. <https://www.twobirds.com/en/insights/2024/china/liability-of-ai-service-providers-for-copyright-infringement/>

*b) Un encadrement strict et assez exhaustif des IA génératives*

En novembre 2022, un décret interministériel sur l'encadrement des technologies de synthèse profondes, également pris dans la continuité des travaux du CAC<sup>1</sup>, vise les contenus engendrés par l'IA générative même si son champ d'application est plus large. Il s'agit notamment de lutter contre les fausses informations et les *deepfakes* en prévoyant un régime de sanctions. Le décret prévoit notamment l'obligation d'ajouter des filigranes ou *watermarking* pour les contenus générés par l'IA. Il peut être noté que ce décret a été publié cinq jours avant le lancement de ChatGPT.

En juillet 2023, la Cyberspace administration of China (CAC) a publié en coopération avec plusieurs ministères et administrations des **mesures administratives provisoires pour la gestion des services d'IA générative**<sup>2</sup>, prévoyant, au terme de vastes consultations, des réglementations spécifiques sur les algorithmes, les *deepfakes* et la standardisation des bases de données. Les dispositions de ce cadre récent – et postérieur à la diffusion des IA génératives auprès du grand public – sont dans la continuité des textes précédents tout en étant plus détaillées.

Les entreprises proposant des services d'intelligence artificielle sont ainsi **soumises à des obligations strictes en termes de transparence, de précision, d'efficacité et de respect de normes éthiques de leurs modèles**, ces entreprises pouvant même faire l'objet de contrôles et d'inspections périodiques les conduisant à expliquer le **détail des sources, des mécanismes de fonctionnement et des modalités d'entraînement de leurs modèles**. Sont également énumérées les attentes lors de la phase d'entraînement des algorithmes, en termes de gestion des données et d'annotation des données d'entraînement.

La **lutte contre les activités illégales est renforcée**, notamment l'interdiction de la création de contenus illégaux ou immoraux ou encore portant atteinte aux droits de la propriété intellectuelle, et, au-delà, les services d'IA générative doivent enfin respecter les **valeurs socialistes**. Les utilisateurs et leurs données sont davantage protégés.

---

<sup>1</sup> « *The Internet Information Service Deep Synthesis Management Provisions* » or « *Deep Synthesis Regulation* » promulgated by the Cyberspace Administration of China, the Ministry of Industry and Information Technology, and the Ministry Public Security, [https://www.gov.cn/zhengce/zhengceku/2022-12/12/content\\_5731431.htm](https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm) et en anglais <https://www.chinalawtranslate.com/en/deep-synthesis/>

<sup>2</sup> « *Interim Administrative Measures for Generative AI Services* », 2023 : [https://www.cac.gov.cn/2023-07/13/c\\_1690898327029107.htm](https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm) et en anglais : <https://www.chinalawtranslate.com/en/generative-ai-interim/> ainsi que les articles « *China's New AI Regulations* » : <https://www.lw.com/admin/upload/SiteAttachments/Chinas-New-AI-Regulations.pdf> et « *How will China's Generative AI Regulations Shape the Future?* » : <https://digichina.stanford.edu/work/how-will-chinas-generative-ai-regulations-shape-the-future-a-digichina-forum/>

Il est réaffirmé que les contenus générés par l'IA doivent clairement indiquer *via* des filigranes ou *watermarking* qu'ils ont été générés par ce biais. Le National Information Security Standardization Technical Committee (dite « TC260 ») a été mandaté pour **préciser les modalités de déploiement de ces techniques de marquage**. Les normes proposées en août 2023 distinguent désormais deux types de techniques de filigrane, l'une explicite, l'autre implicite, c'est-à-dire imperceptible à l'œil humain mais pouvant être extraite comme une métadonnée à travers des méthodes techniques<sup>1</sup>. Il est prévu que ce marquage implicite inclut au moins le nom du système ayant offert le service et peut contenir des détails additionnels tels qu'un identifiant unique. En cas de téléchargement du contenu sous la forme d'un fichier, les métadonnées doivent obligatoirement prévoir des informations complémentaires, telles que des détails sur le système ayant offert le service, l'heure et la date de production, ou encore un identifiant unique.

En septembre 2023, le ministère chinois de la science et de la technologie a adopté avec le concours de neuf agences gouvernementales des « Mesures pour la revue éthique des activités scientifiques et technologiques »<sup>2</sup>, qui mentionnent notamment les **évaluations éthiques rigoureuses des activités recourant à l'IA**, à travers par exemple la mise en place de comités d'éthique.

Plus récemment, le TC260 a rendu public le 29 février 2024 des « Exigences fondamentales de sécurité pour les services d'IA générative »<sup>3</sup>. Dans ce document, **31 risques sont identifiés** à travers cinq catégories : les contenus violant les valeurs socialistes, les contenus discriminatoires, les contenus portant atteinte au droit commercial, les contenus portant atteinte aux droits individuels et les contenus ne respectant pas les obligations réglementaires en termes de sécurité. Ce texte oblige les fournisseurs de services d'IA à prendre plusieurs mesures en ce sens et à mettre en œuvre différents mécanismes d'évaluation.

En complément des règles nationales, les provinces et municipalités chinoises édictent également des règles locales. Ainsi, Shenzhen, Shanghai, Guangdong, Jiangsu et Zhejiang ont adopté des mesures pour encadrer l'IA<sup>4</sup>.

---

<sup>1</sup> « Practical Guidelines for Cybersecurity Standards – Method for Tagging Content in Generative Artificial Intelligence Services » promulgated by China's National Information Security Standardization Technical Committee (« TC260 »), 2023, <https://www.tc260.org.cn/upload/2023-08-08/1691454801460099635.pdf>

<sup>2</sup> « Notice of Issuing the Trial Measures for Scientific and Technological Ethics Review, 2023, [https://www.gov.cn/zhengce/zhengceku/202310/content\\_6908045.htm](https://www.gov.cn/zhengce/zhengceku/202310/content_6908045.htm) et en anglais <https://lawinfochina.com/display.aspx?id=42015&lib=law&SearchKeyword=&SearchCKeyword=>

<sup>3</sup> « Basic Safety Requirements for Generative Artificial Intelligence Services », 2024, <https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf> et en anglais, <https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai-final/>

<sup>4</sup> Cf. par exemple les règles éthiques assez complètes adoptées par le Conseil de Shenzhen, « Regulations for the Promotion of the Artificial Intelligence Industry in Shenzhen Special Economic Zone », <https://cset.georgetown.edu/publication/regulations-for-the-promotion-of-the-artificial-intelligence-industry-in-shenzhen-special-economic-zone/>

*c) Un rival sérieux des États-Unis intéressé par l'AI Act*

La Chine est **leader dans des domaines clés d'application de l'IA** tels que la santé, les voitures autonomes, l'automatisation industrielle, les villes intelligentes ou surtout les technologies de reconnaissance faciale. La Chine fait la course en tête avec les États-Unis en **apprentissage automatique** et les devance même en **nombre de publications**, en **analyse de données avancée** et en **algorithmes et logiciels d'accélération**. Les États-Unis sont toujours en avance dans la conception des circuits intégrés et dans le traitement du langage naturel. La Chine vise donc à développer des modèles de LLM rivalisant avec ChatGPT, comme WuDao 2.0 et ErnieBot.

Elle cherche aussi à **s'affranchir de sa dépendance aux GPU américaines**, surtout que les États-Unis interdisent depuis trois ans l'exportation vers la Chine des puces les plus avancées<sup>1</sup> et même des outils de gravure. La Chine recommande donc depuis 2024, à travers des instructions de son ministère de l'industrie et des technologies de l'information, de privilégier ses alternatives nationales, comme les puces conçues par Huawei<sup>2</sup>.

Les géants de l'IA chinois, les BATXH – pour **Baidu, Alibaba, Tencent, Xiaomi, et Huawei** – dominant la chaîne de valeur de l'IA dans le pays et possèdent leurs propres centres de recherche. Huawei, par exemple, développe des puces pour pallier les restrictions à l'exportation de technologies de lithographie avancée, bien que l'accès à des fonderies avancées reste un défi.

La production scientifique en IA en Chine est en croissance, avec des recherches axées sur les algorithmes, la vision par ordinateur, les systèmes de recommandation, et les robots. Les chercheurs chinois publient plus que les chercheurs américains en apprentissage automatique. La France y est reconnue pour son haut niveau académique en IA, avec des accords de formation entre les deux pays et un intérêt des entreprises pour nos chercheurs.

Il faut, enfin, relever que la Chine a mis en place une **stratégie à long terme pour non seulement attirer et retenir ses talents mais même favoriser leurs retours**. Il s'agit essentiellement, d'une part, du plan « Made in China 2025 » dont le but est de faire de la Chine le leader mondial des technologies de pointe, créant ainsi de nombreuses opportunités pour les ingénieurs et les chercheurs et, d'autre part, du concept de « rêve chinois », un concept nationaliste visant à renforcer le sentiment d'appartenance et à encourager les Chinois de la diaspora à revenir contribuer au développement de leur pays.

---

<sup>1</sup> Ces restrictions d'exportation portent par exemple sur les GPU H100 et A100 de Nvidia, sur la puce Gaudi 2 d'Intel ou les nouveaux GPU d'AMD, cf. <https://www.usine-digitale.fr/article/nvidia-ne-pourra-plus-exporter-ses-gpu-optimises-pour-l-ia-vers-la-chine.N2183758>

<sup>2</sup> L'information vient de la presse chinoise elle-même, cf. cet article en anglais du South China Morning Post du 7 octobre 2024 : <https://www.scmp.com/tech/article/3281254/tech-war-china-advocates-use-local-ai-chips-over-those-us-powerhouse-nvidia>

Le gouvernement chinois a, de plus, mis en place des programmes pour « les Chinois de retour » en vue de faciliter le retour de chercheurs et d'entrepreneurs chinois. En outre, tout comme en Inde, les zones économiques spéciales (ZES) chinoises jouent un rôle crucial dans l'attraction des investissements étrangers et la création d'emplois qualifiés.

L'IA est, en conclusion, perçue en Chine comme un **moteur de croissance** et un **outil stratégique**. La rivalité avec les États-Unis est souvent mise en avant par les médias, tandis que la protection des données et de la vie privée n'est pas une préoccupation majeure pour le public. Les régulations européennes, comme le RGPD et l'*AI Act*, intéressent les régulateurs chinois pour leurs approches ambitieuses. Ils suivront donc avec intérêt la mise en œuvre du règlement.

### 3. Quelques autres initiatives intéressantes

Comme l'indique Florence G'Sell, auditionnée par vos rapporteurs à l'Université de Stanford, les initiatives prises en matière de régulation de l'IA permettent de distinguer **trois groupes de pays** :

- ceux qui font le choix d'adopter un **cadre strict de régulation** de l'IA, comme l'UE, la Chine, le Canada ou le Brésil ;

- ceux qui font le choix d'**avancer progressivement** dans la voie de la régulation, comme les États-Unis, le Japon, la Corée du Sud et l'Inde ;

- et ceux qui, pour le moment, **excluent la perspective de régulations contraignantes**, comme le Royaume-Uni, Israël, l'Arabie Saoudite et les Émirats arabes unis.

Ces trois cas de figure seront présentés successivement (sans l'UE, les États-Unis et la Chine vus précédemment), en mobilisant notamment les réponses aux questionnaires de vos rapporteurs et leurs déplacements ainsi que les travaux de Florence G'Sell déjà mentionnés, à travers la présentation du Canada et du Brésil, puis du Japon, de la Corée du Sud et de l'Inde et, enfin, du Royaume-Uni, d'Israël, de l'Arabie Saoudite et des Émirats arabes unis.

#### a) Le Canada

Le Canada, pays doté d'une recherche fondamentale en IA d'excellent niveau, est toujours en train de mettre au point sa **première législation complète** en matière d'IA. Le vote du projet de loi sur l'intelligence artificielle et les données (*Artificial Intelligence and Data Act* - AIDA) a pris du retard puisque, déposé en 2022, il a seulement passé le stade de la deuxième lecture fin 2023. Il s'agit d'assurer le développement et l'utilisation de systèmes d'IA responsables au Canada.

Ce projet de loi devrait se traduire *in fine* par un **cadre assez proche du règlement de l'Union européenne** du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle. Les systèmes d'IA y seront distingués des modèles de *Machine Learning*. Et un **encadrement spécifique sera élaboré pour les systèmes d'IA à usage général** (modèles d'IA à usage général dans l'Union européenne). On y retrouvera aussi une **approche basée sur les niveaux de risque**. Un système d'IA pourra être à usage général ou pas et à haut niveau de risque ou pas. **Les systèmes conjuguant usage général et haut niveau de risque seront particulièrement encadrés.**

Le projet de loi prévoit également la création d'un **commissaire à l'intelligence artificielle et aux données**, qui aura pour mission d'aider le ministre de l'innovation, des sciences et de l'industrie et sa tutelle le ministère de l'innovation, des sciences et du développement économique, à mettre en œuvre les dispositions du texte. Ce commissaire pourra prononcer des injonctions et même infliger des amendes.

Avant l'adoption définitive du projet de loi sur l'intelligence artificielle et les données, **présenté il y a déjà deux ans**, le Canada a adopté pour l'heure, en septembre 2023, un **code de conduite pour le développement et la gestion responsables des systèmes d'IA générative avancés.**

Ce code, qui repose sur la **participation volontaire des entreprises**, sert depuis un an de dispositif provisoire en attendant que le Canada finalise son cadre législatif. Il est d'une nature similaire à l'*AI Pact* de l'Union européenne.

Le Québec, de son côté, a avancé plus vite que le gouvernement fédéral avec une réforme des lois sur la protection des renseignements personnels, obligeant à informer les utilisateurs de l'utilisation de processus de prise de décision automatisés ou qui recueillent de l'information à l'aide d'outils technologiques permettant de les identifier, de les localiser ou d'effectuer leur profilage. De plus, en janvier 2024, le Conseil de l'innovation du Québec a remis au ministère de l'économie un rapport pour un encadrement réglementaire de l'IA.

#### *b) Le Brésil*

L'action du gouvernement fédéral brésilien débute en **2019** avec le lancement de la **stratégie brésilienne d'intelligence artificielle** (dite « EBIA »). Cette stratégie consiste notamment en un travail de préparation de lois fédérales brésiennes pour **promouvoir une utilisation responsable et éthique de l'IA au Brésil**. Dès 2020, le Congrès national brésilien a commencé à examiner un **projet de loi** visant à établir le « *cadre juridique de l'intelligence artificielle* ». Ce projet de loi a été le premier d'une **série de quatre projets** proposés au Congrès. Des exceptions à cet encadrement sont prévues pour le *data mining*.

Le choix du Brésil d'établir un véritable cadre juridique pour l'intelligence artificielle a été **directement influencé par la réflexion conduite au sein de l'Union européenne** depuis 2018 puis par le contenu du projet de règlement sur l'intelligence artificielle de la Commission européenne d'avril 2021, devenu le règlement de l'Union européenne du 13 juin 2024.

En avril 2024, le Sénat brésilien a annoncé que sa commission temporaire interne sur l'intelligence artificielle avait publié un nouveau rapport préliminaire avec une proposition mise à jour pour l'un des quatre projets de loi. Une innovation clé de cette version alternative du projet de loi était la proposition d'un système de surveillance, autour d'un **système national de régulation et de gouvernance de l'IA** (SIA), coordonné par une autorité compétente nommée par le pouvoir exécutif.

Le 8 mai 2024, l'autorité brésilienne de protection des données (*Autoridade Nacional de Proteção de Dados*, ANPD) a publié une proposition d'amendement au projet de loi prévoyant la notion de « système d'IA à usage général », proche de la notion de modèle d'IA à usage général retenue par l'AI Act. Elle a en effet défini ces systèmes comme un « *système d'IA basé sur un modèle formé avec des bases de données à grande échelle, capable d'effectuer une grande variété de tâches différentes et de servir à différentes fins, y compris celles pour lesquelles il n'a pas été spécifiquement conçu ou formé, et qui peut être utilisé dans différents systèmes ou applications* ».

Le cadre juridique choisi par le Brésil pour réguler l'intelligence artificielle reprend aussi **l'approche selon les niveaux de risque**, dans le même esprit que les dispositions contenues dans le règlement de l'Union européenne du 13 juin 2024.

### *c) Le Japon*

Le Japon ne dispose pas actuellement d'une loi globale réglementant le développement ou l'utilisation des systèmes d'IA. Toutefois, le gouvernement japonais a établi plusieurs séries de **lignes directrices non contraignantes** qui sont généralement applicables à des activités précises et a encouragé les efforts volontaires des parties prenantes de l'IA.

Il aborde au fond les risques associés à l'IA de deux manières : par les **lignes directrices non contraignantes** précitées et par **l'application des lois sectorielles existantes**. Le Japon a par exemple élaboré des lignes directrices sur la protection des données et des lois sur le droit d'auteur afin de faciliter la conformité au droit des parties prenantes de l'IA. La loi japonaise sur le droit d'auteur autorise en principe l'utilisation d'œuvres protégées par le droit d'auteur pour la formation de modèles d'IA sans l'autorisation du titulaire du droit d'auteur.



En juillet 2021, le ministère de l'économie, du commerce et de l'industrie (METI) a publié un rapport indiquant que « *du point de vue de l'équilibre entre le respect des principes de l'IA et la promotion de l'innovation, et du moins à l'heure actuelle, à l'exception de certains domaines spécifiques, la gouvernance de l'IA devrait être conçue principalement avec de la soft law, favorable aux entreprises qui respectent les principes de l'IA* ». Le document précise que « *les exigences horizontales juridiquement contraignantes pour les systèmes d'IA sont jugées inutiles pour le moment* ».

Cette approche, reposant sur des lignes directrices non contraignantes, découle de la conviction que les lois contraignantes ne peuvent pas suivre le rythme rapide et la complexité du développement de l'IA et pourraient même étouffer l'innovation en matière d'IA.

En février 2024, les autorités japonaises ont toutefois lancé des **discussions sur l'élaboration d'une législation contraignante**, un cadre qui imposerait différentes obligations aux développeurs de modèles fondamentaux à grande échelle.

#### *d) La Corée du Sud*

La Corée du Sud a annoncé des objectifs ambitieux pour devenir un leader mondial de la technologie de l'IA. Elle ne dispose actuellement d'**aucune loi ou politique spécifique pour encadrer l'intelligence artificielle**. Pour parvenir à être un leader mondial, son approche de la régulation de l'IA est en effet animée par un principe qui peut être résumé par la formule « *Autoriser d'abord, réglementer ensuite* ».

Une législation pourrait néanmoins se dessiner puisque son Assemblée nationale se prépare à discuter d'un projet de loi nationale sur l'IA. En attendant, la Corée du Sud veille à la mise en œuvre de ses normes éthiques en matière d'IA et prend des mesures pour mettre en œuvre sa stratégie pour une IA digne de confiance. Les agences coréennes ont pris des mesures proactives pour établir des **lignes directrices sur l'IA** et appliquer des mesures de **protection des données personnelles** aux principaux acteurs de l'IA.

#### *e) L'Inde*

L'Inde est l'un des pays asiatiques où l'économie numérique est la plus développée. Ce secteur a connu une croissance rapide au cours des dernières années, notamment à travers une forte augmentation des services numériques. L'Inde a également fait du développement et de l'adoption de l'intelligence artificielle une **priorité dans ses initiatives politiques pour l'avenir**.

En mars 2024, le gouvernement indien a annoncé une dotation budgétaire de plus de 1,25 milliard de dollars pour la « *India AI Mission* », qui recouvrira divers aspects de l'IA, notamment les capacités du pays en termes d'infrastructures informatiques, la formation, l'innovation, les jeux de données et le développement d'une IA sûre et fiable.

Les organismes gouvernementaux indiens ont entrepris diverses actions, telles que la publication de rapports ou la promotion d'initiatives visant à favoriser des pratiques et principes responsables en matière d'IA. Ces initiatives peuvent relever de soutiens au niveau fédéral comme au niveau des États et d'autres organismes de réglementation.

L'Inde a accueilli le **Sommet du Partenariat mondial sur l'IA**, *Global Partnership on AI* ou GPAI, à New Delhi en décembre 2023 et préside le Sommet en 2024. L'un des résultats du Sommet de 2023 a été la « Déclaration 2023 des ministres », dans laquelle les membres se sont engagés à continuer à œuvrer à la promotion d'une IA sûre, sécurisée et digne de confiance dans leurs pays.

Le gouvernement indien est également conscient des éventuels risques liés à ces technologies. Après plusieurs années de délibération, il a récemment promulgué une **loi nationale sur la protection des données**. Des **cadres réglementaires** pour répondre aux risques de l'IA devraient être prochainement élaborés. Des avis ont d'ores et déjà été émis en 2023 par le ministère indien de l'électronique et des technologies de l'information (MEITY) concernant l'IA générative et les *deepfakes*. Bien qu'il existe une incertitude quant à la nature juridiquement contraignante ou pas de ces avis, ils donnent une idée du sérieux avec lequel le gouvernement indien se préoccupe de l'IA générative et de la manière dont ces préoccupations pourraient être traitées dans les futures législations et réglementations.

Ces avis, réitérant les obligations déjà existantes en vertu de la loi indienne concernant les mesures que les plateformes doivent prendre pour **identifier et prévenir la désinformation**, exigent ainsi des plateformes en ligne qu'elles prennent des mesures pour lutter contre les *deepfakes* et autres contenus de désinformation. Outre l'identification des contenus de désinformation, la loi interdit de mettre en ligne tout contenu de désinformation et oblige les plateformes à agir sur ces contenus dans les 36 heures. Il s'agit donc en réalité essentiellement d'étendre les obligations existantes aux *deepfakes*. Le MEITY a publié un message sur Twitter (rebaptisé X en juillet 2023) indiquant que ses avis de décembre 2023 ont été publiés après la tenue de deux consultations avec les parties prenantes concernées, connues sous le nom de « Dialogues numériques sur l'Inde », portant notamment sur la question de la lutte contre les *deepfakes*. Un des avis du MEITY conseille par ailleurs aux plateformes de se conformer à diverses exigences de modération du contenu en ligne afin de continuer à bénéficier d'une immunité de responsabilité (protection de la sphère de sécurité) en vertu de la loi indienne.

Enfin, un avis a été émis par le MEITY le 1<sup>er</sup> mars 2024, exigeant que « *tout modèle d'intelligence artificielle/LLM/IA générative, logiciel ou algorithme non fiable* » demande préalablement à son déploiement en Inde « *l'autorisation explicite* » du gouvernement indien. Les entreprises ont également été invitées à soumettre des rapports sur l'état de conformité de leurs systèmes avant le 15 mars 2024.

Cet avis a fait l'objet de nombreuses critiques. Outre une mise en question de sa base juridique et de sa portée précise, diverses parties prenantes ont fait valoir que ses exigences étaient anti-innovation et avaient un impact négatif sur la croissance de l'écosystème indien de l'IA. Cela a incité le ministre indien des technologies de l'information à publier à nouveau un message sur X/Twitter, pour clarifier la portée de l'avis du MEITY puis à publier **le 15 mars 2024 un nouvel avis de deux pages**, remplaçant le précédent. Le nouvel avis a **supprimé l'obligation de demander l'approbation du gouvernement pour les modèles d'IA générative** mais impose aux entreprises de soumettre un rapport sur l'état de conformité des systèmes. Le caractère juridiquement contraignant du nouvel avis reste toujours aussi incertain. Il s'agirait en fait pour les plateformes de plus de 500 000 utilisateurs enregistrés de :

- ne pas afficher ou permettre aux utilisateurs de partager des contenus illégaux ;
- ne pas autoriser de parti pris ou de discrimination ;
- ne pas menacer l'intégrité du processus électoral ;
- d'étiqueter les contenus générés par l'IA qui peuvent inclure de la désinformation ou des *deepfakes* ;
- d'utiliser des métadonnées pour identifier tout utilisateur qui modifie les informations.

L'Inde cherche, par ailleurs, à **faire revenir les talents** sur son sol<sup>1</sup>. Longtemps confrontée à un exode de ses meilleurs cerveaux, elle a entrepris des actions significatives pour inverser cette tendance :

- le programme gouvernemental « *Start up India, Stand up India* » qui vise à promouvoir l'esprit d'entreprise et à créer un écosystème favorable aux *start-up*. L'objectif est de fournir un environnement propice à l'innovation et de retenir les talents sur le territoire indien ;
- le programme « *Diaspora connect* » qui cherche à renforcer les liens entre la diaspora indienne et son pays d'origine, en encourageant le retour des talents et les investissements ;
- l'initiative « *Make in India* » qui encourage les entreprises manufacturières à investir en Inde, créant ainsi de nouveaux emplois qualifiés et des opportunités pour les ingénieurs et les chercheurs indiens ;
- les zones économiques spéciales (ZES), qui bénéficient d'avantages fiscaux et réglementaires pour attirer les investissements étrangers et créer des pôles d'excellence.

---

<sup>1</sup> Cf. les liens suivants : <https://www.globalbusinessculture.com/fr/india/the-great-indian-brain-drain> et <https://les-yeux-du-monde.fr/actualites-analysees/34055-back-to-bangalore-brain-gain-inde>

*f) Le Royaume-Uni*

L'écosystème du Royaume-Uni en matière d'IA est le **troisième au monde** derrière ceux des États-Unis et de la Chine. Depuis 2022, le secteur technologique britannique est le troisième à atteindre une valorisation de plus de 1 000 milliards de dollars, après les États-Unis et la Chine. Une fiscalité avantageuse encourage ce dynamisme (EIS, SEIS, Business Asset, Disposal Relief...) tout comme la présence de plus de 400 incubateurs. Le secteur de l'IA compte 3 170 entreprises (60 % spécialisées dans l'IA et 40 % diversifiées) dont environ 50 *start-up* dans la seule IA générative. Il générerait 3,7 milliards de livres de valeur ajoutée pour l'économie britannique dont 71 % de revenus réalisés par les très grandes entreprises (Microsoft, Google, DeepMind, IBM, etc.). Le Royaume-Uni cherche à **garder ses talents** et à **attirer les talents étrangers**.

Le **monde de la recherche y est performant et reconnu**, bien intégré au secteur privé, notamment autour d'un arc de pointe Londres-Oxford-Cambridge. Certains secteurs à fort potentiel pour l'IA comme les services financiers et les services professionnels sont puissants à Londres. 55 % des entreprises de l'IA se situent donc à Londres même. Un quart environ se trouve en périphérie de Londres.

L'approche du Royaume-Uni en matière de **régulation de l'IA** est **plutôt pro-innovation que pro-réglementation**. Les autorités ne souhaitent pas perturber leur riche écosystème. Le Royaume-Uni a annoncé en 2021 un plan sur dix ans pour devenir une superpuissance mondiale de l'IA. Ce plan met l'accent sur un investissement solide dans la recherche et le développement ainsi que sur un cadre de gouvernance qui donnera la priorité à l'innovation et à la gestion des risques.

La stratégie du gouvernement britannique se concentre principalement sur la **promotion de l'innovation**. Au lieu de mettre en œuvre une réglementation générale, le gouvernement britannique privilégie une approche souple, contextuelle et intersectorielle fondée sur des principes. Le Royaume-Uni a ainsi pris des mesures spécifiques pour soutenir la sécurité des produits, la cybersécurité et d'autres domaines et a demandé aux agences publiques de soumettre des **orientations conformes à l'approche pro-innovation de la réglementation** pour traiter les risques liés à l'IA. Le gouvernement britannique reconnaît que des exigences contraignantes pourraient éventuellement être nécessaires pour atténuer les dommages potentiels liés à l'IA, mais il a également déclaré qu'**il n'introduira une législation que lorsqu'une telle mesure lui semblera pleinement justifiée**.

L'ancien Premier ministre Rishi Sunak, dans son discours d'octobre 2023 devant la Royal Society, a résumé l'approche du Royaume-Uni par une question : « *Comment pouvons-nous rédiger des lois qui aient du sens pour quelque chose que nous ne comprenons pas encore pleinement ?* » Il a fait valoir qu'il ne se précipiterait pas pour réglementer l'IA à ce stade.

Le Royaume-Uni continue de publier **des conseils et des précautions en termes de sécurité** pour les développeurs de l'IA, inspirés par les principes de l'OCDE. Des rapports parus en avril 2024 pourraient inciter le gouvernement britannique à prendre des mesures minimales pour adopter sa propre législation ou réglementation sur l'IA, malgré les propos antérieurs du Premier ministre.

Les étapes de construction de la politique du Royaume-Uni en matière d'IA peuvent être rappelées brièvement. Dès 2015, le pays se dote d'un institut national pour la science des données et l'IA, le *Alan Turing Institute*, une entité de droit privé indépendante mais financée par le gouvernement<sup>1</sup>. En 2017, l'IA est identifiée comme l'un des secteurs prioritaires de la stratégie industrielle nationale (*UK Industrial Strategy*). En 2016, est mis en place **l'Institut pour une IA responsable** (*Responsible AI Institute* ou *RAI Institute*) avec un partenariat public-privé efficace<sup>2</sup>, qui s'est depuis déployé dans un grand réseau conduisant plusieurs initiatives dédiées à la promotion de l'IA, dans une approche multidisciplinaire et multisectorielle. En 2018, un **Bureau national de l'IA** est créé au sein du ministère de la science, de l'innovation et de la technologie (*UK Office for AI*).

Avec le *AI Sector Deal* de 2019, un paquet d'investissements de 1,3 milliard de livres est prévu. La même année, le gouvernement se dote d'un Conseil de l'IA (*UK AI Council*).

En 2020 est créé le *Digital Regulation Cooperation Forum*, un **forum de coopération volontaire entre les régulateurs du secteur numérique**. Son rôle est de coordonner sans fournir de directives à ses membres. Dans un rapport, le *Communications and Digital Committee* de la *House of Lords* estime que la création du forum va dans le bon sens mais qu'il **ne dispose pas des pouvoirs et moyens suffisants** pour répondre aux défis de la régulation de l'IA et du numérique.

En 2021 est adoptée une **stratégie nationale pour l'IA**, un plan sur dix ans pour faire du Royaume-Uni une « *superpuissance de l'IA* ». Les soutiens à cet objectif s'accroissent depuis lors. L'IA est ainsi identifiée en 2023 comme **priorité pour la stratégie de croissance** du pays. Elle est, de plus, l'une des cinq technologies prioritaires du *Science and Technology Framework*.

En 2022 est lancée une initiative, appelée *AI Standards Hub*, dirigée par le *Alan Turing Institute* et soutenue par la *British Standards Institution* et le *National Physical Laboratory*. Elle vise à orienter les débats sur la normalisation **et à développer des standards nationaux mais aussi internationaux** en rassemblant les différents acteurs de la normalisation (gouvernement, régulateurs, industrie, consommateurs, société civile, etc.).

---

<sup>1</sup> Le site du *Alan Turing Institute* : <https://www.turing.ac.uk/>

<sup>2</sup> Le site du *RAI Institute* : <https://www.responsible.ai/>

Elle repose sur quatre piliers : un observatoire qui se dote d'une base de données compilant les normes existantes ou en cours d'élaboration partout dans le monde ; une plateforme communautaire pour faciliter les connexions, la coordination, l'échange d'idées et la résolution collaborative de problèmes entre les parties prenantes ; la constitution d'un matériel d'apprentissage en ligne ainsi que des formations en présentiel afin de familiariser les acteurs avec les normes ; la construction d'une base de recherches et d'études sur les grands enjeux liés au développement de normes dans l'IA.

### Le UK AI White Paper sur l'encadrement de l'IA

Les principes clés du *AI White Paper* pour encadrer de manière souple et pro-innovation les usages de l'IA sont les suivants :

- l'encadrement ne doit pas porter sur les technologies elles-mêmes ou les modèles mais sur les usages de ces modèles et des systèmes d'IA ;
- la régulation de ces usages est confiée à des régulateurs sectoriels et pas à un régulateur unique ;
- cinq principes horizontaux centralisés et non contraignants doivent guider l'action des régulateurs, à savoir 1. la sûreté, la sécurité et la robustesse ; 2. la transparence et l'explicabilité du processus décisionnel ; 3. la compatibilité des systèmes d'IA avec les lois britanniques existantes ; 4. l'existence d'une chaîne de responsabilités clairement établie et d'une gouvernance permettant le contrôle des systèmes d'IA ; 5. la contestabilité et l'existence de voies claires de recours.

En juin 2023, dans un rapport intitulé « *Generative AI Framework for HMG* », le *Central Digital and Data Office* du Royaume-Uni a publié des **directives spécifiques sur l'IA générative** pour la fonction publique britannique. Le rapport énumère 10 principes pour guider l'utilisation responsable de l'IA générative au sein du gouvernement. Ces principes vont de recommandations éthiques (utiliser l'IA générative de manière légale, éthique et responsable) à des recommandations plus pratiques (comprendre comment gérer l'intégralité du cycle de vie de l'IA générative ou avoir les compétences et l'expertise nécessaires pour créer et utiliser l'IA générative).

Le 26 octobre 2023, le roi a donné son assentiment à la promulgation de la **loi sur la sécurité en ligne**, qui vise surtout la modération de contenu. La loi impose une obligation légale de diligence à deux catégories de services en ligne pour limiter les contenus préjudiciables : les services qui partagent du contenu généré par l'utilisateur (Facebook par exemple) et les services de recherche (comme Google).

En mars 2023, un *AI White Paper* réitère et précise l'option d'une régulation souple et pro-innovation de l'IA.

À la fin de l'année 2023, dans le cadre du nouveau **programme britannique pour une IA responsable** (*Responsible AI Programme* appelé *RAI UK*) a été lancé un appel à projets pour financer des projets de recherche autour de l'IA responsable<sup>1</sup>. *RAI UK* cherche aussi à diffuser à l'échelle mondiale les bonnes pratiques applicables aux systèmes d'IA.

En novembre 2023, lors du sommet de Bletchley Park, a été lancé le **UK AI Safety Institute, dont des responsables ont été rencontrés par vos rapporteurs**. Ian Hogarth en est le directeur général et une centaine de chercheurs y sont rémunérés pour effectuer des tests sur les modèles des principales entreprises d'IA, sur la base de leur participation volontaire.

Le *UK AI Safety Institute* poursuit trois objectifs :

- mener des recherches sur la sécurité des systèmes d'IA ;
- effectuer des tests de sécurité sur les modèles d'IA, avant et après leur mise sur le marché ;
- promouvoir le développement d'un écosystème d'entreprises tournées vers des IA responsables et sécurisées.

Sur la base du volontariat, les entreprises participent à ces tests. Un bac à sable réglementaire a également été mis en place.

Plus récemment, en février 2024, 80 millions de livres ont été alloués au lancement de **neuf centres de recherche** sur l'IA à travers le pays (recherche mathématique et informatique et IA appliquée à la science, à l'ingénierie et aux données), 9 millions de livres pour un **partenariat avec les États-Unis** sur la recherche en IA responsable et 2 millions de livres pour le *Arts and Humanities Research Council* (AHRC) afin de **financer la recherche en sciences sociales sur l'IA**.

Par ailleurs, avec les *AI Safety Summits*, dont la première édition s'est tenue à Bletchley Park en novembre 2023, le Royaume-Uni cherche à se **placer au centre de la régulation internationale**. Rishi Sunak a ainsi déclaré vouloir faire de son pays « *le foyer de la réglementation mondiale en matière de sécurité de l'IA* ». Ses objectifs étaient à la fois d'installer les institutions britanniques **au cœur des dispositifs de régulation internationale** de l'IA ; d'**attirer les regards** de l'ensemble des investisseurs internationaux vers l'écosystème britannique de l'IA ; et de **consacrer pour la première fois un sommet international à la question des risques extrêmes posés par les systèmes d'IA de pointe**, qu'il s'agisse de mauvais usages de la technologie par des acteurs malveillants (attaques et biosécurité notamment), de futures avancées technologiques vers l'IA générale ou de perte de contrôle humain sur les machines.

---

<sup>1</sup> Le site de *RAI UK* : <https://rai.ac.uk/> et la page dédiée à ses programmes de recherche : <https://rai.ac.uk/research/>



Plusieurs institutions internationales ainsi que 30 pays, dont les États-Unis et la Chine, y étaient représentés. Les participants au sommet avaient pour principales préoccupations : la nécessité d'agir en vue d'une compréhension commune de l'IA de pointe (la rareté des travaux scientifiques évaluant les risques existentiels est soulignée), le fait de traiter des risques actuels et des risques extrêmes futurs, de progresser vers la normalisation et l'interopérabilité en IA, d'associer les entreprises à la mise en place de tests de sécurité.

Ce premier sommet a débouché sur une **déclaration commune**, la *Bletchley Declaration*, qui assure la description des risques potentiels de l'IA et propose de développer un réseau de chercheurs à ce sujet. Il a aussi conduit à un **rapport commun** sur l'état de la recherche sur les risques extrêmes posés par les modèles de fondation, dit *State of Science Report* et a appelé à la création de *AI Safety Institutes*, qui auront pour mission de renforcer la capacité du secteur public à mener des recherches sur la sécurité de l'IA et à effectuer des tests de sécurité avant et après la mise sur le marché des modèles.

La **pérennisation de cette première initiative** a également été actée, ce qui s'est traduit par l'organisation en 2024 d'un sommet sur la sécurité de l'IA en Corée du Sud en 2024 et la préparation d'un autre sommet en France au début de l'année 2025, sur lequel nous reviendrons.

Le *soft power* britannique passe par le fait de **diffuser son modèle de régulation de l'IA à l'international**, pour peser dans les instances multilatérales et sur l'édiction de standards internationaux.

En conclusion, l'approche britannique souple et pro-innovation doit contribuer à faire du Royaume-Uni « *la place la plus attractive pour l'IA dans le monde* ». Le gouvernement souhaite **tirer profit du Brexit et de sa liberté réglementaire** pour proposer un cadre plus souple que l'*AI Act* européen et **attirer des capitaux étrangers**.

Cette approche a **des avantages** (un parti pris pro-innovation, une souplesse et une adaptabilité aux évolutions des technologies, une connaissance fine de leurs secteurs par les régulateurs, etc.) mais aussi des **limites** (difficulté pour les régulateurs à exercer la charge qui leur sera confiée tant matériellement que techniquement, absence de cadre harmonisé entre les régulateurs et potentiels conflits d'interprétation, ce qui pourrait conduire à une sécurité et une prévisibilité moindre pour les entreprises).

De plus, pour rester une superpuissance de l'IA, le Royaume-Uni a encore des **défis à relever**, comme :

- **une offre souveraine de GPU**, avec l'annonce d'un investissement d'un milliard de livres dans le secteur des semi-conducteurs ;

- **l'accès à la puissance de calcul**, en 2022, le Royaume-Uni représentait 1,3 % des parts mondiales de capacités de calculs contre 2,46 % pour la France ;

- **l'accès aux données**, bien que les organismes publics, y compris le NHS, possèdent des données précieuses, celles-ci sont souvent mal organisées et difficiles d'accès ;

- **un niveau d'investissement dans l'IA à accroître**, étant actuellement largement inférieur à ceux des États-Unis et de la Chine.

Comme dans d'autres contextes nationaux, la législation britannique sur la **propriété intellectuelle** applicable à l'IA générative **nécessitera des éclaircissements**, qu'il s'agisse du régime des données utilisées pour entraîner les modèles, des droits de propriété du contenu produit à l'aide de l'IA générative ou, encore, des risques de violation d'un droit d'auteur.

*g) Israël*

L'industrie israélienne de l'IA générative progresse rapidement (un rapport place même l'écosystème de capital-risque de l'IA générative du pays au troisième rang mondial) en accord avec une **stratégie nationale pour l'IA ambitieuse** et un volet stratégique au sein de la politique de défense.

La volonté de **protéger son secteur technologique** (20 % du PIB israélien environ) a conduit Israël à **refuser un encadrement strict de l'IA**. Israël a donc choisi d'adopter une approche de régulation douce et de gouvernance de l'IA, basée sur les risques et spécifique à chaque secteur.

Il n'existe donc **pas de lois ou de réglementations spécifiques** qui régissent directement l'IA dans ce pays. Israël ne dispose pas non plus d'autorité de régulation de l'IA. Le ministère israélien de l'innovation, de la science et de la technologie (MIST) sert d'agence exécutive de la stratégie nationale d'IA et collabore étroitement avec le ministère israélien de la justice (MOJ). En 2022, le MIST et le MOJ ont publié un projet de document d'orientation sur l'IA. Après avoir mené des consultations publiques, les deux ministères ont publié un document d'orientation en décembre 2023 intitulé « Innovation responsable : la politique d'Israël en matière de réglementation et d'éthique de l'intelligence artificielle ». Ce document décrit l'approche du pays en matière de gouvernance et de politique de l'IA. L'innovation responsable est un terme qui reflète la volonté du pays de protéger et de favoriser son industrie technologique en pleine croissance tout en restant attaché à des principes non contraignants.

Le document d'orientation recommande aux régulateurs de **formuler leurs politiques sur la base des principes de l'OCDE** pour garantir la fiabilité de la technologie de l'IA. Israël a en effet officiellement approuvé les principes de l'OCDE en matière d'IA, et a fait de ces lignes directrices internationales une base pour les acteurs israéliens de l'IA. Cette approche vise à renforcer la croissance, le développement durable, l'innovation, le bien-être social et la responsabilité.

En outre, elle souligne l'importance du respect des droits fondamentaux et des intérêts publics, de la garantie de l'égalité, de la prévention des préjugés et du maintien de la transparence, de la clarté, de la fiabilité, de la résilience, de la sécurité et de la sûreté. En particulier, le document d'orientation préconise l'adoption d'une approche fondée sur les risques au moyen d'évaluations des risques menées par les régulateurs sectoriels concernés, conformément aux principes de l'OCDE

Le document recommande également qu'Israël établisse des **directives nationales pour atténuer les abus potentiels du secteur privé**, tels que la discrimination, le manque de surveillance humaine, l'explicabilité insuffisante, la divulgation inadéquate, les problèmes de sécurité, les lacunes en matière de responsabilité et les violations de la vie privée.

Pour répondre à ces préoccupations, le document préconise d'**éviter une législation horizontale de grande envergure** et d'opérer plutôt dans le cadre de réglementations sectorielles spécifiques. Il appelle aussi à la création d'un **centre de coordination des politiques d'IA sous l'égide du ministère de la justice**, qui fonctionnerait comme un organe interministériel pour faciliter la coordination entre les différents départements et les agences concernées. Ce centre serait également chargé de conseiller les régulateurs, de faciliter le dialogue et le partage des connaissances avec le monde universitaire et l'industrie, et d'aider les régulateurs à identifier les applications et les défis de l'IA au sein des secteurs réglementés.

En conclusion, bien qu'aucune obligation spécifique ne soit actuellement imposée en Israël aux développeurs, aux déployeurs ou aux utilisateurs de systèmes d'IA, ces acteurs doivent se préparer à la **mise en œuvre d'éventuelles normes** qui seraient en harmonie avec les principes de l'OCDE.

#### *h) L'Arabie Saoudite*

Le Royaume d'Arabie saoudite n'a **pas encore adopté de cadre juridique pour la gouvernance de l'IA**, préférant se concentrer sur la **croissance et l'investissement**. Il a cependant mis à jour ses lois sur le droit d'auteur et les données personnelles pour relever les défis posés par l'IA.

Le pays a établi une **autorité spécialisée** et a adopté des **principes éthiques** pour guider le développement de l'IA.

L'Autorité saoudienne de l'intelligence artificielle (SDAIA) est ainsi chargée de superviser les ambitions du pays en matière d'IA et de préparer de nouveaux cadres législatifs et réglementaires pour l'IA.

Les Principes d'éthique de l'IA de l'Arabie saoudite tentent de régir son secteur croissant de l'IA. Cependant, la **portée juridique** de ces lignes directrices n'est **pas claire**.

*i) Les Émirats arabes unis (EAU)*

Parmi tous les pays arabes, les Émirats arabes unis (EAU) jouent un **rôle de premier plan dans le développement des technologies d'intelligence artificielle**. Dès octobre 2017, le pays a clairement exprimé son ambition de construire son écosystème d'IA dans le cadre de sa stratégie nationale pour l'intelligence artificielle. L'objectif de cette stratégie est de mettre l'accent sur les moyens d'améliorer la compétitivité des EAU en matière d'IA dans la région et dans le monde. Une priorité de second rang consiste à garantir une gouvernance et une réglementation efficaces. Un **Conseil de l'intelligence artificielle et de la blockchain** a notamment été mis en place.

Avec un financement du conseil de recherche sur les technologies avancées du gouvernement d'Abou Dhabi, le pays a produit **le LLM en open source Falcon 180B**, du nom de l'oiseau national des Émirats. Lors du lancement de Falcon par l'Institut d'innovation technologique des EAU en septembre 2023, Hugging Face a salué son arrivée comme « *le plus grand LLM disponible en libre accès, avec 180 milliards de paramètres* ». La création de Falcon a marqué une étape importante dans la stratégie nationale des Émirats pour intégrer le club des leaders mondiaux de l'IA.

Privilégiant l'innovation et la compétitivité, les Émirats n'ont adopté **ni loi ni réglementation spécifique** de l'IA. Ils privilégient des « **bacs à sable réglementaires** » plutôt qu'un cadre juridique contraignant : cette stratégie **favorise le développement technologique** avec des **tests en direct** des technologies **dans un environnement contrôlé** sous la **supervision** directe d'un régulateur. Il n'est pas surprenant que cette approche favorable aux développeurs reçoive le soutien marqué d'entreprises d'IA comme OpenAI.

## II. UNE DIZAINNE DE PROJETS DE GOUVERNANCE INTERNATIONALE NON COORDONNÉS

La gouvernance globale de l'IA est **insuffisante** en dépit des annonces répétées de nombreuses organisations internationales en la matière. Non seulement les **propositions se multiplient de manière stérile car non coordonnée** mais l'on assiste au **creusement de la fracture numérique mondiale** sous l'effet de l'IA, très inégalement distribuée sur notre planète, tant du point de vue de sa production que de son utilisation.

Outre leurs propres investigations, audits, déplacements et questionnaires, vos rapporteurs ont pu s'appuyer pour cette partie sur les travaux comparatifs de Florence G'sell, rencontrée à l'Université de Stanford<sup>1</sup>.

---

<sup>1</sup> Cf. son rapport précité d'octobre 2024.

Ses recherches montrent que plusieurs catégories d'actions internationales peuvent être distinguées :

- le fait de **rédiger des traités internationaux ou des lignes directrices mondiales**, comme le traité contraignant du Conseil de l'Europe, l'accord non contraignant de l'ONU appelé « Pacte numérique mondial » ou les recommandations de l'Unesco ;

- le **soutien aux politiques nationales et internationales** par des recommandations et des travaux d'experts, comme l'illustrent les contributions de l'OCDE et du Partenariat mondial sur l'intelligence artificielle ;

- les **discussions internationales dans des forums diplomatiques restreints** comme le G7, le G20, les BRICS ou le Conseil du commerce et des technologies UE-États-Unis (TTC) ;

- la **coordination de l'action des États membres d'organisations supranationales** en matière de régulation de l'IA comme l'illustrent les initiatives de l'UE ou de l'Union africaine ;

- ou encore le fait de **convoquer des représentants des États et du monde économique autour d'une question spécifique**, à l'instar des sommets sur la sécurité de l'IA.

#### ***A. LA RÉFLEXION TECHNIQUE LA PLUS ABOUTIE : LES PRINCIPES, RECOMMANDATIONS ET MÉTRIQUES DE L'OCDE***

Dès 2016, l'Organisation de coopération et de développement économiques (OCDE) a réuni un forum de prospective sur l'IA puis a organisé en 2017 une grande conférence sur l'IA (« L'IA. Machines intelligentes, politiques intelligentes »). En 2018 et 2019, elle s'est appuyée sur un groupe d'experts sur l'IA pour élaborer, en mai 2019, des principes de l'IA et des recommandations de l'OCDE pour les politiques publiques.

### **1. Les principes, les recommandations et la classification des systèmes**

Adoptés en 2019 et amendés en 2024, les principes et les recommandations sur l'intelligence artificielle de l'OCDE<sup>1</sup> fournissent des **lignes directrices non contraignantes aux États membres et à toutes les parties prenantes** dans le cadre d'une approche responsable pour une IA digne de confiance. Certaines des recommandations concernent spécifiquement la gouvernance de l'IA et les politiques publiques de l'IA à mettre en place au niveau national. Une mise à jour des principes a été effectuée en 2024.

---

<sup>1</sup> *Recommandations de l'OCDE sur l'IA de 2019* : <https://oecd.ai/en/assets/files/OECD-LEGAL-0449-en.pdf> ; *Principes sur l'IA* : <https://www.oecd.org/en/topics/ai-principles.html> ; *Mise à jour des principes en 2024* : <https://www.oecd.org/newsroom/oecd-updates-ai-principles-to-stay-abreast-of-rapid-technological-developments.htm>

Prendre des décisions d'encadrement de l'IA nécessite de savoir de quoi l'on parle : il faut pouvoir décrire ce que l'on souhaite exactement réguler. Ainsi, l'OCDE propose un cadre qui permet de **classer les systèmes d'IA selon différentes dimensions** aux stades de leur conception et de leur utilisation<sup>1</sup>. **Tous les pouvoirs publics devraient prendre en compte cet outil de classification ou une méthode similaire (tournée davantage vers la chaîne de valeur de l'IA par exemple) pour mettre en place des politiques publiques de régulation de l'IA.**

L'OCDE fait valoir que tout modèle d'IA se décompose en **quatre phases distinctes** : le contexte (1), les données d'entrées (2), les modèles d'IA (3), les tâches demandées ainsi que les sorties de l'IA (4). Bien que ces phases soient distinctes, elles s'influencent mutuellement et ne peuvent pas être traitées de manière totalement indépendante.

Le **contexte** désigne l'environnement socio-économique dans lequel l'IA est déployée. Il désigne aussi bien l'utilisateur potentiel du modèle, que les parties prenantes à son développement, ou le secteur économique qu'il sert. De ce contexte dépendent **les données et les valeurs d'entrées** du modèle d'IA pour lesquelles il est important de connaître la nature mais également la façon dont elles sont collectées, ou dont elles sont amenées à évoluer, s'il s'agit de données dynamiques. De ces données récoltées dépend **le modèle d'IA**, qui est lui-même, un critère de classification. Le modèle peut être de type symbolique, donc basé sur une série de règles logiques, ou connexionniste et donc appuyé sur des calculs de type statistique. S'il est connexionniste, le modèle peut être soumis à un apprentissage supervisé ou semi-supervisé par exemple. Enfin, selon le type de modèle, les sorties ne sont pas les mêmes et il est évidemment important **d'évaluer les modèles à la lumière de leur production et des tâches qui leur sont demandées.**

Par exemple, une **IA appliquée au secteur des banques et assurances** peut être considérée comme particulièrement sensible car elle prend place dans un contexte (1) pouvant influencer de façon importante la vie des personnes qu'elle concerne. Elle prendrait en entrée des données personnelles, relevés de compte ou données de santé par exemple (2). Selon que le modèle utilisé est symbolique ou connexionniste (3), on n'a pas la même possibilité d'expliquer les choix réalisés en sortie par le système d'IA (4), ce qui peut représenter une injustice pour le client concerné.

L'OCDE pose **cinq principes généralistes** que doivent respecter les systèmes d'IA tout au long de leur cycle de vie et qui concernent l'ensemble des acteurs concernés par l'IA. Elle formule aussi **cinq recommandations spécifiques à destination des décideurs politiques.**

---

<sup>1</sup> Cf. cette classification si utile sur le site de l'OCDE : <https://oecd.ai/en/classification> une vidéo a également été réalisée : <https://youtu.be/-S5dCR9z5rI>

Les principes généralistes à destination de tous les acteurs de l'IA sont conçus à partir des valeurs morales nécessaires au développement d'une IA saine. Ces principes sont : « **Croissance inclusive, développement durable et bien-être** » (1.1), « **Droits de l'homme, vie privée, équité** » (1.2), « **Transparence, explicabilité** » (1.3), « **Robustesse, sécurité, sûreté** » (1.4) et « **Responsabilité** » (1.5). Ces principes font qu'une intelligence artificielle doit être dédiée à la prospérité et être bénéfique à l'humanité et à la planète (1.1), dans le respect des droits de l'homme, de la dignité humaine, de la vie privée et sans induire de discriminations (1.2). Ses sorties doivent être explicables au mieux et de la façon la plus transparente possible (1.3) tout en protégeant les données des utilisateurs du modèle ou des personnes dont les données ont servi d'entraînement au modèle, la sécurité et la sûreté du système doivent être garanties (1.4). Enfin, les individus et organisations qui déploient des systèmes d'IA doivent toujours être responsables du fonctionnement, donc des résultats de ces derniers et être capables d'en rendre compte (1.5).

Les **cinq recommandations pour les décideurs politiques** doivent permettre à ceux-ci d'anticiper les transformations liées au développement de l'IA. L'organisation recommande ainsi : un **investissement dans la recherche et le développement** de l'IA (2.1), l'encouragement d'un **écosystème numérique** pour l'IA (2.2), la fourniture d'un environnement de **politiques publiques propices à l'IA** (2.3), la construction et la **préparation de capacités humaines pour la transformation du marché du travail** (2.4) et la **coopération internationale pour une IA de confiance** (2.5).

Plus concrètement, cela signifie que les gouvernements devraient envisager des **investissements** de long terme dans la recherche et le développement d'IA mais également encourager la recherche publique comme la recherche privée, et ce, dans des domaines techniques mais également juridiques ou sociétaux, y compris à travers des investissements dans des systèmes ouverts dits *open source* (2.1). Les gouvernants devraient également encourager la création d'un **écosystème permettant de faciliter le partage de technologies ou d'infrastructures** entre les acteurs des systèmes d'IA de confiance (2.2). Un environnement de **politiques publiques propices** devrait également être construit, permettant de passer de la recherche et développement aux produits en passant, par exemple, par des mécanismes d'expérimentation de type bac à sable (2.3). Il faut également préparer la **transition du marché du travail** impliquée par l'IA en travaillant avec les parties prenantes à ce changement et en promouvant le dialogue social durant la mise en place de l'IA au travail (2.4). Enfin, plutôt qu'un travail strictement national, la gouvernance de l'IA devrait être le fruit de **coopérations internationales** permettant le partage d'informations ou le développement de normes internationales (2.5).



Le système de classification proposé par l'OCDE prend en compte **l'ensemble du cycle de vie du modèle d'IA** et permet de définir précisément ce qu'est un système d'IA et d'en donner des limites claires. Couplé à des **principes éthiques** également définis, il offre des ressources très utiles pour encadrer l'IA, surtout que l'OCDE y ajoute des recommandations pour le contenu des politiques publiques à mettre en place. L'organisation a aussi proposé de définir des **métriques d'analyse** et des **outils** pour corriger les problèmes pouvant subvenir. L'objectif est de maximiser, grâce à ces outils, le respect des principes éthiques par l'ensemble des parties prenantes, dont les gouvernants, tout au long du cycle de conception et d'utilisation de l'IA.

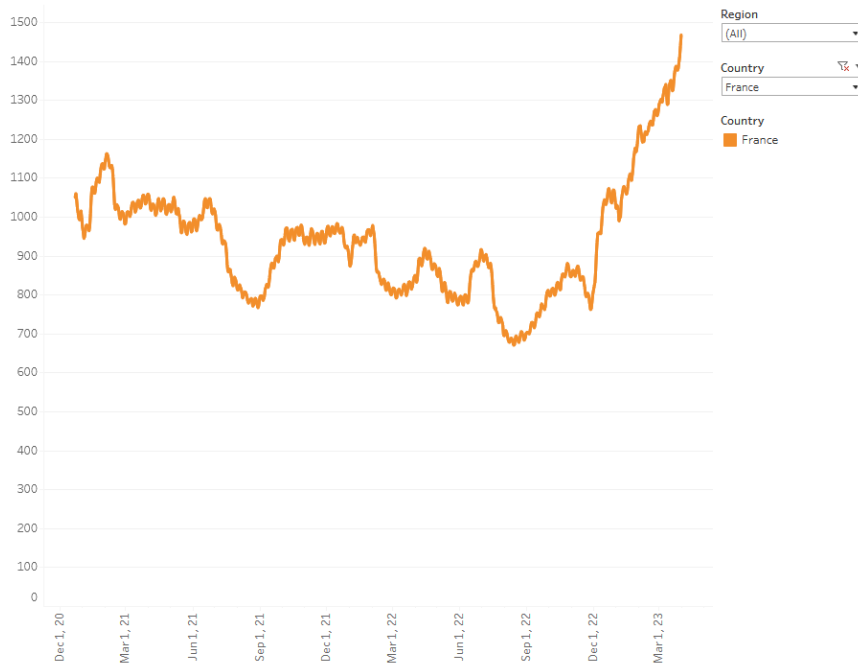
## 2. La méthodologie des métriques

L'OCDE a commencé par recenser un **ensemble de métriques d'analyse et d'outils provenant d'organismes publics ou privés**, permettant d'évaluer et de corriger d'éventuels défauts présents dans l'une des quatre phases des systèmes d'IA mentionnées ci-dessus. L'organisation a répertorié des méthodologies pour **104 métriques** et **125 outils techniques de correction des systèmes d'IA**.

L'utilisation d'une métrique d'évaluation dépend du type de modèle utilisé, le *Word Error Rate* ne peut s'appliquer qu'aux modèles de reconnaissance de discours par exemple. Elle dépend également de l'objectif évalué lors du passage du test. L'OCDE en répertorie huit, fondés sur les principes éthiques définis dans son cadre d'analyse. Certaines métriques relèvent de plusieurs principes, mais certains principes n'ont aucune métrique : responsabilité (aucune métrique disponible), équité (20 métriques), bien-être humain (aucune métrique), performance (78 métriques), gouvernance des données et de la vie privée (1 métrique), robustesse et sécurité numérique (7 métriques), sûreté (1 métrique) et transparence et explicabilité (11 métriques). On peut remarquer une **répartition inégale des métriques selon les catégories**, la performance des modèles possédant par exemple bien plus de métriques que les autres catégories.

En plus de ces métriques, le site de l'OCDE propose des **outils d'écoute sociale** relatifs à l'IA, par exemple celui concernant les tweets sur X (ex-Twitter) mentionnant l'intelligence artificielle.

### Moyenne mensuelle du nombre de tweets sur X (ex-Twitter) mentionnant l'intelligence artificielle en France



**Légende :**

- en abscisse, la date

- en ordonnée, la moyenne mobile sur 30 jours du nombre de tweets mentionnant l'IA

**Source :** OECD.AI

On trouve également une mesure du **nombre d'incidents** concernant l'IA (AIM) qui est un outil d'écoute sociale puisqu' il s'appuie sur le traitement automatique d'articles de presse dans le monde, et permet de voir quels sont les moments d'engouement médiatique autour de la notion d'intelligence artificielle. La mesure des incidents peut se décliner en fonction de la **nature de l'incident** (c'est-à-dire en fonction du principe éthique que l'IA n'a pas respecté), mais également de **sa sévérité** (menace non physique, danger, blessure ou mort), du **type de dommage** causé (physique, psychologique, etc.) et de la **partie prenante** victime de l'incident (gouvernement, consommateur, travailleur, minorité, etc.).

Le cadre de travail de l'IA de l'OCDE permet donc de **définir l'IA au-delà de sa simple définition technique** : la mise en place de politiques publiques selon ses recommandations exige par exemple de s'intéresser aux relations entre l'IA et son contexte ; le contexte influence les données et les résultats influencent l'environnement). Cette définition sociotechnique concerne un grand nombre de parties prenantes lors de la conception et l'utilisation de modèles d'IA.

Toutes les parties prenantes sont concernées par des enjeux éthiques majeurs liés à l'IA : les systèmes d'IA doivent être **bénéfiques pour l'homme et son environnement**, ils doivent respecter la **dignité humaine** et faire en sorte qu'**aucun biais** ne puisse nuire aux utilisateurs. Les données d'entraînement ainsi que les données des utilisateurs doivent être protégées de façon robuste et les organisations doivent toujours être responsables des produits d'un système d'IA. Pour cela, les décideurs publics doivent mettre en place des initiatives qui permettent d'encourager la recherche et développement, ainsi que de créer un écosystème sain. Les décisions doivent être prises en accord avec les multiples parties prenantes lors de la création des systèmes d'IA comme pendant leur phase de fonctionnement, qu'il s'agisse du secteur public comme du secteur privé. La **gouvernance mondiale de l'IA** doit être mise en place pour permettre la création de normes internationales ainsi qu'un partage plus vaste des bonnes pratiques.

Des métriques et outils sont déjà à disposition pour évaluer et corriger les défauts des systèmes d'IA, toutefois comme il a été vu, **tous les principes ne disposent pas de métriques permettant leur évaluation** : la performance est clairement le critère le plus évalué alors que les aspects éthiques ou la responsabilité sont difficilement voire pas du tout mesurés. En conclusion, à côté de l'évaluation des modèles eux-mêmes, **l'écoute sociale peut faire partie des outils techniques** qui permettent d'observer l'environnement de l'IA, notamment le rapport de la société aux technologies.

### 3. L'Observatoire des politiques publiques de l'IA

L'OCDE a créé un Observatoire des politiques publiques de l'IA<sup>1</sup> pour aider les décideurs politiques et les experts en IA à travers un centre de ressources complet sur les politiques et les normes applicables à l'IA, tout en promouvant les lignes directrices de l'OCDE puisqu'il surveille aussi la manière dont les pays respectent et mettent en œuvre les principes et les recommandations politiques de l'organisation.

L'évaluation biennale conduite par l'Observatoire, intitulée « L'état de mise en œuvre des principes de l'OCDE en matière d'IA », indique qu'en 2024 **plus de 50 pays avaient mis en œuvre des stratégies nationales en matière d'IA**, dont beaucoup font directement référence aux principes de l'OCDE<sup>2</sup>. Sur les 46 adhérents aux principes de l'OCDE en matière d'IA, 41 avaient mis en place une stratégie nationale et trois étaient en train d'en élaborer une. Le rapport indique qu'en mai 2024, plus de 1 020 programmes d'action avaient été initiés dans 70 pays membres et non membres, ce qui témoigne de l'attention accrue portée à la gouvernance de l'IA depuis 2019.

---

<sup>1</sup> Cf. cette présentation de l'Observatoire des politiques publiques de l'IA : <https://oecd.ai/en/about/background>

<sup>2</sup> OCDE, 2024, « Report on the Implementation of the OECD Recommendation on Artificial Intelligence » : [https://one.oecd.org/document/C/MIN\(2024\)17/en/pdf](https://one.oecd.org/document/C/MIN(2024)17/en/pdf)

Le site de l'Observatoire des politiques de l'IA de l'OCDE propose également un référentiel en temps réel qui suit les paysages réglementaires de l'IA de 69 pays différents<sup>1</sup>. Il fournit également des outils pour l'audit des systèmes d'IA<sup>2</sup> et un *Global AI Incident Monitor* (AIM)<sup>3</sup>. Pour soutenir davantage l'AIM et les travaux de l'OCDE visant à recueillir des rapports sur les incidents liés à l'IA, l'Observatoire des politiques de l'IA a publié un rapport sur la définition des incidents liés à l'IA et des termes associés qui propose des distinctions importantes entre les incidents liés à l'IA et les dangers de l'IA<sup>4</sup>.

## B. LE CADRE MULTILATÉRAL EN CONSTRUCTION

### 1. La contribution multiforme mais encore inachevée de l'Organisation des Nations unies (ONU)

L'Organisation des Nations unies (ONU) souhaite **s'associer avec l'OCDE** pour avancer sur le chemin d'une régulation mondiale de l'IA<sup>5</sup>, même si elle a semblé adopter une **position prudente quant à la gouvernance globale de l'IA** jusqu'ici, dans le cadre de son mandat en faveur de la paix et de la sécurité dans le monde. Il s'agit surtout pour elle de trouver des **consensus au sein de la communauté internationale** et de voter des résolutions qui posent un **cadre non contraignant** en vue du respect de ces consensus, sans aller jusqu'à une régulation réelle de l'IA par le droit international. Lors du déplacement de vos rapporteurs aux États-Unis, la représentante du bureau de liaison ONU-Unesco a confirmé que les institutions onusiennes n'en sont pas encore au stade de la construction effective d'une organisation internationale pour la gouvernance de l'IA. Une avancée importante vient cependant d'être faite très récemment avec le **Pacte numérique mondial de septembre 2024**.

Avant cela, l'IA n'avait pas été écartée du champ de travail de l'ONU puisque l'on dénombrait environ 300 projets en lien avec l'IA dans l'ensemble de l'ONU et de ses institutions.

La première de ces contributions est celle de l'**Organe consultatif de haut niveau sur l'intelligence artificielle** (*High-Level Advisory Body on AI*) mis en place par le Secrétaire général des Nations unies en octobre 2023.

---

<sup>1</sup> Cf. OCDE, « National AI Policies & Strategies » : <https://oecd.ai/en/dashboards/overview>

<sup>2</sup> Cf. OCDE, « Catalogue of Tools and Metrics for Trustworthy AI » : <https://oecd.ai/en/catalogue/tools?terms=audit&approachIds=1&objectiveIds=2&orderBy=dateDesc>

<sup>3</sup> Cf. OCDE, « AI Incidents Monitor » : <https://oecd.ai/en/incidents>

<sup>4</sup> Cf. OCDE, « Defining AI incidents and related terms » : <https://doi.org/10.1787/d1a8d965-en>

<sup>5</sup> L'ONU semble avoir ouvert un dialogue constructif avec l'OCDE ainsi qu'avec le G20 mais elle reste étrangère aux travaux conduits dans les cadres du G7 et du PMIA/GPAI qui seront évoqués plus loin.

Le rapport provisoire de ce groupe multipartite a été publié en décembre 2023 sous le titre « *Interim Report : Governing AI for Humanity* », suivi d'un rapport définitif en septembre 2024 : « *Governing AI for Humanity* »<sup>1</sup>. Ce dernier a été rendu public à l'occasion du Sommet sur le Futur qui s'est tenu les 22 et 23 septembre 2024 au siège de l'ONU à New York en présence de plus de 130 chefs d'État et de gouvernement et qui a surtout été l'occasion d'adopter un **Pacte numérique mondial** et d'annoncer le **lancement d'un travail en commun entre l'ONU et l'OCDE afin de renforcer la gouvernance mondiale de l'IA**.

Un autre aspect est celui de la **place accordée aux droits de l'homme**. En novembre 2023, le Haut-Commissaire des Nations unies aux droits de l'homme, Volker Türk, appelait ainsi à porter la plus grande attention aux risques liés à l'IA, en insistant sur son impact en matière de droits humains. Dans cette déclaration, il proposait des **évaluations des modèles d'IA** dans les domaines dans lesquels les technologies pourraient avoir des effets significatifs, affirmant que tout projet de gouvernance mondiale de l'IA devrait **prendre en compte les droits de l'homme**<sup>2</sup>.

L'Assemblée générale des Nations unies a apporté sa contribution à la discussion internationale avec le **vote de plusieurs résolutions** appelant à la coopération pour garantir la sécurité des systèmes d'IA.

Le 21 mars 2024, elle a adopté une première résolution, soutenue par plus de 120 États membres, dont la France et surtout les États-Unis, à la tête de l'initiative. La résolution vise à « *saisir les possibilités offertes par des systèmes d'intelligence artificielle sûrs, sécurisés et dignes de confiance pour le développement durable* »<sup>3</sup>. Non contraignante, elle préconise une approche **de régulation axée sur la sécurité, le respect des droits de la personne et des libertés fondamentales et l'inclusivité**. Elle souligne l'importance de diffuser partout dans le monde les capacités de l'IA, en particulier dans les **pays en développement**. Elle affirme également que les droits de l'homme doivent être respectés tout au long de la chaîne de valeur de l'IA, en ligne et hors ligne, demandant à tous les États membres de s'abstenir ou de cesser d'utiliser des systèmes d'IA contraires au droit international des droits de l'homme, ou qui présentent un risque pour ces derniers. L'Assemblée générale considère que les systèmes d'IA sont des outils qui peuvent être mobilisés par les États pour **répondre aux objectifs de développement durable 2030**.

---

<sup>1</sup> Cf. les liens vers ces deux rapports des Nations unies sur l'IA et sa gouvernance : [https://www.un.org/sites/un2.un.org/files/ai\\_advisory\\_body\\_interim\\_report.pdf](https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf) ainsi que [https://www.un.org/sites/un2.un.org/files/governing\\_ai\\_for\\_humanity\\_final\\_report\\_en.pdf](https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf)

<sup>2</sup> Cf. l'article, « Volker Türk appelle à une gouvernance attentive aux risques liés à l'intelligence artificielle et mettant l'accent sur les droits humains », 2023, OHCHR : <https://www.ohchr.org/fr/statements-and-speeches/2023/11/turk-calls-attentive-governance-artificial-intelligence-risks>

<sup>3</sup> Résolution 78/49 du 21 mars 2024, « Saisir les possibilités offertes par des systèmes d'intelligence artificielle sûrs, sécurisés et dignes de confiance pour le développement durable » : <https://documents.un.org/doc/undoc/ltid/n24/065/93/pdf/n2406593.pdf>

Dans le même temps, elle souligne les risques potentiels des technologies d'IA dans les domaines de **la vie privée, des données personnelles** ou encore **du droit d'auteur**, ainsi que les risques de **biais** induits par l'IA. Aussi, elle invite à prendre des mesures d'évaluation et de gouvernance des modèles, en promouvant des modèles sûrs, sécurisés et dignes de confiance.

L'Assemblée générale des Nations unies a poursuivi sa réflexion avec le vote le 1<sup>er</sup> juillet 2024 d'une autre résolution visant à « *intensifier la coopération internationale en matière de renforcement des capacités dans le domaine de l'intelligence artificielle* »<sup>1</sup>. Initiée par la Chine, cette résolution poursuit un objectif spécifique de **solidarité internationale** afin de permettre aux États membres de combler leurs lacunes éventuelles en termes de développement de l'IA. Elle vise à favoriser une coopération étroite en matière d'IA en encourageant davantage le partage de connaissances, les transferts de technologie, la formation et des recherches collaboratives au sein de la communauté internationale. La résolution demande aux États membres de mettre en place autant que possible des plans de **renforcement de leurs capacités souveraines dans le cadre de leurs stratégies nationales relatives à l'IA**.

Sur un autre aspect, en novembre 2023, la commission sur le désarmement et la sécurité internationale de l'Assemblée générale des Nations unies a adopté une **résolution contre les armes létales autonomes**<sup>2</sup>.

Le **Pacte numérique mondial**, adopté en septembre 2024 par les participants au Sommet sur le Futur<sup>3</sup>, vingt ans après le Sommet mondial des Nations unies sur la société de l'information, trace une feuille de route pour une coopération numérique à l'échelle mondiale poursuivant à la fois l'objectif d'exploiter l'immense potentiel des technologies numériques et celui de combler les fractures numériques existantes.

Non contraignant, ce Pacte numérique mondial propose un premier cadre global préparatoire à une gouvernance mondiale des technologies numériques et de l'intelligence artificielle. Il en définit les objectifs, les principes, les engagements et les actions permettant de développer un avenir numérique ouvert, libre et sûr pour tous, en soulignant les avantages que les technologies numériques apportent à l'humanité.

---

<sup>1</sup> Résolution 78/311 du 1<sup>er</sup> juillet 2024, « Intensifier la coopération internationale en matière de renforcement des capacités dans le domaine de l'intelligence artificielle » : <https://documents.un.org/doc/undoc/gen/n24/197/27/pdf/n2419727.pdf>

<sup>2</sup> Communiqué de la commission sur le désarmement et la sécurité internationale de l'Assemblée générale des Nations unies, le 1<sup>er</sup> novembre 2023, « First Committee Approves Resolution on Lethal Autonomous Weapons, as Speaker Warns "An Algorithm Must Not Be in Full Control of Decisions Involving Killing" » : <https://press.un.org/en/2023/gadis3731.doc.htm>

<sup>3</sup> Cf. la présentation du pacte sur cette page : <https://www.un.org/techenvoy/global-digital-compact>



Le Pacte, qui a bénéficié de l'aide de l'UE et pour lequel la Commission européenne s'est fortement investie<sup>1</sup>, se déclinera à travers :

- un **comité scientifique international indépendant sur l'IA** pour promouvoir la compréhension scientifique (la question du rôle futur de l'Organe consultatif de haut niveau sur l'intelligence artificielle au sein de l'ONU est donc posée) ;

- un **dialogue mondial sur la gouvernance de l'IA** impliquant les gouvernements des États membres et toutes les parties prenantes concernées, sera lancé en marge des conférences et réunions pertinentes des Nations unies ;

- l'engagement des États membres à prendre plusieurs mesures d'ici 2030, telles que le **développement de mécanismes de financement et d'incitations pour connecter les 2,6 milliards de personnes** ne bénéficiant pas d'Internet, l'établissement de **garanties pour prévenir et traiter tout impact négatif sur les droits de l'homme** découlant de l'utilisation des technologies numériques émergentes et la fourniture et la facilitation de l'accès à des **informations scientifiques indépendantes pour lutter contre la désinformation**.

## 2. Le travail spécifique de l'Organisation des Nations unies pour l'éducation, la science et la culture (Unesco)

Dans les domaines de la science, de la culture et de l'innovation, qui sont le cœur de ses compétences, l'Unesco, a initié plusieurs réflexions en matière d'intelligence artificielle<sup>2</sup>. Pour rester dans son champ de compétences, l'Unesco a principalement proposé la définition de règles éthiques, d'application large. Ainsi, Gabriela Ramos, sous-directrice générale pour les sciences sociales et humaines de l'Unesco affirme : « *Dans aucun autre domaine, la boussole éthique n'est plus pertinente que dans celui de l'intelligence artificielle* ».

Un travail préliminaire a été réalisé en 2019 par sa Commission mondiale d'éthique des connaissances scientifiques et des technologies (COMEST)<sup>3</sup>.

---

<sup>1</sup> Communiqué de la Commission européenne : <https://digital-strategy.ec.europa.eu/en/news/united-nations-members-adopted-global-digital-compact-shaping-safe-and-sustainable-digital-future>

<sup>2</sup> Unesco, 2019, « *International conference on Artificial intelligence and Education, Planning education in the AI Era : Lead the leap: Final report* » ; Pedró, F., Subosa, M., Rivas, A., & Valverde, P., 2019, « *Artificial intelligence in education: Challenges and opportunities for sustainable development* » ; Miao, F., & Holmes, W., 2020, « *International Forum on AI and the Futures of Education, developing competencies for the AI Era* », Synthesis report ; ou encore Miao, F., Wayne Holmes, Ronghuai, H., & Hui, Z., 2021, « *AI and education: Guidance for policy-makers* ».

<sup>3</sup> COMEST, 2019, « *Étude préliminaire sur l'Éthique de l'intelligence artificielle* » : [https://unesdoc.unesco.org/ark:/48223/pf0000367823\\_fre](https://unesdoc.unesco.org/ark:/48223/pf0000367823_fre)



Puis l'Unesco a publié le 23 novembre 2021 ses **recommandations éthiques en matière d'intelligence artificielle**, adoptées par 193 pays<sup>1</sup>. Dix principes éthiques, **non contraignants**, assez proches de ceux préconisés par l'ONU, ciblent les systèmes d'IA par rapport aux domaines centraux de l'Unesco : éducation, science, culture, communication et information.

Pour permettre la mise en œuvre concrète de ces recommandations éthiques, une **méthodologie** a été **mise à la disposition des États**<sup>2</sup> (appelée « méthode d'évaluation de l'état de préparation ») et une autre a été préparée spécifiquement **pour les entreprises**<sup>3</sup> (appelée « évaluation de l'IA éthique »). Un accord a été signé le 5 février 2024 lors du deuxième Forum mondial de l'Unesco sur l'IA par huit entreprises du secteur numérique et Audrey Azoulay, directrice générale de l'Unesco<sup>4</sup>.

Dans le cadre de la Semaine de l'apprentissage numérique qui s'est tenue à Paris en septembre 2023, l'Unesco a publié un *Guide pour l'IA générative dans l'éducation et la recherche*<sup>5</sup>, qui présente les technologies d'IA générative et les différents modèles actuellement disponibles. Il propose des **recommandations pour encadrer ces technologies à la lumière de principes éthiques en promouvant l'inclusion et l'équité**. Il met l'accent sur une approche centrée sur l'humain, en prônant surtout une **vigilance sur les usages de l'IA dans les contextes éducatifs, ces usages devant être éthiques, sûrs, justes et dotés de sens**. Il propose des mesures pour **intégrer de manière responsable l'IA dans les activités d'enseignement, d'apprentissage et de recherche**, notamment avec une explication pédagogique des technologies d'IA générative, une présentation de leurs enjeux éthiques et politiques et de leurs perspectives d'encadrement. Le rapport fournit des **exemples d'usages de l'IA générative apportant des bénéfices pour la pensée critique** ainsi que pour la **créativité** dans l'éducation et la recherche, tout en en atténuant les risques lors de la **conception de programmes, dans l'enseignement et les activités d'apprentissage**.

---

<sup>1</sup> Unesco, 23 novembre 2021, « Recommandations sur l'éthique de l'intelligence artificielle » : [https://unesdoc.unesco.org/ark:/48223/pf0000381137\\_fre](https://unesdoc.unesco.org/ark:/48223/pf0000381137_fre)

<sup>2</sup> Cf. la méthodologie pour les États proposée au lien suivant : <https://www.unesco.org/fr/articles/methode-devaluation-de-letat-de-preparation-un-outil-de-la-recommandation-sur-lethique-de>

<sup>3</sup> Cf. la méthodologie pour les entreprises proposée au lien suivant : <https://www.unesco.org/en/articles/ethical-impact-assessment-tool-recommendation-ethics-artificial-intelligence>

<sup>4</sup> Il s'agit de GSMA, INNIT, Lenovo Group, LG AI Research, Mastercard, Microsoft, Salesforce et Telefonica, cf. le lien suivant : <https://www.unesco.org/fr/articles/ethique-de-lia-8-geants-de-la-tech-sengagent-appliquer-la-recommandation-de-lunesco>

<sup>5</sup> Cf. le guide publié en septembre 2023 par l'Unesco sur l'IA dans l'éducation et la recherche : <https://www.unesco.org/fr/articles/orientations-pour-lintelligence-artificielle-generative-dans-leducation-et-la-recherche>

À la suite de ces recommandations, l'Unesco a créé en février 2024, en partenariat avec d'autres institutions comme l'Institut Alan Turing ou l'Union internationale des télécommunications (UIT), un **Observatoire international de l'éthique et de la gouvernance de l'intelligence artificielle**<sup>1</sup> (en anglais *Global AI Ethics and Governance Observatory*). Sur la base des recommandations de 2021, l'observatoire a pour objectif de fournir aux parties prenantes des ressources pour discuter des dilemmes éthiques liés à l'IA qui se présentent à eux, ainsi que de leurs implications sociétales.

### 3. Les Principes pour l'IA du G20

En relation avec l'ONU et l'Unesco, le G20 – qui représente 85 % de l'économie mondiale et 80 % de la population mondiale – a adopté, lors du sommet d'Osaka de 2019, des « **Principes pour l'IA** », qui reprennent les priorités fixées par l'OCDE<sup>2</sup>. Depuis, le G20 a mis l'intelligence artificielle à l'ordre du jour de chacune de ses rencontres.

Le sommet du G20 de Rio de Janeiro des 18 et 19 novembre 2024, est l'occasion de **poursuivre la discussion sur les principes de l'IA et sur les moyens par lesquels le G20 peut contribuer à la réflexion internationale sur l'IA**, en intégrant la dimension du **développement**, de la **fracture numérique mondiale** et de la **justice sociale**, à laquelle le Brésil est attaché<sup>3</sup>. Le groupe de travail sur l'économie numérique du G20 a d'ailleurs organisé en avril 2024 une conférence sur « l'intelligence artificielle pour le développement durable et la réduction des inégalités ».

---

<sup>1</sup>Cf. l'article du 6 février 2024, « Unesco launches Global AI Ethics and Governance Observatory at the 2024 Global Forum on the Ethics of Artificial Intelligence », in Digital Skills & Jobs Platform : <https://digital-skills-jobs.europa.eu/en/latest/news/unesco-launches-global-ai-ethics-and-governance-observatory-2024-global-forum-ethics>

<sup>2</sup> La croissance inclusive, le développement durable et le bien-être ; des valeurs centrées sur l'humain et l'équité ; la transparence et l'explicabilité ; la robustesse, la sécurité et la sûreté ; la responsabilité. Cf. cet article du 8 juin 2019, « G20 ministers agree on guiding principles for using artificial intelligence », Japan Times, <https://www.japantimes.co.jp/news/2019/06/08/business/g20-ministers-kick-talks-trade-digital-economy-ibaraki-prefecture/>

<sup>3</sup> Cf. la note du think tank américain Center for AI and Digital Policy, « G20 and Artificial Intelligence » : <https://www.caidp.org/resources/g20>

## C. LES FORUMS FERMÉS DU G7, DU PARTENARIAT MONDIAL SUR L'INTELLIGENCE ARTIFICIELLE ET DU CONSEIL DU COMMERCE ET DES TECHNOLOGIES UE-ÉTATS-UNIS

### 1. Le G7 et son « processus d'Hiroshima »

Le G7 a lancé en mai 2023 un dispositif appelé « processus d'Hiroshima » qui vise à définir de **grands principes pour régir l'utilisation de l'intelligence artificielle générative**<sup>1</sup>. Il s'agit d'un cadre d'action non contraignant fondé sur les principes suivants : promouvoir une IA sûre et digne de confiance ; fournir des orientations aux organisations qui développent et utilisent les systèmes d'IA ; analyser les possibilités et les défis que représente l'IA ; et promouvoir la coopération pour le développement d'outils et de pratiques en matière d'IA.

Cinq mois plus tard, le 30 octobre 2023, éclairé par un rapport commandé à l'OCDE sur ce processus d'Hiroshima<sup>2</sup>, le G7 a publié « Les principes directeurs internationaux du processus d'Hiroshima pour les organisations développant des systèmes d'IA avancés » et « Le code de conduite international du processus d'Hiroshima pour les organisations développant des systèmes d'IA avancés » et annoncé des projets de recherche conjoints sur l'IA générative avec le Partenariat mondial sur l'intelligence artificielle et l'OCDE.

Les **principes directeurs** et le **code de conduite** forment – avec le rapport commandé à l'OCDE et ces projets de recherche conjoints sur l'IA générative – le **cadre du processus d'Hiroshima**<sup>3</sup>. Ils abordent la « *conception, le développement, le déploiement et l'utilisation de systèmes d'IA avancés* » et intègrent un large éventail de principes internationaux existants, offrant un ensemble de **lignes directrices plus détaillées que les principes de l'OCDE** auxquels ils se réfèrent explicitement.

L'approche est fondée sur les risques tout au long du cycle de vie de l'IA, en commençant par des évaluations des risques avant le déploiement et des stratégies d'atténuation. Les développeurs et les déployeurs de modèles et de systèmes d'IA sont ainsi tenus de mettre en œuvre des procédures de gestion des risques, ainsi que des contrôles de sécurité robustes, y compris sous la forme d'exercices internes de *red teaming*. L'importance d'une surveillance, d'un signalement et d'une atténuation continus des abus et des incidents est soulignée. En outre, les principes et le code de conduite identifient les domaines prioritaires de la recherche et du développement de l'IA, tels que l'authentification du contenu, la protection des données personnelles et l'établissement de normes techniques.

---

<sup>1</sup> Cf. le « G7 Leaders' Statement on the Hiroshima AI Process », 2023 : <https://digital-strategy.ec.europa.eu/fr/library/g7-leaders-statement-hiroshima-ai-process>

<sup>2</sup> Rapport de l'OCDE du 7 septembre 2023, « G7 Hiroshima Process on Generative Artificial Intelligence : Towards a G7 Common Understanding on Generative AI » : <https://doi.org/10.1787/bf3c0c60-en>

<sup>3</sup> Cf. ce site qui présente le processus et tous les livrables du « Hiroshima AI Process Comprehensive Policy Framework » : <https://www.soumu.go.jp/hiroshimaaiprocess/en/index.html>

Les deux documents se réfèrent aux mêmes principes mais ne sont pas destinés aux mêmes publics : les **principes directeurs s'adressent à toutes les parties prenantes** tandis que le **code de conduite est d'abord destiné aux entreprises** développant ou déployant des systèmes d'IA.

Le cadre du processus d'Hiroshima est **non contraignant** et repose sur la **participation volontaire des États et des entreprises**. De nombreuses entreprises ont manifesté leur intérêt mais pas encore les grandes sociétés américaines de type MAAAM, ou comme Nvidia et IBM. Parmi les entreprises développant des modèles de pointe, seul **Anthropic s'est engagé à mettre en œuvre le code de conduite** du processus d'Hiroshima<sup>1</sup>.

Certains observateurs ont dénoncé un cadre **trop long sur les bonnes intentions et insuffisamment détaillé, concret et opérationnel**<sup>2</sup>. Le Président américain s'est pourtant en partie inspiré de ce cadre pour son *Executive Order* de 2023, rédigé avec l'appui du Nist. La présidente de la Commission européenne a, de son côté, estimé que le code de conduite était complémentaire à l'*AI Act*. En juin 2024, le G7 a décidé de préparer un outil de suivi de mise en œuvre du code de conduite.

Compte tenu de l'importance du sujet, à l'occasion de la réunion des ministres de l'industrie, de la technologie et du numérique du G7 les 14 et 15 mars 2024 consacrée à la future réglementation mondiale de l'IA, il a été décidé d'**ouvrir les débats à différentes organisations et plusieurs autres pays** : l'OCDE, le Programme des Nations unies pour le développement (PNUD), l'Unesco, l'Union internationale des télécommunications (UIT) et l'Envoyé du Secrétaire général des Nations unies pour la technologie ainsi que le Brésil, la Corée du Sud, l'Ukraine et les Émirats arabes unis ont été invités à se joindre à la conférence. Les participants sont convenus de mettre à jour les Principes directeurs et le Code de conduite du processus d'Hiroshima en fonction des nouveaux développements de l'IA.

Le 2 mai 2024, un an après le lancement du processus d'IA d'Hiroshima, le Premier ministre japonais Kishida Fumio a annoncé la création du **club des amis du processus de l'IA d'Hiroshima**, un groupe de 49 pays qui soutiennent l'esprit du cadre et ses lignes directrices volontaires.

En conclusion, même si la mise en œuvre de ce cadre reste incertaine, elle est néanmoins devenue un signe important de la **mobilisation internationale en faveur de la gouvernance mondiale de l'IA**.

---

<sup>1</sup> Cf. Anthropic, 30 novembre 2023, « *Thoughts on the US Executive Order, G7 Code of Conduct, and Bletchley Park Summit* » : <https://www.anthropic.com/news/policyrecap-q4-2023>

<sup>2</sup> Cf. Enza Iannopollo, 2023, « *The G7 AI Guidelines: Long On Good Intentions, Short On Detail And Substance* », Forrester : <https://www.forrester.com/blogs/the-g7-ai-guidelines-long-on-good-intentions-short-on-detail-and-substance/>

## 2. L'expertise du Partenariat mondial sur l'intelligence artificielle (PMIA) ou Global partnership on artificial intelligence (GPAI)

Le Partenariat mondial sur l'intelligence artificielle (PMIA ou, en anglais, *Global partnership on artificial intelligence* ou GPAI) est une initiative internationale proposée en 2018 au sommet du G7 par **la France et le Canada**.

Ce partenariat, qui s'inspirait du Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC ou en anglais *Intergovernmental Panel on Climate Change* ou IPCC), est officiellement fondé en 2020 et compte aujourd'hui **vingt-neuf membres**.

Il vise à promouvoir une IA responsable respectueuse des droits de l'homme et de la démocratie. Le partenariat est organisé autour de quatre thèmes : IA responsable et gouvernance de la donnée, basés à Montréal ; innovation et commercialisation, et futur du travail à Paris.

Le secrétariat du PMIA est assuré par l'OCDE, dont le siège est à Paris. Ce partenariat donne lieu à des **réunions annuelles** et à la publication d'un **rapport annuel** sous l'égide de son **groupe d'experts** (le GIEC du PMIA ou GPAI IPCC a ainsi pris le nom de GPAI MEG pour *Multi-stakeholders Experts Group*).

En 2024, le PMIA a lancé un **partenariat avec l'OCDE** et son prochain **sommet** qui se tiendra à Belgrade en Serbie en **décembre 2024** devrait réunir 44 pays et l'OCDE ainsi que des entreprises, des organisations internationales et des chercheurs.

## 3. Le Conseil du commerce et des technologies (CCT) UE-États-Unis

Le Conseil du commerce et des technologies (CCT) UE-États-Unis a été créé en 2021 pour **encourager le commerce bilatéral et réduire l'écart entre la régulation américaine souvent non contraignante et la législation européenne, souvent ambitieuse**, comme l'est celle de l'*AI Act* par exemple. L'UE est le premier marché d'exportation des États-Unis et les États-Unis sont le 2<sup>e</sup> plus grand importateur de produits européens. Cette économie transatlantique pèse plus de 1 100 milliards de dollars et justifie la **recherche de principes et de normes harmonisés**, autour de valeurs partagées (comme la protection des données personnelles ou les usages éthiques de l'IA). Depuis 2021, des réunions ont lieu régulièrement (sept à ce jour) conduisant l'UE et les États-Unis à mettre en place différentes initiatives.

Après une feuille de route conjointe sur les outils d'évaluation et de mesure pour une IA et une gestion des risques dignes de confiance adoptée en 2022, le CCT a mis en place trois groupes de travail pour avancer sur la voie de trois projets de collaboration en vue de faire converger les politiques en matière de risques et de créer des outils identiques et appropriés.

L'objectif est de créer une **taxonomie et une liste de termes communs**, de développer des **outils techniques (dont des métriques et des méthodologies de mesure)**, des **normes et des standards internationaux** et de **surveiller et mesurer les risques existants et les risques émergents** liés à l'IA. En avril 2024, afin d'avancer plus vite sur la sécurité de l'IA, les normes et les usages éthiques, un nouveau dialogue collaboratif est créé par le CCT entre le *EU AI Office* et le *US AI Safety Institute*.

Trois autres initiatives peuvent être mentionnées : des recherches sur les technologies améliorant la **protection de la vie privée** (depuis 2022) ; un rapport sur **l'impact de l'IA sur l'avenir des mains-d'œuvre américaines et européennes** (publié le 5 décembre 2022) suivi d'une Talent for Growth Taskforce chargée de l'amélioration de l'offre de formation permanente et de la diversification des pratiques de recrutement ; la préparation d'un **code de conduite sur l'IA à destination des entreprises** (mai 2023), abandonnée au motif que le processus d'Hiroshima du G7 poursuit le même objectif (ce travail du G7 a bien abouti le 30 octobre 2023).

#### D. LES INITIATIVES NON OCCIDENTALES

##### 1. La Proposition chinoise de gouvernance mondiale de l'IA

En octobre 2023, le président Xi Jinping a présenté une initiative pour la gouvernance mondiale de l'IA témoignant de l'effort stratégique de la Chine pour **construire et infléchir la future gouvernance mondiale de l'intelligence artificielle**<sup>1</sup>.

Cette initiative souligne l'engagement de la Chine à favoriser une **gouvernance équitable de ces technologies par la coopération internationale** en préconisant une approche équilibrée qui tienne compte à la fois des opportunités et des risques de l'IA. Elle se distingue par l'importance qu'elle accorde à la collaboration internationale en soulignant la nécessité de **veiller à ce que les pays en développement aient une voix significative dans la gouvernance mondiale de l'IA**, permettant à toutes les nations, quels que soient leur statut économique ou leur système politique, de participer au développement et à la supervision de l'IA. L'Initiative s'oppose ainsi aux monopoles technologiques et promeut la coopération mondiale pour empêcher l'utilisation abusive des technologies de l'IA.

---

<sup>1</sup> Cf. Wang Cong & Yin Yeping, « China launches Global AI Governance Initiative, offering an open approach in contrast to US blockade », *Global Times*, Oct. 18, 2023, <https://www.globaltimes.cn/page/202310/1300092.shtml> ; cf. aussi l'initiative en chinois [https://www.cac.gov.cn/2023-10/18/c\\_1699291032884978.htm](https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm) ou sa traduction en anglais [http://gd.china-embassy.gov.cn/eng/zxhd\\_1/202310/t20231024\\_11167412.htm](http://gd.china-embassy.gov.cn/eng/zxhd_1/202310/t20231024_11167412.htm)

## 2. Le travail des BRICS

Les BRICS représentent plus de la moitié de la population mondiale et plus du quart de la richesse mondiale. Composé de cinq pays (Brésil, Russie, Inde et Chine en 2009, rejoints par l’Afrique du Sud en 2011), le groupe s’est étendu le 1<sup>er</sup> janvier 2024 à l’Égypte, à l’Iran, aux Émirats arabes unis et à l’Éthiopie, et une vingtaine d’autres pays ont demandé leur adhésion<sup>1</sup>.

Depuis 2015, les BRICS ont jeté les bases de leur coopération en matière d’IA dans un « *Memorandum of Understanding on Science, Technology, and Innovation* », y soulignant que les technologies de l’information et des communications étaient un domaine essentiel de coopération. Dans leur **déclaration commune de 2017**, ces pays ont pour la première fois **explicitement mentionné l’IA** et en ont fait un domaine dans lequel ils devraient renforcer leurs efforts de coopération. Depuis, l’IA est régulièrement évoquée lors des sommets annuels et des rencontres ministérielles de l’organisation.

Une nouvelle étape a été franchie lors du sommet de Johannesburg en 2023, avec l’annonce par le président chinois Xi Jinping de la formation d’un **Groupe d’étude sur l’IA** sous l’égide de l’Institut des BRICS pour les réseaux du futur. Il s’agit à la fois de **surveiller les avancées technologiques** en matière d’IA, de **favoriser l’innovation** et d’établir un **cadre international solide pour la gouvernance de l’IA**. Le président Xi Jinping a annoncé que le groupe élaborerait des cadres et des normes de gouvernance de l’IA bénéficiant d’un large consensus. D’autres initiatives sur l’IA sont en cours au sein des BRICS, la Nouvelle Banque de Développement basée à Shanghai dirigée par l’ancienne présidente du Brésil Dilma Rousseff, dont l’Algérie est également membre, s’est par exemple dotée d’un **groupe de travail sur l’économie numérique et les investissements dans les applications d’IA**.

## 3. La stratégie de l’Union africaine

Les 55 États membres de l’Union africaine ont lancé, en 2013, « l’Agenda 2063 », un plan de développement de l’Afrique pour atteindre un développement socio-économique inclusif et durable en 50 ans.

---

<sup>1</sup> L’Indonésie, l’Arabie saoudite, le Mexique, la Corée du Sud, le Viêt-Nam, la Turquie, le Bahreïn, le Bangladesh, la Biélorussie, la Bolivie, Cuba, le Honduras, le Kazakhstan, le Koweït, le Nigeria, la Palestine, la Serbie, le Sénégal, la Thaïlande, le Venezuela et l’Argentine (dont le nouveau Président a cependant renoncé à la candidature souhaitant se rapprocher des États-Unis). L’Algérie a vu sa demande d’adhésion rejetée en août 2023 et a renoncé à sa candidature.



L'Agence de développement de l'Union africaine dans le cadre du Nouveau Partenariat pour le développement de l'Afrique (AUDA-NEPAD), a publié le 29 février 2024, lors de la conférence « AI Dialogue » un **livre blanc intitulé « Régulation et adoption responsable de l'IA en Afrique vers la réalisation de l'Agenda 2063 de l'Union africaine »**<sup>1</sup>, élaboré pendant deux ans par un groupe de haut niveau sur les technologies émergentes (*African Union's High-Level Panel on Emerging Technologies* ou APET).

En s'inspirant des lignes directrices de l'Unesco, le document exhorte toute future stratégie de l'Union africaine en matière d'IA à **intégrer des principes éthiques** et demande aux pays africains de mettre en œuvre des **stratégies nationales d'IA responsables** en mettant l'accent sur des outils juridiques qui renforceront les valeurs d'équité, de sécurité, de confidentialité et de sûreté. Cependant, il n'aborde pas directement les modalités de gouvernance ni le détail des défis réglementaires et juridiques spécifiques de l'IA générative.

En février 2024, lors d'une réunion du Conseil exécutif de l'Union africaine (44<sup>e</sup> session), la Commission de l'Union africaine a appelé à une **stratégie continentale pour l'IA**, avec une feuille de route complète permettant aux nations africaines de développer de manière responsable les technologies d'IA. Un groupe de travail a été chargé de développer cette stratégie continentale en s'appuyant sur le livre blanc de l'AUDA-NEPAD. Le Comité technique de la communication et des technologies de l'information de l'Union africaine lors de sa réunion de juin 2024 a adopté à l'unanimité la stratégie continentale pour l'IA de l'Union africaine ainsi que le Pacte numérique africain, document distinct détaillant la stratégie de l'Afrique pour gérer son avenir numérique et promouvoir le progrès sociétal global.

Le Conseil exécutif de l'Union africaine a formellement adopté ces documents lors de sa 45<sup>e</sup> session qui s'est tenue à Accra en juillet 2024<sup>2</sup>. La commissaire aux infrastructures, à l'énergie et à la numérisation de l'Union africaine, Amani Abou-Zeid a déclaré qu'ils « *fourniront des orientations sur l'utilisation de la technologie pour trouver des solutions aux défis de l'Afrique, aideront à accélérer de nombreux projets et programmes, et protégeront contre l'utilisation non éthique de la technologie : la technologie doit nous aider à préserver notre identité, nos langues et nos cultures, et nous être utile plutôt que de nous nuire* ». Il s'agit aussi de créer plus globalement un environnement favorable au développement et à l'utilisation des technologies numériques et d'aider les gouvernements des différents États membres à élaborer leurs politiques en matière d'IA et de régulation du secteur numérique.

---

<sup>1</sup> Auda-Nepad, « Regulation and Responsible Adoption of AI in Africa Towards Achievement of African Union Agenda 2063 » : <https://dig.watch/resource/auda-nepad-white-paper-regulation-and-responsible-adoption-of-ai-in-africa-towards-achievement-of-au-agenda-2063>

<sup>2</sup> Cf. cet article : <https://pouvoirsafrique.com/article/1249/blunion-africaine-lance-une-strategie-pour-lia-et-un-pacte-pour-le-numeriqueb>

## ***E. LES AUTRES PROJETS DE GOUVERNANCE MONDIALE : CONSEIL DE L'EUROPE, FORUM ÉCONOMIQUE MONDIAL, INITIATIVES ÉMANANT DU SECTEUR PRIVÉ...***

### **1. La Convention-cadre sur l'IA du Conseil de l'Europe**

Avec ses 48 États membres et riche de ses plus de 220 conventions internationales, le Conseil de l'Europe a préparé depuis la fin de l'année 2019 à travers sa commission spéciale sur l'IA une **Convention-cadre sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit**, qui a été adoptée le 17 mai 2024. Un rapport explicatif a été joint au texte<sup>1</sup>.

Il s'agit du **premier traité international juridiquement contraignant** visant à garantir une utilisation des systèmes d'intelligence artificielle pleinement conforme aux droits humains, à la démocratie et à l'État de droit. Il **nécessitera cependant d'être régulièrement ratifié** par chacun des États qui auront décidé de le signer.

Ce traité international qui vise à garantir une IA respectueuse des droits fondamentaux a été **ouvert à la signature le 5 septembre 2024**, lors d'une conférence des ministres de la justice des États membres du Conseil de l'Europe organisée à Vilnius. À ce stade, il a été signé par Andorre, la Géorgie, l'Islande, la Norvège, la République de Moldova, Saint-Marin, le Royaume-Uni ainsi qu'Israël, les États-Unis d'Amérique et l'Union européenne.

La Convention-cadre vise à garantir que les activités menées aux différentes étapes du cycle de vie des systèmes d'intelligence artificielle sont pleinement compatibles avec les droits humains, la démocratie et l'État de droit, tout en étant favorable au progrès et aux innovations technologiques. Elle apporte des compléments aux normes internationales existantes relatives aux droits humains, à la démocratie et à l'État de droit et a surtout pour but de pallier à tout vide juridique qui pourrait résulter d'avancées technologiques rapides. Afin de résister au temps, la Convention-cadre ne régule pas la technologie et est neutre sur le plan technologique. Une conférence des parties fera office de commission exécutive du traité et facilitera la coopération entre les signataires.

La Convention-cadre du Conseil de l'Europe est critiquée par l'Assemblée parlementaire du Conseil de l'Europe, le Contrôleur européen de la protection des données et plusieurs organisations de la société civile du fait que, sous la pression de certains États membres ou observateurs (notamment les États-Unis, le Royaume-Uni, le Canada et le Japon), **les obligations pesant sur les entreprises privées sont en réalité facultatives**.

---

<sup>1</sup> Le lien suivant contient les deux documents : <https://www.coe.int/fr/web/artificial-intelligence/la-convention-cadre-sur-l-intelligence-artificielle>

Les parties au traité ont en effet le choix entre appliquer la Convention-cadre aux acteurs privés ou bien traiter des risques et des impacts découlant des activités menées par les acteurs privés d'une manière conforme à l'objet et au but de la convention, en prenant « *d'autres mesures pour se conformer aux dispositions du traité tout en respectant pleinement leurs obligations internationales en matière de droits de l'homme, de démocratie et d'État de droit* ».

## **2. L'Alliance pour la gouvernance de l'IA proposée par le Forum économique mondial**

Le Forum économique mondial a lancé une **Alliance pour la gouvernance de l'IA**<sup>1</sup>. Déterminée à promouvoir une IA inclusive, éthique et durable, cette Alliance de **603 membres, dont 463 organisations**, se concentre sur le développement des innovations, l'intégration des technologies d'IA dans les secteurs économiques et les impacts pratiques de l'adoption de l'IA.

Elle veille à la collaboration de plusieurs groupes de travail thématiques en vue d'**accélérer les progrès technologiques et sociétaux** avec des systèmes d'IA sûrs et avancés et de **rationaliser la gouvernance de l'IA** grâce à des cadres réglementaires solides et la promotion de **normes techniques**.

## **3. Des principes et bonnes pratiques proposés par les entreprises au Partnership on AI lancé en 2016 par sept géants de l'IA**

Les entreprises, notamment les MAAAM, ont non seulement publié **leurs propres principes et standards en matière d'IA** (par exemple Google, Microsoft, Meta et Anthropic<sup>2</sup>) qui prévoient des bonnes pratiques en amont du déploiement des systèmes d'IA (nettoyage des données, vigilance sur les sources et les données synthétiques ; recours à des évaluations et à des tests des modèles, y compris en *red teaming* ; lutte contre les biais ; alignement des modèles avec des principes éthiques) et en aval (politique de sécurité ; vigilance à la mise sur le marché et après ; transparence et accès aux sources par les RAG par exemple ; *watermarking* des contenus générés) mais elles ont aussi pris des initiatives en faveur de la gouvernance globale de l'IA.

---

<sup>1</sup> World Economic Forum, « AI Governance Alliance » : <https://initiatives.weforum.org/ai-governance-alliance/home>

<sup>2</sup> Cf. ces quatre exemples de références en matière d'AI safety, Google, « Our Principles » : <https://ai.google/responsibility/principles/> ; Microsoft, « Principles and Approach » : <https://www.microsoft.com/en-us/ai/principles-and-approach> ; Meta, « Responsible AI : Driven by Our Belief that AI Should Benefit Everyone » : <https://ai.meta.com/responsible-ai/> ; et Anthropic, « Make Safe AI Systems, Deploy Them Reliably » : <https://www.anthropic.com/research>

Un *Partnership on AI* est créé en 2016. Fondée par **Amazon, Facebook, Google, DeepMind, Microsoft et IBM**, rejoints par **Apple** en 2017, cette association regroupe désormais plus d'une centaine de structures, non seulement des entreprises mais aussi des associations et des organismes du monde de la recherche, ainsi que des universitaires<sup>1</sup>. Selon sa propre définition, elle se présente comme un « *centre de ressources pour les décideurs politiques, par exemple pour mener des recherches qui éclairent les meilleures pratiques en matière d'IA et pour explorer les conséquences sociétales de certains systèmes d'IA, ainsi que les politiques entourant le développement et l'utilisation de ces systèmes* ». Il a publié de nombreux articles, rapports et recommandations depuis sa création, dont, en 2023, un **rapport sur l'utilisation des données synthétiques** et un **guide pour le déploiement de modèles de fondation sûrs**. Google, Meta, Microsoft, Apple, OpenAI, le Alan Turing Institute et le Ada Lovelace Institute, ont, avec de nombreuses autres entreprises et organisations, fait part de leur soutien à ces travaux relatifs à la sécurité des modèles d'IA.

Le *Partnership on AI* est membre du consortium américain de l'*AI Safety Institute*, a le statut de membre du conseil consultatif de la société civile au sein du réseau des experts de l'OCDE et a eu le statut d'observateur au sein du comité chargé de la rédaction du traité international sur l'IA du Conseil de l'Europe.

#### **4. Le Forum sur les modèles de pointe ou *Frontier Model Forum* et les autres initiatives**

Créé en 2023, le **Forum sur les modèles de pointe**, ou *Frontier Model Forum*, est un partenariat restreint entre les entreprises américaines développant les systèmes d'IA les plus avancés : **Microsoft, OpenAI, Google et Anthropic**, qui se proposent de définir les conditions du développement d'IA sûres et responsables<sup>2</sup>. Il s'agit en effet d'aider à :

- **faire progresser la recherche sur la sécurité de l'IA** afin de promouvoir le développement responsable des modèles de pointe et minimiser les risques potentiels ;
- identifier les **meilleures pratiques de sécurité** pour les modèles de pointe ;
- **partager les connaissances** avec les décideurs politiques, les universitaires, la société civile et d'autres pour faire progresser le développement responsable de l'IA ;
- soutenir les efforts visant à tirer parti de l'IA pour relever les plus **grands défis sociétaux**.

---

<sup>1</sup> Cf. son site actif depuis déjà huit ans : <https://www.partnershiponai.org>

<sup>2</sup> Cf. l'annonce sur le blog de Microsoft le 26 juillet 2023, « Microsoft, Anthropic, Google, and OpenAI launch Frontier Model Forum » : <https://blogs.microsoft.com/on-the-issues/2023/07/26/anthropic-google-microsoft-openai-launch-frontier-model-forum/> ainsi que le site d'OpenAI dédié au Frontier Model Forum : <https://openai.com/index/frontier-model-forum>

Le Forum vise aussi la **divulgence responsable des vulnérabilités ou des capacités dangereuses au sein des modèles de pointe**.

Le Forum a d'ores et déjà publié des notes. Les membres du Forum et leurs partenaires ont également créé un Fonds pour la sécurité de l'IA doté initialement de plus de 10 millions de dollars en vue de soutenir la recherche indépendante sur la sécurité de l'IA, comme de nouvelles évaluations des modèles et de nouvelles techniques de *red teaming* des modèles d'IA. Ce Fonds se veut un élément de l'accord sur les engagements volontaires en matière d'IA signé à la Maison-Blanche dans le cadre de l'*Executive Order* de 2023. Le 1<sup>er</sup> avril 2024, le Forum a annoncé qu'il avait accordé la première série de subventions au titre du Fonds.

De son côté, Meta a créé en décembre 2023 avec IBM et plus de 50 membres fondateurs une **Alliance pour l'IA** (*AI Alliance*). Ils ont été rejoints par 20 autres membres en avril 2024. Cette initiative se singularise par sa volonté de promouvoir la science ouverte et l'innovation ouverte en IA<sup>1</sup>. Elle a déjà mis en place des groupes de travail (l'un porte sur les outils de sécurité et de confiance, l'autre sur les politiques publiques).

Le consortium **MLCommons** lancé en 2020 est quant à lui composé de 125 membres et vise la coopération dans la communauté du *Machine Learning*. Il a produit un *benchmark* apprécié des aspects software et hardware de l'apprentissage machine et offre des jeux de données et des outils utiles à tous les chercheurs. Il se propose également de définir les bonnes pratiques et les normes des systèmes d'IA relevant du *Machine Learning*. En 2023, ce consortium a lancé un groupe de travail sur la sécurité des systèmes d'IA, avec un intérêt pour l'évaluation des LLM selon le cadre proposé par le CRFM de Stanford. Anthropic, Coactive AI, Google, Inflection, Intel, Meta, Microsoft, Nvidia, OpenAI, Qualcomm Technologies sont, avec d'autres entreprises, au sein de ce groupe de travail qui a annoncé une première preuve de concept (*Proof of Concept*) en avril 2024.

Microsoft a initié en 2021 une **association autour de l'origine des contenus**, la *Coalition for Content Provenance and Authenticity* (C2PA), avec Adobe, Arm, BBC, Intel, and Truepic. Depuis, de nombreuses entreprises comme Google, OpenAI, xAI ou Sony, ont rejoint la C2PA.

Microsoft et OpenAI ont créé en 2024 un **fonds pour la résilience sociétale**, le *Societal Resilience Fund*, qui avec 2 millions de dollars vise à encourager **l'éducation** et la connaissance de l'IA auprès des électeurs et des communautés fragiles. Ce fonds sert par exemple à verser des subventions à des associations conduisant des projets éducatifs comme le *Partnership on AI* ou la C2PA.

---

<sup>1</sup> Cf. le site de l'AI Alliance, « Building the Open Future of AI » : <https://thealliance.ai>

Toujours en 2024 a été créée la *Tech Coalition*, autour d'Adobe, Amazon, Bumble, Google, Meta, Microsoft, OpenAI, Roblox, Snap Inc. et TikTok. Ce groupe finance des recherches sur la **pédopornographie, l'exploitation d'enfants et les abus sexuels en ligne en lien avec l'IA générative**.

Enfin, un *Tech Accord to Combat Deceptive Use of AI in Elections*, un accord pour **lutter contre l'utilisation trompeuse de l'IA lors des élections**, réunit, depuis février 2024, 20 signataires, dont Google, Meta, Microsoft, IBM, xAI, Anthropic, OpenAI et StabilityAI autour d'un ensemble de huit engagements pour lutter contre la désinformation politique par l'IA, notamment dans le cadre des élections qui doivent se tenir dans plus de 40 pays en 2024.

Au total, il faut retenir cette **volonté des entreprises du secteur de l'IA de s'organiser en groupes sectoriels**, soit pour servir de **centre de ressources** aux décideurs politiques comme le *Partnership on AI*, soit pour **soutenir la recherche sur la sécurité de l'IA**, identifier les **bonnes pratiques** pour les modèles de pointe et **partager les connaissances** avec les parties prenantes, comme le *Frontier Model Forum*. Certaines initiatives visent des **aspects plus thématiques** : éducation, science ouverte, authenticité des contenus, pédopornographie, désinformation à caractère politique, etc. Toutes ces démarches sont des exemples d'autorégulation comme l'a expliqué Florence G'sell mais l'on peut déplorer qu'elles ne conduisent pas, le plus souvent, à des normes précises et encore moins à des règles juridiquement contraignantes. Leur intérêt réside peut-être plutôt dans leur capacité volontaire ou non à **fournir des informations précieuses aux législateurs et régulateurs qui envisagent des cadres juridiques contraignants pour les systèmes d'IA**.

## **F. LE CADRE EN CONSTRUCTION DES SOMMETS POUR LA SÉCURITÉ DE L'INTELLIGENCE ARTIFICIELLE ET DES INSTITUTS DE SÉCURITÉ DE L'IA**

### **1. Un réseau international d'agences pour la sécurité de l'intelligence artificielle**

Les sommets pour la sécurité de l'intelligence artificielle (en anglais, *AI safety summits*) sont des **conférences internationales d'initiative britannique visant à anticiper et encadrer les risques potentiels liés à l'intelligence artificielle**. Elles ont conduit plusieurs pays, après le Royaume-Uni et les États-Unis en novembre 2023, à mettre en place des *AI Safety Institutes*.



Le Canada, l’Australie, le Japon, la Corée du Sud, Singapour, le Kenya, la France et l’UE ont ainsi rejoint les États-Unis et le Royaume-Uni au sein d’une **nouvelle coordination internationale**, appelée « *International Network of Cooperation of the National AI Safety Institutes* » (AISI)<sup>1</sup>, dont la première réunion s’est tenue à San Francisco les 21 et 22 novembre 2024, réunissant les experts de neuf pays et de l’Union européenne.

La secrétaire américaine au commerce, Gina Raimondo avait annoncé la naissance de ce réseau lors du sommet pour la sécurité de l’intelligence artificielle organisé par la Corée du Sud à Séoul en mai 2024. La réunion de San Francisco a permis le lancement effectif et les modalités de cette coopération internationale des AISI en vue d’avancer concrètement en termes de **sécurité** de l’IA, de **normes techniques**, d’échanges de **bonnes pratiques** et de **partage de connaissances**.

## 2. Du sommet de Bletchley Park au rapport de Yoshua Bengio

Ces sommets sont relativement récents, la première édition ayant eu lieu en **novembre 2023 à Bletchley Park**, situé à mi-chemin d’Oxford et de Cambridge, lieu symbolique pour l’informatique et l’IA car associé à la mémoire d’Alan Turing. Ils réunissent à la fois des chefs d’États et de gouvernements, mais également des représentants d’organisations internationales comme Antonio Guterres, secrétaire général de l’ONU, et des représentants d’entreprises privées comme Elon Musk ainsi que du monde de la recherche comme Yoshua Bengio.

Les cinq objectifs du premier sommet étaient de parvenir à un **consensus sur les risques associés à l’IA de pointe** ; de **faire progresser la coopération internationale** par le biais de cadres de travail nationaux et internationaux ; de déterminer des **mesures de sécurité** adaptées aux entreprises du secteur privé ; d’identifier des domaines de recherche collaborative en matière de sécurité de l’IA ; et de mettre en lumière les aspects bénéfiques de l’IA.

---

<sup>1</sup> Un think tank américain, le Center for Strategic and International Studies (CSIS), vient de publier le 30 octobre 2024 un rapport sur les enjeux de ce réseau, cf. Gregory C. Allen and Georgia Adamson, 2024, « The AI Safety Institute International Network : Next Steps and Recommendations » : <https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations>



Le premier sommet, dont une partie des discussions a été rendue publique<sup>1</sup>, a débouché sur la rédaction d'une **déclaration commune** de l'ensemble des participants appelée « la Déclaration de Bletchley ». Cette déclaration, plutôt pessimiste, a souligné le **besoin urgent d'une collaboration internationale** pour gérer les risques potentiels associés aux systèmes d'IA de pointe et a reconnu **l'existence de risques existentiels pour l'humanité** qui pourraient être induits par un développement rapide et non maîtrisé de l'intelligence artificielle. La déclaration appelle ainsi à se concentrer sur la notion de sécurité de l'intelligence artificielle tout au long de sa chaîne de valeur lors des prochains sommets, organisés de façon semestrielle.

Le sommet a également débouché sur une **déclaration relative aux tests de sécurité**, signée par l'Union européenne, dix pays (États-Unis, Royaume-Uni, France, Allemagne, Italie, Canada, Australie, Japon, Corée du Sud et Singapour) et des entreprises d'IA de premier plan, telles que OpenAI, Google, Anthropic, Amazon, Mistral, Microsoft et Meta, qui **pourront s'engager à soumettre leurs modèles de pointe aux gouvernements pour qu'ils réalisent des tests de sécurité**. Le document qui n'est pas juridiquement contraignant réitère qu'**il incombe aux gouvernements de financer ces tests et d'évaluer les nouveaux modèles d'IA développés par les entreprises avant leur mise sur le marché**. Il invite à se concentrer sur les évaluations des « *risques liés à la sécurité nationale* », plutôt que sur les dommages potentiels causés par les utilisations au quotidien.

---

<sup>1</sup> Florence G'sell a utilisé les résumés des tables rondes pour en faire une courte synthèse. La première table ronde s'est intéressée aux menaces à la sécurité mondiale, telles que la biosécurité et la cybersécurité, et a appelé à une collaboration intersectorielle urgente. La deuxième table ronde a discuté de l'imprévisibilité de l'évolution des capacités de l'IA, en soulignant les avantages pour les soins de santé, mais aussi les risques substantiels et en insistant sur la nécessité d'une surveillance et de tests de sécurité rigoureux. La troisième table ronde a exploré les risques existentiels potentiels liés à la perte de contrôle de l'IA avancée, en préconisant des tests de sécurité complets et des recherches plus poussées. La quatrième table ronde a abordé les risques sociétaux, notamment les menaces pour la démocratie et les droits de l'homme. Elle a recommandé d'impliquer le public dans les efforts de recherche. La cinquième table ronde a souligné la nécessité d'un développement rapide de politiques de sécurité de l'IA et l'importance de la réglementation gouvernementale, en notant que les politiques des entreprises à elles seules sont insuffisantes. D'autres tables rondes ont discuté des rôles et des actions nécessaires de la part des différentes parties prenantes pour faire face aux risques et aux opportunités de l'IA. La table ronde 6 a souligné la nécessité pour les décideurs politiques nationaux d'équilibrer les risques et les opportunités grâce à une gouvernance rapide, agile et innovante, tout en favorisant la collaboration internationale malgré des contextes nationaux différents. La table ronde 7 s'est concentrée sur les priorités de la communauté internationale, notamment le développement d'une compréhension commune des capacités et des risques de l'IA, la coordination de la recherche sur la sécurité et la garantie des avantages généralisés de l'IA. La table ronde 8 a souligné l'importance pour la communauté scientifique de comprendre les risques existants, de collaborer avec les gouvernements et le public et d'éviter la concentration du pouvoir. La table ronde 9 a souligné la nécessité de développer les compétences publiques et d'améliorer les capacités techniques gouvernementales pour optimiser les avantages potentiels de l'IA. Enfin, les tables rondes 10 et 11 ont porté sur la collaboration internationale dans la lutte contre la désinformation et les deepfakes et sur la garantie que toutes les régions du monde bénéficient du potentiel de l'IA.

Ces discussions ont également conduit à établir un **groupe international de 75 experts de 30 pays** chargé de rédiger un rapport annuel sur la politique et la régulation de l'IA, selon un format proche de celui utilisé par le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC). Le Partenariat mondial sur l'intelligence artificielle (PMIA ou, en anglais, *Global partnership on artificial intelligence* ou GPAI) le propose aussi et en octobre 2023, Eric Schmidt, ancien PDG de Google, avec plusieurs autres dirigeants du secteur, avait également proposé un tel groupe d'experts.

Le sommet de Bletchley Park a, enfin, demandé plus spécifiquement la préparation par ce groupe, en vue du prochain sommet, d'un **rapport précis sur l'état de la science concernant l'IA de pointe** (« *State of the Science Report on Frontier AI* »).

Le sommet suivant s'est déroulé six mois plus tard, en mai 2024, en visioconférence et en présentiel à Séoul et a donné lieu à la publication d'un **rapport scientifique international sur la sécurité de l'IA dirigé par Yoshua Bengio**, que ce dernier a pu présenter à vos rapporteurs lors de son audition.

Il a affirmé que la préparation de ce rapport a permis de réunir une sorte d'équivalent du GIEC mais pour l'IA et a fourni une évaluation actualisée et scientifiquement fondée de la sécurité des systèmes d'IA de pointe. Ce document est destiné aux décideurs publics et vulgarise les connaissances scientifiques dans le domaine<sup>1</sup>.

Le rapport met en évidence plusieurs points clés concernant les modèles d'IA avancés. Il met l'accent sur la double nature de l'IA, son **potentiel pour améliorer le bien-être, l'économie et la science** mais aussi ses **dangers** : l'utilisation malveillante de l'IA peut entraîner de la **désinformation** à grande échelle, des opérations d'influence, des **fraudes** et des escroqueries, tandis que des systèmes d'IA défectueux pourraient produire des **décisions biaisées** affectant des groupes ou des personnes à raison de leur race, sexe, culture, âge ou handicap. Si les capacités de l'IA progressent rapidement, des défis fondamentaux demeurent pour les chercheurs comme la **compréhension du fonctionnement interne des IA** ou ses **modes de raisonnement**, son rapport à la causalité par exemple.

Le rapport souligne **l'incertitude qui entoure l'avenir de l'IA**, avec de nombreux scénarios possibles : une **évolution lente** ou des progrès extrêmement rapides entraînant des **risques systémiques** (perturbations du marché du travail et inégalités économiques), voire des **risques existentiels** (perte de contrôle de l'IA et conséquences catastrophiques sur l'humanité). Si des méthodes techniques, telles que des tests, le *red teaming* et l'audit des données d'entraînement, peuvent atténuer certains risques, elles ont des limites et ne permettent pas de traiter l'ensemble des risques.

---

<sup>1</sup> Yoshua Bengio et al., 2024, « *International Scientific Report on the Safety of Advanced AI* » : [https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international\\_scientific\\_report\\_on\\_the\\_safety\\_of\\_advanced\\_ai\\_interim\\_report.pdf](https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf)

Le sommet de Séoul a également été l'occasion pour 16 entreprises d'IA de s'engager à se doter d'un **référentiel en matière de sécurité** et de structures de contrôle et de gouvernance<sup>1</sup>.

En clôture de la réunion, le Royaume-Uni et la Corée du Sud, hôtes du sommet, ont obtenu des participants l'engagement de poursuivre et d'approfondir la **recherche sur la sécurité de l'IA et les seuils de risque pour les modèles de pointe**<sup>2</sup>.

Des pays, comme **l'Italie ou l'Allemagne**, signataires de la déclaration finale du sommet de Séoul en faveur de la coopération internationale en matière de sécurité d'IA, ont, en outre, fait part de leur volonté de rejoindre le réseau des AISI (*International Network of Cooperation of the National AI Safety Institutes*). Cependant, une annonce de septembre 2024 du secrétaire d'État américain Antony Blinken expliquait que le Kenya serait le seul nouveau membre du réseau à ce stade.

### 3. Un sommet en France les 10 et 11 février 2025

**La France est chargée de l'organisation du prochain sommet** qui se tiendra les 10 et 11 février 2025, et à ce titre, le Président de la République a nommé Anne Bouverot, coprésidente de la Commission de l'intelligence artificielle en 2024, comme son « envoyée spéciale » chargée de l'organisation du sommet. Vos rapporteurs peuvent d'ores et déjà affirmer que le sommet sera organisé par « verticales » thématiques et **n'aura donc pas comme sujet exclusif la sécurité en matière d'intelligence artificielle**.

Vos rapporteurs considèrent qu'il s'agit d'une bonne nouvelle dans la mesure où, même si elle pose des questions certaines, la question de la sécurité n'est pas le seul sujet digne d'intérêt en matière d'intelligence artificielle. Les pouvoirs publics à travers le monde doivent **se saisir du sujet de l'IA de la façon la plus complète possible**, et ce sommet international peut être une occasion d'y parvenir.

C'est pourquoi, en plus des **recommandations générales** concernant les nouveaux développements de l'intelligence artificielle, vos rapporteurs ont également des **recommandations plus spécifiquement destinées au Président de la République et aux organisateurs du futur sommet des 10 et 11 février 2025 de l'intelligence artificielle**. Ces recommandations visent à s'assurer qu'une vision française éclairée de l'intelligence artificielle pourra s'exprimer à travers ce sommet.

---

<sup>1</sup> Communiqué du 21 mai 2024 « Historic first as companies spanning North America, Asia, Europe and Middle East agree safety commitments on development of AI » :

<https://www.gov.uk/government/news/historic-first-as-companies-spanning-north-america-asia-europe-and-middle-east-agree-safetycommitments-on-development-of-ai>

<sup>2</sup> Communiqué du 22 mai 2024 « New commitment to deepen work on severe AI risks concludes AI Seoul Summit » : <https://www.gov.uk/government/news/new-commitmentto-deepen-work-on-severe-ai-risks-concludes-ai-seoul-summit>

### III. LES PROPOSITIONS DE L'OFFICE

Vos rapporteurs ont élaboré **18 recommandations**, dont **cinq sont consacrées à la préparation du prochain sommet sur l'intelligence artificielle** que la France organisera les 10 et 11 février 2025. Ils ont choisi de **limiter le nombre de préconisations** de manière à pouvoir communiquer plus efficacement. Ils demandent au gouvernement et au Président de la République de traduire ces recommandations en **mesures effectives rapidement opérationnelles**.

#### A. LES PROPOSITIONS À SOUTENIR DANS LE CADRE DU FUTUR SOMMET DE L'IA

##### 1. Faire reconnaître le principe d'une approche transversale de l'IA et renoncer à l'approche exclusivement tournée vers les risques

La France doit profiter de l'organisation du prochain sommet pour contribuer à établir, selon une approche transversale, un **cadre international de référence pour la future gouvernance mondiale des systèmes d'intelligence artificielle**. Les sommets pour la sécurité de l'intelligence artificielle (*AI safety summits*) se sont tenus à la suite d'une initiative britannique visant à anticiper et encadrer les risques de l'intelligence artificielle dont les risques existentiels, il faut maintenant aller plus loin.

Le Sommet pour l'action sur l'IA des 10 et 11 février 2025 doit avoir le souci d'élargir les problématiques abordées et de poser les jalons de cette gouvernance mondiale. C'est pourquoi au-delà de la sécurité, il faut mettre à profit le fait que cinq thèmes essentiels seront l'objet du sommet<sup>1</sup> (l'IA au service de l'intérêt public avec la question des infrastructures ouvertes ; l'avenir du travail ; la culture ; l'IA de confiance ; la gouvernance mondiale de l'IA) pour faire accepter au plus haut niveau **le principe d'une approche transversale des enjeux de l'IA**, qui devra se matérialiser solennellement dans une déclaration finale des participants.

La France étant chargée de l'organisation du sommet, elle pourra non seulement élargir ponctuellement les problématiques et les perspectives du travail collectif mais surtout **consacrer cette logique multidimensionnelle des problématiques de l'IA**. Les interrogations sur les **risques existentiels** de l'IA, dont la concrétisation reste très incertaine, **ne doivent pas prendre une part trop grande au sein de cet événement international** et des futurs sommets : il convient plutôt de se préoccuper d'enjeux plus directs. En effet, les systèmes d'intelligence artificielle présentent d'ores et déjà des risques réels et certains.

---

<sup>1</sup> Cf. la présentation du futur sommet sur le site de l'Élysée : <https://www.elysee.fr/sommet-pour-l-action-sur-l-ia>

L'Office a déjà eu l'occasion de consacrer plusieurs travaux à ce sujet, notamment son rapport pionnier de 2017, où il se prononçait en faveur d'une intelligence artificielle maîtrisée, utile et démystifiée, prônant l'usage éthique d'algorithmes sûrs, transparents et non discriminatoires. Il y réclamait une politique publique de l'intelligence artificielle permettant la diffusion de ces technologies dans l'économie et la société française, à travers une politique de formation initiale et continue ambitieuse, une vigilance à l'égard de la domination par la recherche privée et les entreprises américaines et l'organisation de grands débats publics évoluant en fonction de l'état de l'art de la recherche en la matière.

Nous estimons à cet égard que les cinq thèmes retenus pour le prochain sommet de l'IA **éludent deux dimensions** à prendre en compte de manière prioritaire :

- **l'éducation**, qui pourrait être ajoutée à la verticale « culture » avec pour intitulé « **éducation et culture** » ;

- **la souveraineté numérique**, qui pourrait être ajoutée à la verticale « l'IA au service de l'intérêt public » avec pour intitulé « **souveraineté numérique et intérêt général** ».

## **2. Proposer de placer la gouvernance mondiale de l'IA sous l'égide d'une seule organisation internationale**

Ce sommet doit être l'occasion d'apporter un minimum de **clarification et de rationalisation dans la dizaine de projets de gouvernance mondiale de l'IA**, c'est pourquoi il est proposé de **placer la gouvernance mondiale de l'IA sous l'égide d'une seule organisation internationale**, à savoir l'ONU, seule organisation pleinement légitime sur le plan multilatéral.

L'importance et la spécificité du sujet invitent à créer une nouvelle **institution spécialisée membre du système des Nations unies**, dont les compétences s'étendraient de la **coordination internationale de la régulation de l'IA à la lutte contre la fracture numérique mondiale**, plutôt qu'à confier ces responsabilités à une agence internationale déjà existante.

Ce serait la **suite logique du Pacte numérique mondial**, du comité scientifique international sur l'IA et du dialogue mondial sur la gouvernance de l'IA que les Nations unies ont mis en place, en septembre 2024.

Il s'agit aussi de l'une des **propositions formulées par plusieurs chercheurs** qui ont évoqué différentes pistes de gouvernance internationale de l'IA<sup>1</sup>. L'approche de l'OCDE doit nourrir le travail de cette future organisation internationale de l'IA.

### **3. Initier le cadre d'une régulation globale et multidimensionnelle de l'IA en s'inspirant des travaux de l'OCDE et de l'UE**

L'approche de la régulation mondiale de l'IA doit être **multidimensionnelle**, ainsi que le présent rapport dans ses développements sur la chaîne de valeur de l'IA ou, différemment, les travaux de l'OCDE le prévoient. L'OCDE ne reprend pas l'idée de chaîne de valeur telle qu'elle est développée dans le présent rapport mais elle propose de distinguer la multidimensionnalité de l'IA selon plusieurs phases.

Puisque les modèles d'IA se décomposent en **au moins quatre phases distinctes et interdépendantes** (l'OCDE retient : le contexte, les données d'entrées, les modèles eux-mêmes et les tâches demandées qui engendrent les sorties de l'IA), **toutes les politiques publiques de régulation de ces technologies devraient prendre en compte la complexité de la chaîne de valeur de l'IA ou cette multidimensionnalité** lors de la définition et de la mise en œuvre de règles relatives à l'IA. Les dimensions en jeu aux stades de la conception, de l'entraînement, des réglages ou encore de l'utilisation des modèles d'intelligence artificielle sont autant d'objets qui justifient une prise en compte spécifique par la régulation.

---

<sup>1</sup> Outre un soutien à l'accès aux systèmes d'IA de pointe (y compris pour des pays en développement) et l'établissement de normes de sécurité internationales, ces chercheurs ont recensé plusieurs cadres institutionnels possibles à l'échelle globale pour relever les défis de l'IA, notamment une commission mondiale sur l'IA, qui faciliterait le consensus des experts sur les opportunités et les risques associés à l'IA avancée, ou une organisation internationale de gouvernance de l'IA multipartite, qui établirait des normes internationales pour la gestion des menaces provenant des modèles d'IA de pointe, soutiendrait leur mise en œuvre et surveillerait potentiellement la conformité avec un futur régime de gouvernance. Un autre modèle proposé est le Frontier AI Collaborative, qui vise à promouvoir et à diffuser l'accès à l'IA avancée dans les sociétés mal desservies. Enfin, un grand projet de sécurité en IA est envisagé pour réunir les meilleurs chercheurs et les meilleurs ingénieurs de premier plan afin d'étudier et d'atténuer les risques techniques liés à l'IA. Cf. l'article de Lewis Ho et al., 2023, « International Institutions for Advanced AI » : <http://arxiv.org/pdf/2307.04699>

#### **4. Annoncer un programme européen de coopération en IA, associant plusieurs pays dont au moins la France, l'Allemagne, les Pays-Bas, l'Italie et l'Espagne**

Il faut profiter du sommet pour annoncer le **lancement d'un grand programme européen de coopération en IA**. Une telle démarche, envisagée depuis 2017, n'a toujours pas connu de traductions concrètes. Cette initiative n'a pas nécessairement à réunir l'ensemble des 27 États membres, mais au moins la France, l'Allemagne, les Pays-Bas, l'Italie et l'Espagne.

Ces pays partagent une **vision assez proche de l'IA et de ses enjeux**, y compris l'adoption de l'IA dans le monde du travail et les objectifs d'aide aux PME et aux *start-up*. Il y a déjà eu des discussions sur l'IA entre la France et l'Allemagne ou entre la France et l'Italie, mais il reste à construire un programme de coopération marquant l'existence d'une **voie européenne de l'IA spécifique** allant plus loin que le soutien à l'innovation ou la régulation prévue par l'*AI Act*, une **approche basée sur l'éthique et la prise en compte des conséquences à court, moyen et long termes de ces technologies**. Pour vos rapporteurs, l'IA ouvre un espace d'opportunités qui peut nous aider à aller vers des sociétés solidaires soucieuses d'usages des technologies conformes aux droits de l'homme et aux valeurs humanistes. C'est un choix politique qui permettra d'éviter les seuls usages capitalistes, sécuritaires et oppressifs de ces technologies.

#### **5. Associer le Parlement à l'organisation du sommet**

Afin de garantir une plus grande légitimité du futur sommet, il est nécessaire d'**associer plus étroitement le Parlement à son organisation**.

La nomination d'**un sénateur** et d'**un député** au sein du comité de pilotage du sommet serait à cet égard un gage important de crédibilité, marquant la volonté de l'exécutif d'accroître le fondement démocratique de la réflexion française sur l'encadrement de l'IA à l'échelle internationale.

Vos rapporteurs et l'Office se tiennent prêts à répondre à toute sollicitation en ce sens.



## B. LES PROPOSITIONS VISANT À FONDER UNE VÉRITABLE POLITIQUE NATIONALE DE L'IA

### 1. Développer une filière française ou européenne autonome sur l'ensemble de la chaîne de valeur de l'intelligence artificielle

La première de nos propositions au niveau national est **un objectif qui doit tous nous mobiliser**, pouvoirs publics nationaux et locaux, décideurs économiques, associations et syndicats : nous devons viser le **développement d'une filière française ou européenne autonome sur l'ensemble de la chaîne de valeur de l'intelligence artificielle**, même sans chercher à rivaliser avec les puissances américaines et chinoises en la matière. En effet, **mieux vaut une bonne IA chez soi qu'une très bonne IA chez les autres**. Il s'agit de se protéger par la maîtrise de l'ensemble des couches de la technologie.

Que ce soit au niveau européen, par l'UE ou avec une coopération renforcée entre quelques pays, ou directement au niveau national, la France doit relever ce défi de **construire pour elle, en toute indépendance, les nombreux maillons de la chaîne de valeur de l'intelligence artificielle**. L'enjeu de la **souveraineté nationale** sur cette chaîne est crucial, c'est une condition à la fois de notre indépendance en général et d'une véritable **autonomie stratégique** sur ces technologies d'IA. Lorsqu'une maîtrise complète de la chaîne ne sera pas possible, notre pays pourra se tourner vers ses partenaires européens pour construire les coopérations nécessaires.

Un article récent de dirigeants de Mc Kinsey, publié le 31 octobre 2024 dans *Les Échos*<sup>1</sup>, se prononce également en faveur « *d'une stratégie holistique pour couvrir l'ensemble de la chaîne de valeur de la technologie* » en Europe, de l'énergie jusqu'aux applications en passant par les semi-conducteurs, les infrastructures et les modèles. L'Europe est aujourd'hui leader dans un seul des segments de la chaîne de valeur de l'IA, celui très précis de la **gravure des puces**, et ce grâce à ASML. Cet article, sans y faire référence explicitement, repose sur le constat dressé dans une note du Mc Kinsey Global Institute d'octobre 2024<sup>2</sup> analysant la place des pays européens dans chacun des maillons de la chaîne de valeur de l'IA.

---

<sup>1</sup> « IA : l'Europe peut encore jouer dans la cour des grands ! » dans *Les Échos* du 31 octobre 2024 : <https://www.lesechos.fr/idees-debats/cercle/opinion-ia-leurope-peut-encore-jouer-dans-la-cour-des-grands-2128995>

<sup>2</sup> La note s'intitule « *Time to place our bets: Europe's AI opportunity* » et vise à encourager la compétitivité de l'Europe sur la totalité de la chaîne de valeur de l'IA en recourant à une approche à plusieurs niveaux que le cabinet qualifie ainsi : « *A holistic approach to help Europe realize generative AI's full potential* ». Cf. Alexander Sukharevsky et al., octobre 2024, « *Time to place our bets: Europe's AI opportunity* », Mc Kinsey Global Institute : <https://www.mckinsey.com/capabilities/quantumblack/our-insights/time-to-place-our-bets-europes-ai-opportunity>

En dehors de la niche très spécifique d'ASML, les entreprises européennes restent dans la course en matière de **conceptions des modèles, d'applications d'IA et de services** mais elles ne représentent que moins de 5 % de parts de marché pour les matières premières (silicium en particulier), la conception des processeurs, la fabrication des puces, les infrastructures de calcul en nuage et les supercalculateurs. Ces filières sont à développer.

Outre l'augmentation des **investissements**, par exemple à travers les marchés publics, comme pour des applications d'IA dans les secteurs de la défense ou de la santé par exemple, Mc Kinsey invite à prendre pleinement position sur le marché des semi-conducteurs (*via* des technologies émergentes notamment), à **lutter contre la fuite des cerveaux** en assurant l'attraction des talents, et à **former davantage** aux métiers des différentes filières de l'IA, par des programmes de requalification de la main-d'œuvre pour la préparer à ces nouveaux défis.

Vos rapporteurs ont retenu de leurs comparaisons internationales qu'il faut commencer à **lutter réellement contre la fuite des cerveaux**. L'Inde et la Chine ont toutes deux compris l'importance pour leurs stratégies nationales en IA de retenir leurs talents et d'inciter au retour de leurs expatriés. En créant des écosystèmes compétitifs favorables à l'innovation, en offrant des opportunités de carrière attractives, en renforçant le sentiment national et surtout à travers divers dispositifs concrets, la France pourra suivre la voie ouverte par ces deux pays et limiter voire inverser la tendance à la fuite des cerveaux. Nous devons aller dans cette direction. **Notre pays n'est pas un centre de formation destiné à préparer les futurs génies de la Silicon Valley.**

## **2. Mettre en place une politique publique de l'IA avec des objectifs, des moyens et des outils de suivi et d'évaluation**

Plutôt que d'annoncer une **stratégie sans objectifs, sans gouvernance et sans outils de suivi**, visant pourtant à « *faire de la France un leader mondial de l'IA* », il convient de mettre en place **une véritable politique publique de l'IA** avec des **objectifs**, des **moyens réels** dont **une gouvernance digne de ce nom**, et, enfin, des **outils de suivi et d'évaluation**. Ces éléments sont aujourd'hui cruellement absents des politiques publiques menées en France en matière d'intelligence artificielle.

Plus largement, la **politique de la Start-up Nation avec son bras armé la French Tech**, aussi élitiste qu'inadaptée, est à **abandonner** au profit d'une **politique de souveraineté numérique**, cherchant à construire notre autonomie stratégique et à mailler les territoires.

### **3. Organiser le pilotage stratégique de la politique publique de l'intelligence artificielle au plus haut niveau**

La stratégie nationale pour l'IA ne dispose pas d'une gouvernance digne de ce nom et, comme l'ont montré les développements du présent rapport, le coordinateur national à l'intelligence artificielle ne représente qu'une **toute petite équipe rattachée à un service de Bercy**, la DGE. Il faudra, au moins, **mieux coordonner la politique publique nationale de l'intelligence artificielle** que nous appelons de nos vœux et lui donner une réelle **dimension interministérielle** avec un rattachement du coordinateur au Premier ministre.

La **nomination d'une secrétaire d'État à l'intelligence artificielle et au numérique va dans le bon sens**, mais il faut aller plus loin que ce premier pas symbolique et définir un pilotage stratégique de la politique publique de l'intelligence artificielle au plus haut niveau avec une coordination interministérielle.

### **4. Former les élèves de l'école à l'Université, former les actifs et former le grand public à l'IA**

Il est indispensable de **lancer de grands programmes de formation à destination des scolaires, des collégiens, des lycéens, des étudiants, des actifs et du grand public** à l'IA. De ce point de vue, les politiques conduites par la Finlande, qui ont été présentées de manière détaillée, sont des modèles à suivre. La démystification de l'IA est une première étape importante et nécessaire pour permettre une adhésion à son développement. Elle favorisera aussi la diffusion de la technologie dans la société et dans nos entreprises. Et des programmes de formation de haut niveau permettront par ailleurs une montée en compétence en IA en France, qui dispose déjà d'atouts importants en la matière.

Il faut aussi promouvoir **une vision scientifiquement éclairée et plutôt optimiste de l'intelligence artificielle**, telle que celle portée par Yann LeCun par exemple, qui est l'un de nos plus grands experts de ces technologies. L'IA générale reste pour l'heure une perspective peu probable et la question des risques existentiels éventuellement posés par ces technologies n'est pas une priorité, même si elle fait écho aux représentations catastrophistes que le grand public se fait souvent de l'IA en lien avec les récits de science-fiction et le cinéma. Vos rapporteurs, comme leurs prédécesseurs de 2017, **jugent indispensable de démystifier l'intelligence artificielle**.

## 5. Accompagner le déploiement de ces technologies dans le monde du travail et la société, notamment par la formation permanente

S'il est difficile de prévoir l'impact précis que l'IA aura sur le marché du travail, comme l'ont montré les développements à ce sujet dans le présent rapport, il faut tout de même **accompagner le déploiement de ces technologies**, notamment l'IA générative, dans le monde du travail, en particulier par des **programmes de formation permanente ambitieux**.

LaborIA s'intéresse aux aspects qualitatifs de ces transformations, mais il ne faut pas oublier les enjeux quantitatifs. C'est pourquoi il est recommandé de **mener régulièrement des études qualitatives et quantitatives sur l'impact de l'IA sur l'emploi, le tissu social (dont les inégalités) et les structures cognitives** en vue d'éclairer les pouvoirs publics et d'anticiper les mutations des pratiques professionnelles et les changements structurels dans les secteurs d'activité. Sur cette base, il sera possible d'**ajuster plus efficacement les programmes de formation permanente** et d'adapter les politiques publiques, par exemple en matière d'éducation, de recherche ou de soutien à l'innovation.

## 6. Lancer un grand dialogue social autour de l'intelligence artificielle et de ses enjeux

Le dialogue social par la négociation collective peut être renouvelé par **l'introduction de cycles de discussions tripartites autour de l'IA** et de ses nombreuses problématiques. Une opération d'envergure nationale, comme un **Grenelle de l'IA**, pourrait également être organisée.

Le dialogue social autour de l'IA devrait aussi **se décliner dans les entreprises avec les salariés, les responsables des systèmes d'information et les DRH** pour permettre une meilleure diffusion des outils technologiques et un rapport moins passionné à leurs conséquences. Comme le présent rapport l'a montré, c'est en effet une occasion de favoriser l'appropriation concrète et réaliste de la technologie et de ses enjeux, en se débarrassant des mythes entourant l'IA.

## 7. Mobiliser et animer l'écosystème français de l'IA

L'écosystème français de l'IA ne doit pas être mobilisé qu'à travers la *French Tech*, des *meet-up* à Station F et l'événement annuel VivaTech. Tous les acteurs de l'IA, **la recherche publique et privée, les grands déployeurs de systèmes** mais aussi **l'ensemble des filières économiques via des correspondants IA** (qui pourraient être des représentants des DSI par secteur) doivent pouvoir faire l'objet d'une grande mobilisation générale.

C'est un peu l'esprit qui avait régné entre décembre 2016 et mars 2017 lors de la préparation du plan France IA voulu par le Président de la République, alors François Hollande, un plan sans doute trop rapidement enterré au profit d'une stratégie de *Start-up Nation* dont les impasses et les lacunes apparaissent de plus en plus nettement avec le temps.

Il y manquait tout de même **une structure d'animation**, raison pour laquelle vos rapporteurs suggèrent de mobiliser l'écosystème français de l'IA autour de **pôles d'animation régionaux**, en relation étroite avec les universités, les centres de recherche, comme Inria, et les entreprises. Des expériences étrangères ayant concrétisé l'approche initiale de France IA peuvent nous inspirer. La structure **NL AI Coalition**, créée par le gouvernement néerlandais et rencontrée par vos rapporteurs à La Haye, rassemble ainsi depuis cinq ans **l'écosystème public et privé de l'IA aux Pays-Bas**, avec le concours du patronat, des universités et des grands centres de recherche. Elle s'appuie sur **sept centres régionaux** et est organisée en **18 groupes de travail thématiques**.

## **8. Reconduire le programme « Confiance.ai » ou mettre en place un projet équivalent**

Le programme « Confiance.ai » s'est interrompu en 2024 alors qu'il ne coûtait pas cher<sup>1</sup> et était efficace<sup>2</sup>. Il visait à **permettre aux industriels d'intégrer des systèmes d'IA de confiance dans leurs *process*** grâce à des méthodes et des outils intégrables dans tout projet d'ingénierie.

Pour ce faire, le programme levait les verrous associés à l'industrialisation de l'IA comme la construction de composants de confiance maîtrisés, la construction de données et/ou de connaissances pour augmenter la confiance dans l'apprentissage ou encore l'interaction générant de la confiance entre l'utilisateur et le système fondé lui-même sur l'IA de confiance. Il réunissait dans une logique partenariale de grands acteurs académiques et industriels français dans les domaines critiques de l'énergie, la défense, des transports et de l'industrie 4.0 comme l'illustre ce graphique.

---

<sup>1</sup> Son budget était de 45 millions d'euros sur quatre ans dont une dotation de l'État de 15 millions d'euros, soit environ 3,75 millions d'euros par an.

<sup>2</sup> Cf. ses actions sur le site de Confiance.ai : <https://www.confiance.ai/>

## Les partenaires du programme Confiance.ai



Source : Confiance.ai

Le programme visait la construction d'une plateforme sûre, fiable et sécurisée d'outils logiciels, qui soit à la fois souveraine, ouverte, interopérable et pérenne dans les secteurs des produits et services critiques les plus concernés, mutualisant les savoir-faire scientifiques et technologiques et contribuant au cadre technique du règlement européen sur l'intelligence artificielle. Dans ce contexte, il est demandé au gouvernement et, en particulier, à la DGE de **reconduire le programme « Confiance.ai »** ou de **mettre très rapidement en place un projet équivalent** capable d'allier des expertises en termes de technologie, d'évaluation et de normalisation de l'intelligence artificielle.

Son **coût d'environ 3,75 millions d'euros par an** est à mettre en regard des 37 milliards d'euros de dépenses proposés par la commission sur l'intelligence artificielle en 2024. Le programme « Confiance.ai » n'en représentait que 1 %, or au vu de ses actions et de ses résultats, il justifie la **poursuite d'un effort budgétaire raisonnable et pourtant rentable**.

## 9. Soutenir la recherche publique en intelligence artificielle selon des critères de transversalité et de diversification des technologies

La recherche privée en intelligence artificielle a pris beaucoup d'avance sur la recherche publique, mais cette dernière doit revenir dans la course. **La soutenir davantage est un impératif.** L'Office juge pertinent de l'orienter vers des **activités transdisciplinaires et, plus globalement, transversales autour de « projets de recherche » en IA.**

La préoccupation à l'égard de la **diversification des technologies** est aussi fondamentale : les avancées en IA se font par la **combinaison** et la **recomposition** de savoirs et de savoir-faire, pas par l'enfermement dans un modèle unique que l'on chercherait à perfectionner. L'**IA générative**, à travers son modèle Transformer, **ne doit pas devenir la priorité du monde de la recherche**, en dépit de l'effet de mode autour de ChatGPT. L'IA symbolique, par exemple, ne doit pas être totalement abandonnée, elle peut s'hybrider avec les IA connexionnistes pour forger de **nouvelles approches logiques**, imbriquant le signifiant et le signifié et, partant, plus proches des raisonnements humains. D'autres technologies permettant d'apporter plus de logique aux systèmes d'IA générative peuvent également inspirer de nouvelles perspectives pour la recherche, comme les modèles « *Mixture of Experts* » (MoE), les arbres de pensées ou *Trees of Thoughts* (ToT) et la génération augmentée de récupération ou *Retrieval Augmented Generation* (RAG). La prise en compte de **modèles de représentation du monde** (« *World Models* ») dans de nouvelles « architectures cognitives » est un autre défi que la recherche devra relever, permettant aux IA de prendre en compte la réalité spatio-temporelle, y compris le monde physique et ses lois. Ce domaine avance lentement mais pourrait connaître une accélération des découvertes<sup>1</sup>.

Pour paraphraser Rabelais qui écrivait que « *Science sans conscience n'est que ruine de l'âme* », vos rapporteurs affirment que « *l'IA sans logique n'est qu'illusion d'intelligence* ».

---

<sup>1</sup> Joshua Bengio, Yann LeCun et d'autres chercheurs travaillent à cette amélioration qualitative. En novembre 2024, un article a ainsi proposé une nouvelle méthode permettant aux IA génératives de modéliser les dynamiques visuelles sans passer par une étape de reconstruction du monde physique dans un espace pixellisé : le Dino-WM pour World Model. Cf. Gaoyue Zhou, Hengkai Pan, Yann LeCun, Lerrel Pinto, 2024, « DINO-WM : World Models on Pre-trained Visual Features enable Zero-shot Planning » : <https://arxiv.org/abs/2411.04983>



## **10. Relever le défi de la normalisation en matière d'intelligence artificielle**

Il faut permettre à la France de **défendre au mieux l'intérêt national** ainsi que **les intérêts de nos entreprises nationales** en matière de normalisation de l'IA, ce qui implique de **mobiliser davantage l'Afnor et surtout le Cofrac**, aujourd'hui désinvesti sur ce sujet.

Il faudrait aussi **maintenir ou augmenter le financement de nos organismes de normalisation** qui remplissent leurs missions dans des conditions parfois difficiles.

Enfin, la France doit **inviter ses partenaires européens à faire preuve d'une plus grande vigilance** dans le choix de leurs représentants dans les comités chargés de la normalisation en IA : **s'appuyer sur des experts issus d'entreprises extra-européennes**, le plus souvent américaines ou chinoises, **n'est pas acceptable**.

## **11. S'assurer du contrôle souverain des données issues de la culture française et des cultures francophones et créer des jeux de données autour des cultures francophones**

Il faut à la fois s'assurer du **contrôle souverain des données issues de la culture française**, notamment des archives détenues par la BNF ou l'Institut national de l'audiovisuel, voire des cultures francophones et créer des *datasets* autour des cultures francophones, en vue d'alimenter l'entraînement de modèles d'IA reflétant notre environnement linguistique et culturel.

Il s'agit d'un **acte de résistance face à la domination linguistique et culturelle anglo-saxonne, en particulier américaine**, qui caractérise l'IA aujourd'hui et qui fait courir un risque grave d'uniformisation culturelle et d'appauvrissement linguistique. Les jeux de données autour des cultures francophones pourront être **constitués avec le concours de tous les pays de la francophonie**.

Les initiatives conduites par certains pays, en particulier par l'Espagne, peuvent contribuer à inspirer notre pays et à aider à la définition du cadre de ces *datasets*. En effet, le gouvernement espagnol a lancé depuis 2022 un plan national autour de la « nouvelle économie de la langue ». Il vise à **placer l'espagnol au cœur de la transformation numérique et de la promotion de la chaîne de valeur de la nouvelle économie de la connaissance et de l'intelligence artificielle**. La France doit **faire la même chose pour le français**. Ainsi, le plan de l'Espagne se décline notamment en constitution de bases de données textuelles dans les langues espagnoles, ces *datasets*, permettant **l'entraînement de modèles et la création de LLM basés sur la langue et la culture espagnoles**.

Comme l'Espagne, qui a réservé une part de son plan à ses langues « co-officielles » (catalan, basque, galicien) à côté de l'espagnol, la France pourrait avantageusement consacrer une part de ces jeux de données autour des cultures francophones aux **langues régionales** (breton, occitan, basque, corse, langues pratiquées dans les territoires ultramarins...) dont l'appartenance au patrimoine national a été consacrée par la Constitution depuis 2008.

## **12. Préparer une réforme des droits de propriété intellectuelle dont le droit d'auteur pour les adapter aux usages de l'IA générative**

Il faut poser les bases d'une **réforme des droits de propriété intellectuelle et du droit d'auteur** pour les adapter aux usages de l'IA générative et aux problèmes posés plus généralement par l'utilisation de l'intelligence artificielle. Notre législation en matière de propriété intellectuelle et de droits d'auteur applicables à l'IA générative nécessite à l'évidence plusieurs éclaircissements. Le rapport d'information de la commission des lois de l'Assemblée nationale déposé en conclusion des travaux de sa mission d'information sur les défis de l'intelligence artificielle générative en matière de protection des données personnelles et d'utilisation du contenu généré s'est consacré à ce travail de clarification. Les 33 recommandations de nos collègues députés, que nous ne récapitulerons pas ici, fournissent des pistes intéressantes qu'il convient d'évaluer et de soumettre à la délibération<sup>1</sup>.

L'objectif d'une telle réforme sera à la fois de **clarifier les régimes juridiques applicables**, de **protéger les ayants droit des données ayant servi à l'entraînement des modèles** mais aussi les **créateurs d'œuvres nouvelles grâce à l'IA**. Cela impliquera donc, d'une part, de trancher l'épineuse question de **l'équilibre entre les intérêts des ayants droit et intérêt des entreprises développant des modèles d'intelligence artificielle**, d'autre part, de lever le doute sur la **frontière entre œuvre originale et contrefaçon ou copie** dans le cas d'un travail créé par IA puisque la réforme devra aussi mettre fin au flou juridique entourant le statut des œuvres créées par l'intelligence artificielle.

La solution des **droits voisins**, utilisée dans les secteurs de la musique et du cinéma, mais aussi de l'information (dont la presse écrite) à la suite des transformations liées au numérique et à Internet, est une perspective dont il faut débattre.

---

<sup>1</sup> Rapport d'information de la commission des lois de l'Assemblée nationale déposé en conclusion des travaux de sa mission d'information sur les défis de l'intelligence artificielle générative en matière de protection des données personnelles et d'utilisation du contenu généré de Philippe Pradal et Stéphane Rambaud, rapport d'information n° 2207, 16<sup>e</sup> législature : [https://www.assemblee-nationale.fr/dyn/16/rapports/cion\\_lois/116b2207\\_rapport-information](https://www.assemblee-nationale.fr/dyn/16/rapports/cion_lois/116b2207_rapport-information)

Face au succès des plateformes de *streaming* et de service de vidéo à la demande, l'enjeu de ces **compensations financières** devient grandissant. Il le sera encore plus avec la diffusion de l'IA. L'UE s'est d'ailleurs dotée en 2019 d'une directive sur le droit d'auteur et les droits voisins dans le marché unique numérique<sup>1</sup>, qu'il faudra faire évoluer au rythme des capacités de création de l'IA.

Techniquement, l'utilisation de données propriétaires et surtout de contenus protégés par le droit d'auteur peut non seulement être **limitée en amont par le nettoyage des données** mais dorénavant **corrigée en aval grâce à une nouvelle technologie** introduite récemment appelée *Model disgorgement* ou *Machine Unlearning*. Les entreprises qui collecteraient illégalement des données pour les utiliser à fin d'entraînement de leurs modèles pourraient par exemple être dans l'obligation non seulement de supprimer les données problématiques et de les abandonner lors de futurs entraînements mais surtout de **mettre à jour leurs modèles en faisant comme si ces données n'avaient jamais été utilisées**. Plutôt que de réentraîner totalement leurs modèles, les entreprises s'appuieraient alors utilement sur cette technique émergente qui pourra encore être perfectionnée<sup>2</sup>.

### **13. Confier à l'OPECST le suivi et l'évaluation régulière de la politique publique conduite par le Gouvernement en la matière**

La politique nationale en matière d'IA conduite par le Gouvernement devrait faire l'objet d'un **suivi** et d'une **évaluation régulière** par l'OPECST. Les aspects scientifiques et technologiques de l'intelligence artificielle ainsi que les enjeux qu'ils soulèvent appellent **une expertise et une vigilance à la croisée des mondes politiques et scientifiques**, c'est donc logiquement à l'OPECST qu'il convient de faire appel.

Dans le rapport précité de la commission des lois de l'Assemblée nationale, nos collègues députés - ayant constaté que le premier travail parlementaire relatif à l'IA avait été fait dans le rapport de l'OPECST de mars 2017 « Pour une intelligence artificielle maîtrisée, utile et démystifiée », qualifié « *d'étude approfondie* » - préconisent dans leur 33<sup>e</sup> recommandation de **confier à l'OPECST un suivi permanent des questions relatives à l'intelligence artificielle**.

---

<sup>1</sup> Cette directive, qui visait notamment à protéger le monde de la presse, impose par exemple aux plateformes numériques telles que Google ou Facebook, de rémunérer les éditeurs et les journaux lorsqu'ils utilisent leurs contenus. Cf. la directive du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique : <https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32019L0790&from=EN>

<sup>2</sup> Cf. Brandon LaLonde, 2023, « Explaining model disgorgement », IAPP : <https://iapp.org/news/a/explaining-model-disgorgement> et Joshua A. Goland, 2023, « Algorithmic Disgorgement : Destruction of Artificial Intelligence Models as the FTC's Newest Enforcement Tool for Bad Data », *Richmond Journal of Law and Technologies*, volume XXIX, n° 2 : <https://ssrn.com/abstract=4382254>

Les rapporteurs Philippe Pradal et Stéphane Rambaud estiment en effet « *souhaitable que le Parlement puisse réaliser un suivi permanent des questions relatives à l'intelligence artificielle, dont les implications sont multiples et évolutives. L'Office parlementaire d'évaluation des choix scientifiques et technologiques semble le mieux à même d'assumer cette responsabilité, d'autant qu'il a déjà travaillé sur ce sujet* ».

**L'actualisation des connaissances sur l'IA et ses enjeux par le présent rapport plaide encore plus en ce sens aujourd'hui.**

## LISTE DES PERSONNES ENTENDUES

### I. INSTITUTIONS

#### **Conseil national du numérique (CNNum)**

- Gilles Babinet, co-président, entrepreneur
- Joëlle Toledano, co-présidente, professeure des universités
- Jean Cattan, secrétaire général
- Agathe Bougon, secrétaire générale adjointe, chargée de la communication
- Anne Alombert, philosophe, enseignante-chercheuse à l'université Paris 8 Vincennes à Saint-Denis
- Olga Kokshagina, chercheuse en management

#### **Commission nationale de l'informatique et des libertés (Cnil)**

- Bertrand Pailhès, directeur des technologies et de l'innovation, ancien coordinateur national pour l'intelligence artificielle
- Félicien Vallet, chef du service Intelligence artificielle
- Chirine Berrichi, conseillère parlementaire

#### **Direction générale des entreprises**

- Loïc Duflot, chef du service de l'économie numérique
- Marie-Léa Rols, chargée de mission pour l'intelligence artificielle

**Guillaume Avrin**, coordinateur national pour l'intelligence artificielle

#### **Commission de l'intelligence artificielle**

- Philippe Aghion, co-président, professeur au Collège de France, à l'Insead et à la London School of Economics
- Anne Bouverot, co-présidente, présidente de l'École normale supérieure
- Cyprien Canivenc, rapporteur général
- Arno Amabile, rapporteur général

#### **Organisation de coopération et de développement économiques**

- Nozha Boujemaa, co-présidente du groupe d'experts sur l'IA de confiance à l'OCDE et membre du groupement d'experts de haut niveau sur l'IA auprès de la Commission européenne
- Angéline Gentaz, chargée de mission sur les politiques publiques de l'IA

### **Association française de normalisation (Afnor)**

- Patrick Bezombes, conseiller pour la stratégie et la gouvernance de l'IA, représentant de la France au CEN-Cenelec pour la normalisation européenne de l'IA

## **II. ANCIENS PARLEMENTAIRES**

**Cédric Villani**, ancien député, ancien président de l'OPECST, mathématicien, professeur des universités, membre de l'Académie des sciences, auteur du rapport « Donner un sens à l'intelligence artificielle » en 2018

**Claude de Ganay**, ancien député, ancien membre de l'OPECST, auteur du rapport « Pour une intelligence artificielle maîtrisée, utile et démystifiée » de l'OPECST en 2017

## **III. EXPERTS ET SCIENTIFIQUES**

**Raja Chatila**, membre du conseil scientifique de l'OPECST, professeur émérite à l'université Paris Pierre-et-Marie-Curie (UPMC), directeur de recherche au Centre national de la recherche scientifique (CNRS), membre du comité national pilote d'éthique du numérique (CNPEN), président de l'Initiative mondiale IEEE sur l'éthique des systèmes autonomes et intelligents, co-président du groupe « IA responsable » du *Global Partnership on Artificial Intelligence* (GPAI), membre du groupe d'experts européen sur l'intelligence artificielle, ancien directeur de l'Institut des systèmes intelligents et de robotique (ISIR)

**Daniel Andler**, membre du Conseil scientifique de l'OPECST, professeur d'université, membre de l'Académie des sciences morales et politiques

### **Institut national de recherche en sciences et technologies du numérique (Inria)**

- Bruno Sportisse, président-directeur général
- Frédéric Alexandre, directeur de recherche à l'Inria, responsable de l'équipe MNEMOSYNE à l'Institut des maladies neurodégénératives (CNRS-Inria-Université de Bordeaux-CHU Pellegrin)
- Yann Ferguson, sociologue, directeur scientifique de LaborIA
- Sandrine Mazetier, directrice générale déléguée à l'appui aux politiques publiques

### **Commissariat à l'énergie atomique et aux énergies alternatives (CEA)**

- François Terrier, vice-président IA, directeur des programmes du List (systèmes numériques)
- Jérôme Bobin, directeur du département d'électronique et informatique pour la physique (Dedip) et du programme NumPEX (logiciels pour supercalculateurs)
- Hélène Latour, chargée de mission Affaires publiques et institutionnelles

**Serge Abiteboul**, directeur de recherche à l'Institut national de recherche en sciences et technologies du numérique (Inria), professeur affilié à l'ENS Paris Saclay, membre de l'Académie des sciences et du Collège de l'Autorité de régulation des communications électroniques, des postes et de la distribution de la presse (Arcep), président du conseil scientifique de la DGFIP, ancien professeur au Collège de France (Chaire d'informatique), ancien membre du Conseil national du numérique, ancien président du conseil scientifique de la Société d'informatique de France, animateur du blog binaire

**Bertrand Braunschweig**, coordinateur scientifique du programme national de recherche sur l'IA de confiance « Confiance.Ai », ancien coordinateur de la stratégie nationale de recherche en IA, ancien directeur du numérique à l'Agence nationale de la recherche (ANR), ancien directeur des centres Inria de Saclay et de Rennes

**Jean-Gabriel Ganascia**, professeur émérite à l'université Paris Pierre-et-Marie-Curie (UPMC), membre de l'Institut universitaire de France, président de l'association française pour l'avancement des sciences (Afas) et des comités d'éthique de Pôle emploi et de Docaposte, ancien président du Comets (Comité d'éthique du CNRS), auteur d'ouvrages sur l'IA destinés au grand public

**Laurence Devillers**, professeure à l'université Paris-Sorbonne et directrice de recherche au CNRS, Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (Limsi de Saclay)

**Laurent Alexandre**, entrepreneur, conférencier, essayiste, membre de l'Académie des technologies

**Christophe Labarde**, journaliste, organisateur de l'exposition *Irruption – Quand l'intelligence artificielle bouleverse la création*

**Pierre Fautrel**, cofondateur du collectif d'artistes Obvious



#### **IV. ENTREPRISES ET ASSOCIATIONS**

##### **Mistral AI**

- Audrey Herblin-Stoop, directrice Affaires publiques
- Marie Pellat, ingénieure, spécialiste de la sécurité de l'IA
- Lucile Saulnier, ingénieure, chercheuse en IA

##### **Atos**

- Vasco Gomes, responsable technologique Cybersécurité

##### **Dataiku**

- Florian Douetteau, co-fondateur et président

##### **LightOn**

- Igor Carron, co-fondateur et président
- Olga Lopusanschi, responsable clients

##### **Preligens**

- Tugdual Ceillier, chercheur expert en IA
- Marc Antoine, responsable Affaires publiques

##### **XXII**

- William Eldin, fondateur et président
- François Mattens, directeur des affaires publiques et des partenariats stratégiques

##### **Microsoft France**

- Philippe Limantour, directeur technologique et cybersécurité
- Lionel Benatia, directeur des affaires publiques

##### **Google France**

- Jean-Marie Boutin, directeur des relations institutionnelles
- Sarah Boiteux, membre de l'équipe Government affairs and public policy

##### **Meta France**

- Martin Signoux, responsable Affaires publiques

### **IBM France**

- Vincent Perrin, directeur technique écosystème, IA et informatique quantique
- Diane Dufoix-Garnier, directrice Affaires publiques
- Thierno-Babacar Ndiaye, alternant au sein de la cellule Affaires publiques

### **Kaspersky**

- Bertrand Trastour, directeur général France
- Amin Hasbini, directeur du centre de recherche Moyen-Orient/ Turquie/ Afrique
- Vincent Prevost, consultant
- Gladys-Anne Okatakyie Yiadom, responsable des affaires publiques

### **SaferAI**

- Siméon Campos, fondateur et président

### **Cybernetica**

- Tariq Krim, fondateur et président

## **V. DÉPLACEMENT À BRUXELLES**

*Mercredi 24 et jeudi 25 avril 2024*

### **Commission européenne**

- Mauritz-Jan Prinz, conseiller au cabinet de M. Thierry Breton, Commissaire européen chargé du marché intérieur, accompagné par des représentants de la DG Connect de la Commission européenne

### **Parlement européen**

- Svetla Tanova, responsable du European Science Media Hub
- Tambiama Madiaga, analyste Politiques numériques au service de recherche du Parlement européen (STOA)

### **Ambassade de France en Belgique**

- François Sénémaud, ambassadeur de France en Belgique
- Séverine Fautrelle, conseillère de coopération et d'action culturelle

### **Représentation Permanente de la France**

- Benoît Blary, conseiller

**Hugues Bersini**, professeur à l'Université libre de Bruxelles, membre de l'Académie royale de Belgique

**Institut Fari (Institut d'intelligence artificielle de la région Bruxelles-capitale)**

- Carl Mörch, directeur

## **VI. DÉPLACEMENT AUX PAYS-BAS**

*Jeudi 25 et vendredi 26 avril 2024*

### **Europol**

- Jean-Philippe Lecouffe, directeur adjoint
- Didier Jacobs, directeur de l'intelligence artificielle
- Emmanuel Kessler, responsable de l'équipe Prévention et sensibilisation du Centre européen de lutte contre la cybercriminalité (EC3)

### **Ministère de l'intérieur (BZK)**

Eva Heijblom, directrice générale Transition numérique des services de l'État

Mark Pryce, responsable Affaires internationales

### **Ministère des affaires étrangères (BZ)**

- Joost Flamand, directeur de la politique de sécurité (DVB)

### **Ministère des affaires économiques et du climat (EZK)**

- Thomas Faber, responsable au sein du département de l'économie numérique

### **Ministère de l'éducation, de la culture et de la science (OCW)**

- Anne-Marie De Rooter, directrice Europe

### **Autorité des données personnelles**

- Sven Stevenson, directeur de programme à la direction des algorithmes de coordination (DCA)

### **Organisation néerlandaise pour la recherche scientifique appliquée (TNO)**

- Christine Balch, responsable Affaires européennes, stratégie d'entreprise
- Anne Fleur van Veenstra, directrice scientifique au Centre d'innovation et de stratégie sociétale (TNO Vector) et professeur chargé de la gouvernance des données et des algorithmes à l'Université de Leyde
- Freek Bomhof, directeur du programme Applications de l'IA

### **AI Amsterdam Coalition**

- Anita Nijboer, ambassadrice

### **Université de technologie de Delft**

- Han La Poutré, professeur, responsable Recherche et stratégie au CWI (Institut national de recherche en mathématiques et en informatique)
- Alessandro Bozzon, professeur
- Fred Herrebout, responsable Affaires publiques

### **Centre d'innovation pour l'intelligence artificielle (Innovation Center for Artificial Intelligence - ICAI)/Université d'Amsterdam**

- Maarten de Rijke, directeur

### **Confédération de l'industrie et des employeurs des Pays-Bas (VNO-NCW)**

- Stefan Leijnen, directeur IA pour les politiques internationales et européennes

### **TechLeap**

- Maxime Lubbers, responsable Intelligence artificielle
- Abba Chaoui, expert Data et IA

### **Fabien Bouhier, entrepreneur et investisseur**

### **Ambassade de France aux Pays-Bas**

- François Alabrune, ambassadeur de France aux Pays-Bas
- Jean-Jacques Pierrat, attaché de coopération scientifique et universitaire
- Marnix Satter, attaché de coopération scientifique et universitaire adjoint
- Manola Ruff, attachée Industrie et innovation
  - *Service économique régional de La Haye*
- Benoît Lemonnier, conseiller financier régional, adjoint à la cheffe du service

### **CCI France/Pays-Bas**

- Sylvie Craenen, directrice générale

### **French Tech Amsterdam**

- Mylena Pierremont, présidente

## **VII. DÉPLACEMENT AU ROYAUME-UNI**

*Du mercredi 21 au vendredi 23 février 2024*

### **Chambre des Lords**

- Timothy Clement-Jones, membre de la Chambre des Lords, porte-parole pour l'Économie numérique

### **AI Safety Institute, ministère de la science, de l'innovation et de la technologie**

- Imogen Schon, directrice adjointe
- Anna Bruvere, responsable pour la coopération politique internationale

### **Innovate UK**

- Sarah El-Hanfey, responsable IA et *Machine learning*

### **Ada Lovelace Institute**

- Fran Bennett, directrice générale par intérim
- Michael Birtwistle, directeur Droit et politiques publiques

### **Université d'Oxford**

- Lionel Tarassenko, professeur de bioingénierie, intelligence artificielle et santé

### **Google DeepMind**

- Ankur Vora, responsable Politique internationale et engagement public
- Dorothy Chou, responsable Affaires publiques
- Jean-Marie Boutin, responsable Affaires publiques Google France

### **BBC**

- Danijela Horak, responsable de la recherche en intelligence artificielle

### **KPMG**

- Dennis Tatarkov, économiste

### **Tony Blair Institute for global change**

- Melanie Garson, responsable Politiques et cybersécurité, professeur de résolution des conflits internationaux et de sécurité internationale au département de science politique de University College London (UCL)

### **Ambassade de France au Royaume-Uni**

- Hélène Duchêne, Ambassadrice de France au Royaume-Uni
  - *Service pour l'enseignement supérieur, la recherche et l'innovation (ESRI)*
- Minh-Hà Pham, conseillère Science et technologie
- Ludovic Drouin, attaché scientifique
- Yanis Gillmann, attaché adjoint Science et innovation
  - *Service économique régional (SER)*
- Karine Maillard, conseillère économique
- Lara Joutard, attachée numérique

## **VIII. DÉPLACEMENT AUX ÉTATS-UNIS (NEW YORK, WASHINGTON D.C., SAN FRANCISCO)**

*Du dimanche 14 au dimanche 21 avril 2024*

### **Congrès des États-Unis**

- Anna Eshoo, représentante (Démocrate-Californie)
- Todd Young, sénateur (Républicain-Indiana)
- Marsha Blackburn, sénatrice (Républicain-Tennessee)
- Mike Rounds, sénateur (Républicain-Dakota du Sud)

### **Office of science and technology policy (OSTP) de la Maison-Blanche**

- Austin Bonner, directrice générale adjointe Technologies et politiques publiques
- Alex Engler, conseiller politique

### **ONU et Unesco**

- Estelle Zadra, officier de liaison entre l'ONU et l'Unesco pour l'Europe et l'Amérique du Nord

### **Bureau de l'UE à San Francisco**

- Joanna Smolinska, directrice adjointe en charge du numérique

### **National Science Foundation (NSF)**

- Michael Littman, directeur de la division de l'informatique et des systèmes intelligents
- Wendy Nilsen, directrice adjointe de la division de l'informatique et des systèmes intelligents
- Katie Antypas, directrice du bureau Cyberinfrastructures avancées
- Dilma da Silva, directrice adjointe pour la science et l'ingénierie de l'informatique
- Roxanne Nikolaus, responsable de section, Bureau de la science et de l'ingénierie internationales
- Sirin Tekinay, responsable des relations entre la recherche américaine et le G7, le G20 et l'Europe

### **Université de Stanford**

- Emmanuel Candès, co-président de l'Institut de la science des données
- Florence G'sell, professeure invitée Programme Gouvernance des technologies émergentes
- Michelle Pualuan, responsable des programmes
- Marietje Schaake, chargée de mission Politique internationale

### **Université de Californie Berkeley - CITRIS (*Center for information technology research in the interest of society*)**

- Alexandre Bayen, directeur
- Camille Crittenden, directrice exécutive
- Sandrine Dudoit, doyenne adjointe de la faculté et de la recherche
- Brandie Nonnecke, directeur du laboratoire pour les politiques publiques, les risques et la sécurité (CITRIS Policy Lab)

### **Université du Maryland**

- Michel Cukier, professeur

### **EveryoneAI**

- Mathilde Cerioli, responsable scientifique
- Anne-Sophie Seret, chef de projet
- Céline Malvoisin, experte en développement du langage

### **Allen Institute**

- Rui Costa, directeur général
- Andy Hickl, directeur technique



### **Meta**

- Yann LeCun, directeur scientifique de l'IA, professeur à New York University
- Nicolas Bièvre, ingénieur en IA

### **Microsoft GitHub**

- Mike Linksvayer, vice-président en charge des politiques publiques
- Peter Cihon, directeur des politiques publiques

### **Google**

- Matthieu Lorrain, chef global de l'innovation créative
- Surya Tubach, cheffe de projet, Google Arts & Culture
- Vincent Vanhoucke, directeur en robotique, Google DeepMind
- Inès Mezerreg, ingénieure en intelligence artificielle, Google Research
- Shervine Amidi, ingénieur

### **OpenAI**

- Elisabeth Proehl, responsable de l'équipe Affaires internationales
- Karen Toro, membre de l'équipe Affaires internationales

### **Anthropic**

- Cathy Dinas, cheffe de cabinet
- Alex Tamkin, chef de projet
- Ashley Zlatinov, responsable des politiques publiques

### **Hugging Face**

- Yacin Jernite, chercheur en IA

### **Nvidia**

- Ned Finkle, vice-président en charge des relations extérieures
- Matt Milner, responsable des affaires publiques

### **Amazon Web Service (AWS)**

- Christian Troncoso, directeur de la politique globale IA
- Sasha Rubel, responsable des politiques publiques en IA/ *machine learning* Europe-Afrique-Moyen-Orient
- Sherry Markus, directrice scientifique d'Amazon Bedrock
- Charlotte Baylac, responsable des affaires publiques d'AWS France

### **Asana**

- Rebecca HINDS, responsable de « The Work Innovation Lab »

### **J&P Morgan Chase & Co**

- Manuela Veloso, responsable de la recherche en IA

### **XBrain**

- Gregory Renard, fondateur et président

### **LaVague**

- Daniel Huynh, président-directeur général

### **Prometheus Computing**

- Frédéric de Vaultx, directeur général

### **Asurion/Simplr**

- Clément Ruin, senior data scientist

### **Appen**

- Alice Desthuilliers, responsable produits

### **Mobioos**

- Nolwenn Godard, vice-présidente

### **Virgile Foussereau, chercheur, Université de Berkeley**

### **Ambassade de France aux États-Unis (Washington D.C.)**

- Laurent Bili, ambassadeur de France aux États-Unis
- Damien Cristofari, premier conseiller
- Alexandre Mirlesse, conseiller politique Chancellerie diplomatique
- Mireille Guyader, conseillère scientifique
- Marc Brocheton, conseiller économique, Entreprise et innovation (Service économique régional)
- Pascal Confavreux, conseiller Presse et communication
- Sébastien Pauly, attaché de défense adjoint - Armement
- Pascal Revel, attaché pour la science et technologie
- Ludovic Francis, attaché scientifique adjoint
- Vincent Lépinay, attaché de coopération universitaire

- A. Smith, attachée de presse
- V. d'Alançon, chargée de mission

**Consulat général de France à San Francisco**

- Frédéric Jung, consul général de France
- Vanessa Bonnet, attachée économique
- Emmanuelle Pauliac-Vaujour, attachée scientifique
- Valentine Asseman, attachée scientifique adjointe

**Commissariat à l'énergie atomique et aux énergies alternatives (CEA)**

- Sunil Félix, conseiller nucléaire États-Unis/Canada

**Centre national de la recherche scientifique (CNRS)**

- Sylvie Tourmente, directrice du bureau États-Unis et Canada

**Centre national d'études spatiales (CNES)**

- Nicolas Maubert, conseiller spatial pour les États-Unis

**Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (Inrae)**

- Laura Gonçalves de Souza, chargée de mission pour les États-Unis



**ANNEXES**  
**LETTRE DE SAISINE DE L'OFFICE**  
**PAR LE BUREAU DE L'ASSEMBLÉE NATIONALE**

RÉPUBLIQUE FRANÇAISE  
LIBERTÉ • ÉGALITÉ • FRATERNITÉ

ASSEMBLÉE NATIONALE

LA PRÉSIDENTE

PARIS, LE - 3 JUIL. 2023

Monsieur le Président et cher collègue,

Lors de sa réunion du 28 juin 2023, le Bureau de l'Assemblée nationale a examiné la demande que vous avez formulée, conjointement avec M. Gérard Longuet, par un courrier du 25 mai 2023, afin que l'Assemblée nationale sollicite de l'Office parlementaire d'évaluation des choix scientifiques et technologiques une étude sur les nouveaux développements de l'intelligence artificielle.

J'ai le plaisir de vous indiquer que, conformément à l'article 6 *ter* de l'ordonnance du 17 novembre 1958 relative au fonctionnement des assemblées parlementaires, le Bureau a fait droit à cette demande.

Je vous prie de croire, Monsieur le Président et cher collègue, à l'assurance de ma considération distinguée.

*Amitié,*



Yaël BRAUN-PIVET

Monsieur Pierre HENRIET  
Président de l'Office parlementaire d'évaluation  
des choix scientifiques et technologiques



## LETTRE DE SAISINE DE L'OFFICE PAR LE BUREAU DU SÉNAT

R É P U B L I Q U E F R A N Ç A I S E



Monsieur Gérard LONGUET  
Premier Vice-président de l'Office parlementaire  
d'évaluation des choix scientifiques et  
technologiques (OPECST)  
Sénateur de la Meuse  
Palais du Luxembourg

LE PRÉSIDENT

Paris, le 10 JUL. 2023

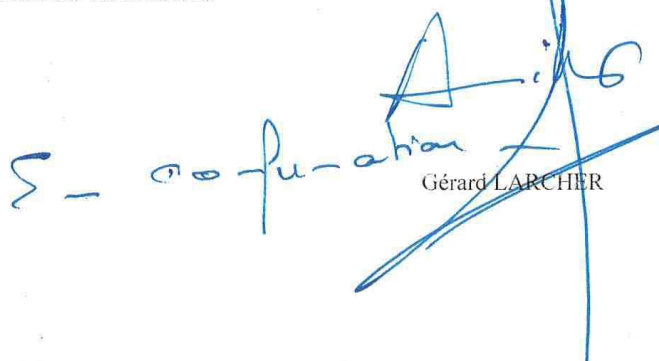
Monsieur le Président,



Par lettre en date du 25 mai dernier, vous m'avez fait part, conjointement avec M. Pierre HENRIET, Président de l'Office parlementaire d'évaluation des choix scientifiques et technologiques (OPECST), de votre souhait qu'à l'occasion du 40<sup>ème</sup> anniversaire de la création de l'Office, les Bureaux des deux assemblées décident l'un et l'autre de saisir l'Office sur un même sujet scientifique majeur d'actualité, à savoir les nouveaux développements en intelligence artificielle.

J'ai le plaisir de vous informer qu'au cours de sa réunion du 5 juillet 2023, le Bureau du Sénat, auquel j'ai soumis cette suggestion, a décidé, conformément à l'article 6<sup>ter</sup>, paragraphe V, de l'ordonnance du 17 novembre 1958 relative au fonctionnement des assemblées parlementaires, de saisir l'OPECST sur ce sujet.

Je vous prie de croire, Monsieur le Président, à l'assurance de mes sentiments les meilleurs.



Gérard LARCHER

*Bureau*