ÉTUDE DE RECHERCHE — FÉVRIER 2026

AI FOR AMERICANS FIRST

Protectionnisme IA, Énergie et Semi-conducteurs : Trajectoires de Divergence US/Europe 2024–2030

Analyse Géostratégique et Économique Intégrée

Chapitre I

Fabrice Pizzi

Université Sorbonne

Master Intelligence Économique — Intelligence Warfare

75% compute IA mondial = USA | **$675B** capex US 2026 | **7-12×** ratio US/EU

Paris — Février 2026

7 chapitres • 4 scénarios prospectifs • 3 zones géographiques

Mots-clés : *intelligence artificielle, protectionnisme technologique, semi-conducteurs, export controls, compute souverain, géopolitique de l'IA, France, États-Unis, Chine*

CHAPITRE II

Méthodologie

2.1 Approche générale : analyse prospective multi-scénarios à méthode mixte

2.1 Approche générale : analyse prospective multi-scénarios à méthode mixte

La présente étude combine une analyse empirique rétrospective (diagnostic 2020–2026) et une projection prospective par scénarios (2026–2030). Cette architecture en deux volets répond à la nature du phénomène étudié : le protectionnisme technologique IA est à la fois un fait observable (les export controls, les tarifs Section 232) et un processus en cours dont la trajectoire future dépend de variables politiques discrétionnaires, de réactions stratégiques européennes, et d'évolutions technologiques partiellement imprévisibles.

Le volet rétrospectif mobilise une méthode quantitative descriptive, fondée sur l'agrégation et la mise en cohérence de données issues de sources institutionnelles (IEA, SIA/WSTS, Eurostat, EIA), de rapports sectoriels (McKinsey, Deloitte, Epoch AI) et de documents réglementaires (Federal Register, BIS, White House). L'objectif est d'établir un socle factuel rigoureux et sourcé, couvrant trois dimensions : la consommation énergétique des data centers, le marché des semi-conducteurs, et la capacité de compute IA installée par région.

Le volet prospectif s'inscrit dans la tradition de la scenario planning telle que formalisée par Schwartz (1991) et pratiquée dans le cadre de Royal Dutch/Shell depuis les années 1970.[1] Cette méthode, relevant de l'école des Intuitive Logics (Bradfield et al., 2005), consiste à construire des scénarios plausibles et internement cohérents, non pas pour prédire l'avenir, mais pour explorer l'espace des possibles et évaluer la robustesse de différentes stratégies face à des évolutions divergentes de l'environnement.[2] Elle est particulièrement adaptée aux situations caractérisées par une forte incertitude politique et technologique, où les modèles économétriques classiques atteignent leurs limites — ce qui est précisément le cas du protectionnisme technologique IA.

Justification du choix méthodologique

Trois raisons fondent le recours à la méthode des scénarios plutôt qu'à une modélisation économétrique pure. Premièrement, les variables clés de l'analyse sont en grande partie politiques et discrétionnaires : la décision d'un président américain d'imposer ou non des quotas GPU à l'Europe ne peut être modélisée par une fonction de régression. Deuxièmement, les interactions entre les dimensions énergétique, technologique et géopolitique sont non linéaires et systémiques : une restriction sur les GPU peut, par effet de cascade, modifier les flux d'investissement énergétique, la localisation des data centers, et la structure concurrentielle de secteurs entiers. Troisièmement, les données disponibles sur le compute installé par région sont partielles et hétérogènes : il n'existe pas de base de données unifiée et publique recensant les FLOPs IA par pays, ce qui rend une calibration économétrique rigoureuse prématurée.

La méthode retenue combine donc la rigueur quantitative du diagnostic empirique (données sourcées, séries temporelles, ratios mesurables) avec la flexibilité qualitative de la construction par scénarios, dans l'esprit de ce que Schoemaker (1995) appelle une « heuristique disciplinée ».[3] Les scénarios ne sont pas des prévisions probabilistes mais des récits stratégiques cohérents, chacun fondé sur des hypothèses explicites et déclinant ses conséquences à travers des métriques mesurables.

2.2 Sources de données : classification et évaluation critique

L'étude mobilise trois catégories de sources, dont la fiabilité et les biais potentiels doivent être explicitement reconnus. Cette transparence méthodologique est conforme aux recommandations du OECD/JRC Handbook on Constructing Composite Indicators (Nardo et al., 2008), qui prescrit une documentation systématique des sources, de leurs limites et de leurs biais dans toute construction d'indicateur composite.[4]

2.2.1 Sources primaires (documents officiels et réglementaires)

Cette catégorie comprend les textes à force normative ou institutionnelle : proclamations présidentielles (Section 232), règles du BIS (AI Diffusion Rule, Entity List), rapports de l'IEA, publications du Parlement européen (EPRS), données statistiques officielles (SIA/WSTS pour les semi-conducteurs, Eurostat et EIA pour l'énergie, RTE pour la France). Ces sources offrent la plus haute fiabilité factuelle mais peuvent contenir des biais de cadrage institutionnel : l'IEA tend à privilégier des scénarios modérés, le Parlement européen à insister sur les risques pour la souveraineté EU.

2.2.2 Sources académiques et think tanks

Cette catégorie inclut les articles peer-reviewed (Farrell & Newman, 2019 ; Bresnahan & Trajtenberg, 1995 ; Brynjolfsson et al., 2019 ; Mügge, 2024) et les publications de think tanks reconnus (Bruegel, Carnegie Endowment, CSIS, OCDE, Federal Reserve Board). Les premiers offrent un ancrage théorique robuste ; les seconds fournissent des analyses de politique publique empiriquement fondées mais potentiellement influencées par les orientations idéologiques de chaque institution. Nous privilégions le croisement de sources d'orientations différentes (Bruegel / Carnegie / Fed) pour limiter ce biais.

2.2.3 Sources industry et consulting

McKinsey, Deloitte, Accenture, Epoch AI et CFG Europe fournissent des données de marché, des projections sectorielles et des estimations de capacité qui ne sont pas disponibles dans les sources publiques. Ces sources présentent un biais systématique potentiel : les cabinets de conseil ont intérêt à amplifier les tendances (pour justifier des missions de transformation), et les estimations de marché sont souvent optimistes.

Nous atténuons ce biais en triangulant les chiffres avec les données institutionnelles et en signalant explicitement les écarts entre sources. Par exemple, les ventes de semi-conducteurs 2024 sont de 627,6 milliards de dollars selon la SIA (périmètre traditionnel) mais de 775 milliards selon McKinsey (périmètre élargi) — un écart de 24 % qui reflète des différences méthodologiques, pas des incohérences.[5]

2.2.4 Données de compute IA : le dataset Epoch AI GPU Clusters

La mesure du compute IA installé par pays constitue le défi méthodologique central de cette étude. Nous nous appuyons principalement sur le dataset Epoch AI GPU Clusters (Pilz, Rahman, Sanders & Heim, 2025), qui recense plus de 500 superordinateurs et clusters GPU à travers le monde pour la période 2019–2025.[6] Ce dataset, disponible en accès libre sous licence Creative Commons Attribution, constitue à ce jour la source la plus complète et la plus systématiquement documentée sur l'infrastructure de compute IA mondiale. Il est utilisé comme référence par le Stanford AI Index Report (2025), par plusieurs rapports de gouvernements, et par des institutions telles qu'OpenAI et DeepMind.

Le dataset couvre pour chaque cluster : le pays d'implantation, le type de puces (H100, A100, GB200, TPU, etc.), la performance computationnelle en FLOP/s 16-bit, le nombre d'équivalents H100, la date de mise en service, le secteur (privé/public), la puissance électrique (MW) et le coût estimé du hardware. Cette granularité permet une agrégation par pays et par année, répondant directement aux besoins de notre variable F(r) dans le CACI.

Limites du dataset Epoch AI. Trois limites doivent être soulignées. Premièrement, la couverture est estimée à 10–20 % de la performance agrégée mondiale de compute IA (mars 2025), avec une hétérogénéité significative selon les entreprises et les types de puces : environ 20–37 % des H100 NVIDIA, 12 % des A100, mais moins de 4 % des TPU Google et une fraction négligeable des puces custom AWS, Microsoft ou Meta.[7] Deuxièmement, les systèmes chinois sont anonymisés (noms supprimés, valeurs arrondies à un chiffre significatif), ce qui limite la finesse de l'analyse pour la Chine. Troisièmement, la localisation physique d'un cluster ne détermine pas l'accès : beaucoup de clusters sont accessibles via des services cloud depuis d'autres pays.

Nous complétons ces données par le OECD Working Paper de Lehdonvirta, Wu, Hawkins et al. (octobre 2025), qui développe une méthodologie pour estimer la disponibilité de compute cloud public pour l'IA par pays, en comptabilisant les régions cloud des principaux fournisseurs disposant d'accélérateurs IA (A100, H100, GB200) dans 39 économies.[8] Cette approche complémentaire distingue le compute installé (Epoch AI) du compute accessible (OCDE), une distinction cruciale pour le CACI.

Catégorie	Sources principales	Données extraites	Biais potentiel
Primaires officielles	IEA, SIA/WSTS, BIS, White House, Eurostat, RTE, Parlement EU	TWh data centers, ventes semis, règles export, conso énergie	*Cadrage institutionnel, conservatisme*
Académiques / think tanks	Bruegel, Carnegie, CSIS, OCDE, Fed, SSRN (Hawkins et al.)	Cadres théoriques, analyses politiques, métriques compute	*Orientation idéologique variable*

Industry / consulting	McKinsey, Deloitte, Accenture, Epoch AI, CFG Europe	Projections marché, estimations capacité, coûts training	*Optimisme systématique, intérêts commerciaux*

Tableau 1. *Classification et évaluation critique des sources mobilisées.*

2.3 Construction des scénarios

La construction des scénarios suit un protocole en quatre étapes, inspiré de la méthodologie de la matrice 2×2 (Schwartz, 1991 ; van der Heijden, 2004) et adapté au contexte géostratégique de l'IA.[5]

Étape 1 — Identification des forces motrices

Nous identifions les forces motrices qui structurent l'évolution du système étudié. Ces forces sont classées en deux catégories, suivant la distinction classique de Schwartz entre éléments prédéterminés et incertitudes critiques.

Les éléments prédéterminés — dont l'évolution est raisonnablement prévisible quels que soient les scénarios — incluent : (i) la croissance continue de la demande mondiale de compute IA, (ii) la hausse structurelle de la consommation énergétique des data centers, (iii) la dépendance européenne sur les fonderies asiatiques et américaines pour le leading-edge, (iv) la concentration du cloud mondial autour de trois hyperscalers US, et (v) l'augmentation exponentielle des coûts d'entraînement des modèles de frontière. Ces tendances sont documentées dans le Chapitre III et constituent le socle commun à tous les scénarios.

Les incertitudes critiques — dont l'évolution dépend de choix politiques, de réactions stratégiques, ou de ruptures technologiques — sont regroupées en deux dimensions :

Dimension 1 : l'intensité du protectionnisme technologique US. Cette dimension couvre un spectre allant du maintien des restrictions actuelles (tarifs ciblés Chine, EU largement exemptée, accès cloud préservé) jusqu'à un durcissement agressif (quotas GPU pour l'EU, restrictions sur les API et modèles, priorisation explicite des livraisons aux entreprises américaines). Les facteurs déterminants incluent l'évolution des relations US-Chine, les pressions du lobby industriel américain, et les arbitrages internes de l'administration Trump entre sécurité nationale et intérêts commerciaux.

Dimension 2 : la capacité de réponse européenne. Cette dimension couvre un spectre allant d'une posture passive (adaptation marginale, acceptation de la dépendance, absence d'investissements massifs dans le compute souverain) jusqu'à une riposte active (Compute Zones à énergie dérogée, AI Factories accélérées, nucléaire SMR pour data centers, partenariats alternatifs Japon-Corée-Taiwan,

révision de l'AI Act pour alléger les coûts de conformité). Les facteurs déterminants incluent la volonté politique EU, la capacité de mobilisation budgétaire, et la vitesse de déploiement des infrastructures énergétiques.

Étape 2 — Matrice 2×2 et génération des scénarios

Le croisement des deux dimensions d'incertitude génère une matrice de quatre scénarios, chacun représentant une combinaison cohérente d'hypothèses :

	EU passive	**EU active (riposte)**
US statu quo renforcé	Scénario A — Dérive lente. Gap stable, dépendance croissante, vassalisation douce.	Scénario B — Rattrapage partiel. EU investit massivement, gap réduit, autonomie renforcée.
US durcissement agressif	Scénario C — Vassalisation. Quotas GPU EU, écart productivité -25 %, délocalisations massives.	Scénario D — Guerre froide technologique. Fragmentation bloc occidental, coûts élevés, autonomie forcée.

Tableau 2. *Matrice des quatre scénarios 2026–2030.*

Le choix de quatre scénarios plutôt que trois est délibéré. Schwartz (1991) et les praticiens de la méthode Shell recommandent de *ne jamais construire trois scénarios*, car l'esprit humain tend à traiter le scénario médian comme le « plus probable », réduisant ainsi l'utilité de l'exercice.[6] La matrice 2×2 force l'analyste à explorer les quadrants extrêmes (C et D), qui sont précisément ceux où les ruptures stratégiques se jouent.

Étape 3 — Développement narratif et quantification

Chaque scénario est développé selon un protocole standardisé en trois composantes :

(a) Un récit stratégique décrivant la séquence d'événements plausibles entre 2026 et 2030 (décisions politiques, réactions des acteurs, évolutions de marché). Ce récit doit être internement cohérent : chaque conséquence découle logiquement des hypothèses posées.

(b) Une quantification des métriques clés (détaillées en section 2.4), calibrée sur les données empiriques 2020–2026 et projetée selon les hypothèses du scénario. Cette quantification est présentée comme un ordre de grandeur plausible, pas comme une prévision ponctuelle.

(c) Des indicateurs d'alerte précoce (*leading indicators*) permettant d'identifier, dès 2026–2027, vers quel scénario la réalité converge. Par exemple : le

volume de GPU livrés à l'EU (indicateur d'intensité protectionnisme US), ou le nombre de GW de Compute Zones autorisées (indicateur de capacité de réponse EU).

Étape 4 — Analyse de sensibilité et robustesse

Pour chaque recommandation formulée au Chapitre VII, nous évaluons sa robustesse face aux quatre scénarios. Une recommandation est dite robuste si elle produit des résultats positifs ou neutres dans au moins trois des quatre scénarios. Cette approche, conforme à la logique de robustesse de la planification par scénarios, privilégie les stratégies qui ne parient pas sur un avenir unique.

2.4 Métriques clés et indicateur original : le CACI

2.4.1 Les six métriques de divergence

Nous définissons six métriques qui seront calculées ou estimées dans le diagnostic empirique (Chapitre III), puis projetées dans chaque scénario (Chapitre V). Ensemble, elles forment un tableau de bord de la divergence US/EU en matière d'IA.

#	Métrique	Définition	Source(s)	Unité
M1	**Compute gap**	Ratio FLOPs IA installés US / EU, normalisé par PIB	Epoch AI, CFG, Top500	Ratio (sans unité)
M2	**Coût relatif du FLOPs**	Prix moyen du TFlop-seconde pour le training, US vs EU	Bruegel, Epoch AI, Cloud providers	$/TFlop·s
M3	**Dépendance cloud**	Part des workloads IA EU exécutés sur infra US (%)	Synergy Research, Accenture	%
M4	**Productivité IA sectorielle**	Gain de productivité attribué à l'IA dans les secteurs clés	McKinsey, Fed Board, Eurostat	% gain
M5	**Contrainte énergétique**	Ratio demande énergétique data centers / capacité disponible	IEA, RTE, Ember	Ratio
M6	**Délocalisations IA**	Part des projets IA critiques EU délocalisés vers les US	Estimation propre (enquetes sectorielles)	% projets

Tableau 3. *Les six métriques de divergence US/EU.*

2.4.2 Le Compute-Adjusted Competitiveness Index (CACI) : fondements théoriques et construction

Ancrage dans la littérature. La construction d'un indicateur composite de compétitivité IA centré sur le compute répond à un besoin identifié par plusieurs courants de recherche convergents. Depuis 2023–2024, la littérature académique et institutionnelle souligne de manière croissante que la capacité de calcul est devenue le facteur de production le plus discriminant pour l'IA de frontière (Sevilla et al., 2022 ; Epoch AI, 2025 ; Pilz et al., 2025). Les export controls américains (BIS, octobre 2022 ; actualisés en 2023 et 2025) placent explicitement le compute avancé au cœur de la compétition géopolitique, tandis que Hawkins, Lehdonvirta & Wu (2025) introduisent le concept de « compute sovereignty » comme dimension structurante de l'autonomie stratégique.[11]

Or, les indices de compétitivité IA existants ne placent pas le compute au centre de leur construction. L'AI Preparedness Index du FMI (Cazzaniga et al., 2024), qui couvre 174 pays, agrège quatre dimensions (infrastructure numérique, capital humain, innovation/intégration économique, régulation/éthique) sans mesurer

directement la capacité de calcul installée.[12] Le Global AI Index de Tortoise Media (2024), qui classe 83 pays sur 122 indicateurs regroupés en trois piliers (Implementation, Innovation, Investment), inclut une composante infrastructure/supercomputing mais la noie dans un indice additif pondéré où le compute n'est qu'un facteur parmi d'autres.[13] Le Stanford AI Index (2025) fournit des données riches sur les tendances de compute mais ne propose pas d'indice composite. Aucun de ces instruments ne formalise le mécanisme par lequel le compute, ajusté de son coût énergétique et rapporté à la capacité d'absorption d'une économie, détermine la compétitivité IA.
Le CACI vise à combler cette lacune en proposant un indicateur parcimonieux mais théoriquement fondé, qui capture l'interaction multiplicative entre trois facteurs : le compute accessible, le coût énergétique qui le contraint, et la capacité économique et humaine à l'exploiter.

Définition formelle. Le CACI pour une région r à la période t est défini comme :

$$\mathbf{CACI(r,t) = [\ F(r,t) \times E(r,t)^{-1}\ ] / [\ PIB(r,t) \times L(r,t)\ ]}$$

où :

F(r,t) = capacité de compute IA installée et accessible dans la région r à la période t, mesurée en FLOP/s agrégés (performance 16-bit). Nous utilisons les données Epoch AI GPU Clusters agrégées par pays et par année, complétées par les estimations OCDE de compute cloud accessible. L'unité retenue est le PetaFLOP/s (10^{15} FLOP/s), incluant les capacités cloud domestiques et les quotas d'accès aux clouds étrangers autorisés.
E(r,t) = coût énergétique moyen du compute dans la région r, en €/MWh pour les data centers. Ce facteur ajuste le compute brut de sa contrainte énergétique : à FLOPs égaux, un pays avec une électricité deux fois plus chère a un CACI deux fois plus bas. Les données proviennent d'Eurostat (tarifs industriels par pays) et de l'EIA (prix US), ajustées pour les Power Purchase Agreements des hyperscalers via les estimations IEA (2025).

PIB(r,t) = produit intérieur brut de la région r (Banque mondiale, Eurostat). La normalisation par le PIB assure la comparabilité entre économies de tailles très différentes : sans elle, les États-Unis et la Chine écraseraient mécaniquement le classement par leur masse économique.

L(r,t) = population active disposant de compétences IA (proxy : diplômés STEM + formations IA certifiées). Ce facteur capture la capacité d'absorption au sens de Cohen & Levinthal (1990) : un compute abondant sans capital humain pour l'exploiter ne produit pas de compétitivité. Les données OCDE sur les diplômés STEM sont complétées par les estimations de LinkedIn Economic Graph sur la densité des compétences IA par pays.[14]

Justification de la forme multiplicative. Le choix d'une agrégation multiplicative (géométrique) plutôt qu'additive (arithmétique) est délibéré et repose sur trois arguments. Premièrement, la théorie des General Purpose Technologies (Bresnahan & Trajtenberg, 1995) postule une forte complémentarité entre les inputs d'innovation : le compute sans énergie abordable ou sans capital humain qualifié ne produit pas de gains de compétitivité, ce qui justifie une forme où la faiblesse d'un facteur pénalise l'ensemble. Deuxièmement, le OECD/JRC Handbook (2008, p. 33) recommande l'agrégation géométrique lorsque les composantes ne sont pas

parfaitement substituables : contrairement à la moyenne arithmétique, la moyenne géométrique ne permet pas qu'un score très élevé sur une dimension compense entièrement un score très faible sur une autre. Troisièmement, des travaux récents sur la construction d'indices IA (Koronakos, Kritikos & Sotiros, 2024 ; analyse du Tortoise GAII via l'intégrale de Choquet) confirment que les dimensions de compétitivité IA présentent des interactions (complémentarités et redondances) qui rendent problématique une agrégation linéaire simple.[15]

Interprétation économique. En forme logarithmique, le CACI s'écrit : ln(CACI) = ln(F) – ln(E) – ln(PIB) – ln(L). Cette transformation est commode pour l'analyse économétrique car elle linéarise la relation et permet d'interpréter chaque coefficient comme une élasticité. L'indicateur est conçu pour être utilisé en comparaison bilatérale : le ratio CACI(US)/CACI(EU) ou CACI(US)/CACI(FR) mesure l'avantage concurrentiel relatif. Un ratio de 7 signifie que, par unité de PIB et à capital humain égal, les acteurs américains disposent de sept fois plus de compute effectif (ajusté du coût énergétique) que les acteurs européens.

2.4.3 Protocole de calibration et sources de données

La calibration du CACI suit un protocole en quatre étapes, conforme aux recommandations du OECD/JRC Handbook (2008) pour la construction d'indicateurs composites : (i) identification et collecte des données brutes, (ii) traitement des valeurs manquantes et normalisation, (iii) agrégation, et (iv) analyse de sensibilité.

F(r,t) — Capacité de compute IA installée. L'estimation suit une approche bottom-up en trois couches. La couche principale provient du dataset Epoch AI GPU Clusters (version février 2026, 746 clusters), qui fournit la performance FLOP/s 16-bit agrégée par pays pour la période 2019–2025. Les parts nationales en mai 2025 sont : États-Unis ~74,5 %, Chine ~14,1 %, Union européenne ~4,8 %, Norvège ~1,8 %, Japon ~1,4 %.[16] La couche complémentaire provient de l'OCDE (Lehdonvirta et al., 2025), qui recense les régions cloud disposant d'accélérateurs IA dans 39 économies, capturant la dimension « accessibilité » que le compute physiquement installé ne reflète pas entièrement. La troisième couche utilise les données partielles publiées par les hyperscalers (Microsoft, Google, Meta, OVHcloud) et les estimations CFG Europe pour la capacité européenne, croisées avec le classement Top500 pour le HPC public.

Procédure d'agrégation de F. Pour chaque pays-année, nous : (1) extrayons du dataset Epoch AI la somme des FLOP/s 16-bit de tous les clusters opérationnels localisés dans le pays, en filtrant les systèmes confirmés (certainty ≥ « Likely ») ; (2) appliquons un facteur d'extrapolation pour corriger la sous-couverture du dataset (~10–20 % du compute mondial), en utilisant les estimations sectorielles d'Epoch AI sur la couverture par type de puce ; (3) ajoutons les capacités cloud accessibles estimées via la méthodologie OCDE pour les pays où le compute cloud étranger représente une part significative de la capacité effective (cas typique des petites économies européennes). Les données brutes et le code de calcul (Python/pandas) sont documentés dans l'annexe méthodologique.[17]

E(r,t) — Coût énergétique. Les données Eurostat (tarifs industriels de l'électricité par pays, bande de consommation IE) et EIA (Average Retail Price of Electricity, Industrial) fournissent la base. Nous ajustons pour les tarifs négociés des gros consommateurs (PPA des hyperscalers), en utilisant les estimations de l'IEA (2025, Energy and AI) sur le mix énergétique des data centers par région. Le coût EU est typiquement 2 à 3 fois celui des US pour l'électricité industrielle, avant PPA.[18] Le

Federal Reserve Board (octobre 2025) documente une corrélation négative significative entre coûts énergétiques et adoption IA au niveau des entreprises européennes.

PIB(r,t) et L(r,t). Le PIB est disponible via la Banque mondiale (World Development Indicators) et Eurostat. Le proxy de capital humain IA L(r,t) combine trois sous-indicateurs : (i) le nombre de diplômés STEM (OCDE, Education at a Glance), (ii) la densité des compétences IA mesurée par LinkedIn Economic Graph (nombre de profils avec compétences IA rapporté à la population active), et (iii) les certifications IA (estimations basées sur les programmes cloud certifiés AWS/Google/Microsoft par pays). Ce proxy est cohérent avec les approches utilisées par le Federal Reserve Board (2025) et par le FMI dans l'AI Preparedness Index (composante Human Capital). Nous reconnaissons un biais en faveur des pays anglophones et des économies où LinkedIn est dominant, et documentons l'effet de ce biais dans l'analyse de sensibilité (section 2.4.5).[19]

2.4.4 Positionnement par rapport aux indices de compétitivité IA existants

Nous situons le CACI dans le paysage des indicateurs de compétitivité IA, en identifiant ses complémentarités et ses différences avec les principaux indices publiés. Cette comparaison est essentielle pour établir la valeur ajoutée spécifique du CACI et pour éviter la multiplication redondante d'indicateurs (Saisana & Tarantola, 2002).

L'AI Preparedness Index du FMI (AIPI) couvre 174 pays (2023) et agrège quatre piliers : infrastructure numérique, capital humain et politiques du marché du travail, innovation et intégration économique, régulation et éthique. L'AIPI ne mesure pas le compute IA installé et ne pondère pas par le coût énergétique. Sa corrélation attendue avec le CACI est positive mais imparfaite : les pays qui scorent haut sur l'AIPI (Singapour, Danemark, Pays-Bas) ne sont pas nécessairement ceux qui disposent du plus de compute effectif par unité de PIB.[20]

Le Global AI Index de Tortoise Media (GAII) classe 83 pays sur 122 indicateurs regroupés en trois piliers (Implementation, Innovation, Investment). Il inclut une composante infrastructure/supercomputing, mais l'agrège linéairement avec des pondérations subjectives (reconnues par Tortoise comme une limite). Le GAII est plus large et multi-dimensionnel que le CACI, mais précisément parce qu'il est large, il dilue le signal du compute dans un composite à nombreux indicateurs. Comme le soulignent Koronakos et al. (2024), la subjectivité des pondérations du GAII peut inverser les classements de pays selon les scénarios de pondération choisis.

Le Stanford AI Index (2025) constitue la référence la plus complète en termes de données brutes (modèles notables par pays, investissements, publications, brevets, tendances de compute). Il ne propose pas d'indice composite mais fournit les séries temporelles utilisées par de nombreux autres indices. Les données publiques du Stanford AI Index sont disponibles via un dossier Google Drive ouvert.[21]

Valeur ajoutée spécifique du CACI. Le CACI se distingue par quatre propriétés : (i) il place le compute au centre plutôt qu'en périphérie de l'indicateur, reflétant le rôle désormais dominant de la capacité de calcul dans l'IA de frontière ; (ii) il intègre explicitement le coût énergétique comme goulot d'étranglement, conformément aux constats de l'IEA (2025) ; (iii) il utilise une agrégation multiplicative théoriquement

fondée plutôt qu'additive ; et (iv) il est parcimonieux (quatre variables), ce qui le rend transparent et reproductible, au prix d'une moindre exhaustivité.

2.4.5 Limites du CACI et analyse de sensibilité

Nous reconnaissons cinq limites majeures de l'indicateur, chacune assortie d'une stratégie d'atténuation.

Première limite : l'opacité de F(r,t). La mesure du compute installé est tributaire de données privées incomplètes. Le dataset Epoch AI ne couvre que 10–20 % du compute mondial, avec une couverture inégale entre secteurs et entreprises. L'attribution nationale est elle-même discutable : une part croissante du compute est détenue par quelques hyperscalers privés qui opèrent à l'échelle mondiale. Atténuation : nous documentons systématiquement les marges d'incertitude, présentons des fourchettes plutôt que des valeurs ponctuelles, et vérifions la stabilité des résultats lorsque F varie de ±30 %.

Deuxième limite : l'hétérogénéité qualitative du compute. Le CACI agrège des FLOPs sans distinguer les générations de GPU (un H200 n'équivaut pas à un A100 en termes d'efficacité énergétique et de performance réelle). Atténuation : le dataset Epoch AI fournit la performance en équivalents H100, ce qui offre déjà une normalisation partielle. Nous proposons en annexe un facteur de pondération par génération de GPU et montrons que l'ajustement modifie marginalement les classements.

Troisième limite : le proxy de capital humain L(r,t). La combinaison STEM + LinkedIn + certifications présente un biais en faveur des pays anglophones et des économies où LinkedIn est largement utilisé. Ce biais sous-estime probablement la Chine et certaines économies d'Asie. Atténuation : nous répliquons l'analyse en utilisant le sous-indice « Human Capital » de l'AIPI du FMI comme proxy alternatif, et montrons la sensibilité des classements à ce choix.

Quatrième limite : la staticité de l'indicateur. Le CACI mesure un état à un instant donné, pas une dynamique. C'est pourquoi nous le calculons sur plusieurs années (2022, 2024, 2026) et le projetons dans chaque scénario, ce qui permet de tracer une trajectoire et d'évaluer l'évolution du gap.

Cinquième limite : l'endogénéité. Les pays qui gagnent en productivité IA investissent massivement dans le compute, créant un risque de causalité inverse : le CACI pourrait capter davantage la conséquence que la cause de la compétitivité. Cette étude, dans son cadre de mémoire, n'instrument pas formellement cette relation (absence de variables instrumentales ou de stratégie GMM). Toutefois, nous notons que le choc exogène des export controls BIS d'octobre 2022 offre un quasi-expériment naturel qui pourrait fonder une stratégie d'identification causale en Difference-in-Differences : la Chine (traitée) subit un plafonnement brutal de F tandis que les États-Unis (contrôle) accélèrent. Nos figures du Chapitre III illustrent cette divergence. Nous identifions le traitement formel de cette endogénéité (DiD, IV, ou GMM Arellano-Bond) comme une piste de recherche prioritaire pour une éventuelle extension publiable de ce travail.[22]

Malgré ces limites, le CACI répond à un besoin identifié dans la littérature : il n'existe à ce jour aucun indicateur formalisé de compétitivité ajustée au compute permettant de comparer systématiquement les régions. La contribution méthodologique réside dans le framework plus que dans la précision des chiffres :

même avec des données approximatives, le CACI rend visible l'écart structurel que les indicateurs traditionnels (PIB, dépenses R&D, brevets) ne capturent pas.

Afrique du Sud qui "explose" avec le CACI score (la courbe qui monte à 100) alors qu'elle est dernière sur les deux autres indices est un résultat très contre-intuitif — mais il est **mécaniquement explicable** par la formule.

la formule : **CACI = [F × E⁻¹] / [PIB × L]**

Le dénominateur c'est **PIB × L**. L'Afrique du Sud a un PIB relativement petit (~400 Md$) et un L (workforce IA) très faible. Donc le dénominateur est minuscule. Si elle a ne serait-ce qu'un peu de compute F (quelques clusters) et une énergie pas trop chère (l'Eskom, malgré le load-shedding, a des tarifs industriels parmi les plus bas du monde, ~0,05-0,07 $/kWh), alors le numérateur divisé par un tout petit dénominateur donne un score artificiellement gonflé.

C'est exactement le type de **biais de normalisation** que le OECD/JRC Handbook (2008) signale : quand on normalise par PIB × L, les petites économies avec un peu d'infrastructure apparaissent disproportionnément fortes. C'est le même problème qu'on voit avec le PIB par habitant de l'Islande ou du Luxembourg dans d'autres indices.

Cela révèle une limite qu'il faut documenter :

Il faudrait soit introduire un **seuil minimum de F** en dessous duquel le CACI n'est pas calculé (éviter les pays où un seul cluster fausse tout), soit pondérer par un **facteur d'échelle** (masse critique), soit simplement restreindre le CACI aux économies dépassant un certain seuil de compute installé (par exemple les 20 premiers pays en FLOPs absolus).

2.5 Périmètre et délimitations

Périmètre géographique. L'analyse se concentre sur la relation bilatérale États-Unis / Union européenne, avec un focus spécifique sur la France. La Chine est traitée comme variable contextuelle (cible principale des export controls US, facteur de pression sur les capacités de production de puces), mais ne fait pas l'objet d'une analyse approfondie. Le Japon, la Corée du Sud et Taïwan interviennent comme acteurs de la chaîne d'approvisionnement des semi-conducteurs.

Périmètre temporel. Le diagnostic couvre la période 2020–2026, les scénarios la période 2026–2030. L'horizon 2030 est choisi car il correspond à la convergence de plusieurs échéances : projections IEA pour l'énergie des data centers, maturité prévue du Chips Act EU, objectifs de la SNIA France 2030, et potentielle arrivée des premiers SMR nucléaires opérationnels.

Périmètre technologique. L'étude couvre l'IA de frontière (modèles de fondation, compute intensif) et ses prérequis matériels (GPU/ASIC, data centers, énergie). Elle intègre la robotique IA comme facteur amplificateur de la demande énergétique. Elle n'aborde pas l'IA embarquée edge (smartphones, IoT), sauf dans la mesure où celle-ci constitue un objectif spécifique de la SNIA française.

2.6 Limites méthodologiques générales

Incertitude politique radicale. Le protectionnisme technologique dépend de décisions politiques discrétionnaires dont la prévisibilité est structurellement faible. Un changement d'administration US en 2028, un accord commercial US-EU inattendu, ou une escalade du conflit US-Chine pourraient invalider certaines hypothèses. C'est précisément la raison pour laquelle nous proposons quatre scénarios plutôt qu'une trajectoire unique.

Ruptures technologiques. L'épisode DeepSeek (janvier 2025), où un modèle chinois a atteint des performances proches des frontières avec un budget d'entraînement sensiblement réduit, illustre la possibilité de ruptures d'efficacité qui modifieraient les termes du problème. L'IEA (2025, Energy and AI) consacre une étude de cas à DeepSeek et conclut que même avec des améliorations d'efficacité significatives, la croissance de la demande absorbe les gains (effet rebond de Jevons).[23]

Opacité des données compute. Le nombre exact de GPU déployés par hyperscaler, la répartition géographique précise des data centers, et les volumes de GPU exportés par région sont des données partiellement ou totalement confidentielles. Nos estimations de compute installé comportent une marge d'erreur significative, que nous documentons systématiquement.

Biais des sources consultants. Comme noté en section 2.2, les sources industry ont un biais d'optimisme systématique. Nous atténuons ce biais par la triangulation mais ne pouvons l'éliminer entièrement.

Ces limites ne compromettent pas la validité de l'analyse. La méthode des scénarios est précisément conçue pour fonctionner dans des environnements de forte incertitude, où l'objectif n'est pas la prédiction mais l'exploration structurée des possibles. Comme le souligne Schwartz, « les scénarios ne sont pas des prévisions ; ce sont des histoires plausibles qui vous aident à réfléchir ».[24] Notre contribution réside dans la rigueur du cadrage, l'explicitation des hypothèses, la transparence des sources de données, et l'originalité de l'indicateur CACI, plus que dans la précision des projections chiffrées.

Notes

[1] Schwartz, P. (1991), The Art of the Long View: Planning for the Future in an Uncertain World, New York, Doubleday. Voir également Wack, P. (1985), « Scenarios: Uncharted Waters Ahead », Harvard Business Review, sept.-oct. 1985, pp. 72-89.

[2] Bradfield, R., Wright, G., Burt, G., Cairns, G. & Van Der Heijden, K. (2005), « The Origins and Evolution of Scenario Techniques in Long Range Business Planning », Futures, 37(8), pp. 795-812.

[3] Schoemaker, P.J.H. (1995), « Scenario Planning: A Tool for Strategic Thinking », MIT Sloan Management Review, 36(2), pp. 25-40.

[4] Nardo, M., Saisana, M., Saltelli, A. & Tarantola, S. (2008), Handbook on Constructing Composite Indicators: Methodology and User Guide, OECD Publishing, Paris. Le Handbook prescrit un protocole en dix étapes : cadre théorique, sélection des données, imputation, analyse multivariée, normalisation, pondération, agrégation, robustesse/sensibilité, retour aux données, présentation/visualisation.

[5] L'écart SIA/McKinsey s'explique par le périmètre : McKinsey (janvier 2026, « Hiding in Plain Sight ») inclut la valeur des captive designers (Apple, Amazon, Tesla) et des opérateurs fabless dont les ventes n'apparaissent pas dans les statistiques WSTS.

[6] Pilz, K.F., Rahman, R., Sanders, J. & Heim, L. (2025), « Trends in AI Supercomputers », arXiv:2504.16026, avril 2025. Le dataset est accessible à https://epoch.ai/data/gpu-clusters sous licence Creative Commons Attribution.

[7] Epoch AI (2025), GPU Clusters Documentation. La couverture est estimée à ~20-37 % des NVIDIA H100, ~12 % des A100, et ~18 % des AMD MI300X, mais moins de 4 % des TPU Google.

[8] Lehdonvirta, V., Wu, B., Hawkins, Z.J., Caira, C. & Russo, L. (2025), « Measuring Domestic Public Cloud Compute Availability for Artificial Intelligence », OECD Artificial Intelligence Papers, No. 49, OECD Publishing, Paris, https://doi.org/10.1787/8602a322-en.

[9] Van der Heijden, K. (2004), Scenarios: The Art of Strategic Conversation, 2nd ed., Chichester, Wiley.

[10] Schwartz (1991), op. cit., pp. 241-243. La règle des « pas trois scénarios » est également endossée par le UK Government Office for Science, Scenario Planning Guidance Note (2009).

[11] Hawkins, Z.J., Lehdonvirta, V. & Wu, B. (2025), « AI Compute Sovereignty: Infrastructure Control Across Territories, Cloud Providers, and Accelerators », SSRN, juin 2025, https://ssrn.com/abstract=5312977. Sevilla, J. et al. (2022), « Compute Trends Across Three Eras of Machine Learning », arXiv:2202.05924.

[12] Cazzaniga, M. et al. (2024), « Gen-AI: Artificial Intelligence and the Future of Work », IMF Staff Discussion Note 2024/001. L'AI Preparedness Index est accessible via le dashboard interactif du FMI : https://www.imf.org/external/datamapper/AI_PI@AIPI/.

[13] Tortoise Media (2024), The Global Artificial Intelligence Index 2024, septembre 2024. Méthodologie disponible à https://www.tortoisemedia.com/data/global-ai. Voir également Koronakos, G., Kritikos, M. & Sotiros, D. (2024), « Mitigating Subjectivity and Bias in AI Development Indices », Expert Systems with Applications, pour une critique des pondérations du GAII via l'intégrale de Choquet.

[14] Cohen, W.M. & Levinthal, D.A. (1990), « Absorptive Capacity: A New Perspective on Learning and Innovation », Administrative Science Quarterly, 35(1), pp. 128-152.

[15] L'agrégation géométrique est également utilisée par le Human Development Index du PNUD depuis 2010, pour des raisons analogues de non-substituabilité entre

dimensions (santé, éducation, revenu). Voir OECD/JRC Handbook (2008), pp. 31-33, pour une discussion comparative des méthodes d'agrégation.

16 Epoch AI (2025), « The US Hosts the Majority of GPU Cluster Performance, Followed by China », juin 2025. Données disponibles à https://epoch.ai/data-insights/ai-supercomputers-performance-share-by-country.

17 Le code Python et les données sont reproductibles. Le dataset Epoch AI est téléchargeable en CSV à https://epoch.ai/data/gpu_clusters.csv (rafraîchissement quotidien). La documentation complète est accessible à https://epoch.ai/data/gpu-clusters-documentation.

18 Le Federal Reserve Board (octobre 2025) documente une corrélation négative significative entre coûts énergétiques et adoption IA au niveau des entreprises européennes. IEA (2025), Energy and AI.

19 Le biais LinkedIn est documenté : dans les pays où LinkedIn est peu utilisé (Chine, Russie, certains pays d'Asie du Sud-Est), la densité des compétences IA est mécaniquement sous-estimée. Ce biais est partiellement corrigé par la composante STEM-OCDE et par les certifications cloud.

20 FMI (2024), AI Preparedness Index Dashboard, https://www.imf.org/external/datamapper/datasets/AIPI. Singapour, Danemark, Pays-Bas et États-Unis occupent les premières places ; la Chine est 31ème (score 0,63).

21 Maslej, N. et al. (2025), « Artificial Intelligence Index Report 2025 », Stanford Institute for Human-Centered AI, avril 2025. Données publiques : https://drive.google.com/drive/folders/1AxxxL9-AsaeMdDKtTNHCR1KqEJTsHCod.

22 La stratégie d'identification causale la plus prometteuse exploiterait le shock exogène des règles BIS d'octobre 2022 dans un cadre Difference-in-Differences (DiD), avec la Chine comme groupe traité et les États-Unis/alliés comme groupe contrôle. Voir également RAND (2025) sur l'impact des controles chips. Pour le GMM, voir Arellano, M. & Bond, S. (1991), « Some Tests of Specification for Panel Data », Review of Economic Studies, 58(2), pp. 277-297.

23 IEA (2025), Energy and AI, consacre une étude de cas à DeepSeek et conclut que même avec des améliorations d'efficacité significatives, la croissance de la demande absorbe les gains (effet rebond de Jevons).

24 Schwartz (1991), op. cit., p. 38. Traduction de l'auteur.

25 Le problème est analogue à celui du « Singapore effect » dans les indices de compétitivité : les petites économies ouvertes et spécialisées (Singapour, Luxembourg, Irlande) dominent systématiquement les classements normalisés par PIB ou population, non pas parce qu'elles sont structurellement supérieures, mais parce que le dénominateur amplifie le signal. Voir Saisana, M. & Tarantola, S. (2002), State-of-the-art Report on Current Methodologies and Practices for Composite Indicator Development, JRC European Commission, pp. 14-16.

26 OECD/JRC Handbook (2008), pp. 27-29 : « When normalising by GDP or population, users should be aware that small economies may obtain extreme values [...] Presenting both raw and normalised data is recommended. » La solution du double classement (intensité/échelle) est également utilisée par Tortoise Media (2024) dans le Global AI Index.