ÉTUDE DE RECHERCHE — FÉVRIER 2026

AI FOR AMERICANS FIRST

FAQ 1

Analyse Géostratégique et Économique Intégrée

Fabrice Pizzi

Université Sorbonne

Master Intelligence Économique — Intelligence Warfare

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

PARTIE I — LA THÈSE ET SON POSITIONNEMENT

Q1. Quelle est votre thèse centrale, en une phrase ?

Les États-Unis ont construit, sous Trump 2.0, une architecture protectionniste à trois étages (export controls, tarifs Section 232, gravité capitalistique) qui transforme l'accès au compute IA en levier géopolitique, créant un ratio de compétitivité US/EU de 7 à 12:1 mesurable par l'indice CACI — ratio qui, sans réponse européenne structurée d'ici 2028, deviendra irréversible.

Q2. Qu'apportez-vous de nouveau par rapport à la littérature existante ?

Quatre contributions originales. Premièrement, l'intégration analytique de quatre dimensions traitées séparément dans la littérature : trajectoires énergétiques des data centers, marché des semi-conducteurs, distribution du compute IA, et chronologie réglementaire US. Aucune étude ne croisait ces quatre dimensions.

Deuxièmement, la proposition du CACI — le premier indice formalisé de compétitivité ajustée au compute, validé économétriquement sur un panel de 12 pays. Ni Hawkins et al. (Oxford), ni le FMI, ni la Fed n'avaient construit un tel indicateur.

Troisièmement, la démonstration que le protectionnisme IA produit des effets paradoxaux systémiques : accélération de l'écosystème chinois alternatif, poussée des Tier 2 vers la Chine, co-financement de la suprématie US par les alliés eux-mêmes.

Quatrièmement, l'analyse comparative de trois zones géographiques (Europe, Amérique du Sud, Asie) montrant des trajectoires de dépendance structurellement différentes et irréductibles à un modèle unique.

Q3. Pourquoi ce sujet est-il pertinent en 2026 ?

Parce que trois événements convergent entre janvier et juillet 2026 qui rendent l'analyse urgente. D'abord, la Section 232 de Trump (15 janvier 2026) impose 25 % de tarifs sur les GPU IA avancés — c'est la première fois que des semi-conducteurs IA sont traités comme un enjeu de sécurité nationale comparable à l'acier. Ensuite, le capex cumulé des Big Tech US atteint 675 milliards de dollars en 2026, créant une concentration sans précédent du compute sur le sol américain. Enfin, la règle finale BIS doit être mise à jour d'ici juillet 2026, avec un potentiel de durcissement vers les alliés Tier 1.

Le sujet n'est pas théorique : il détermine concrètement si une PME industrielle française pourra ou non accéder au compute nécessaire pour rester compétitive en 2028-2030.

Q4. Pourquoi venir de la cybersécurité pour traiter ce sujet ? Quelle légitimité ?

Précisément parce que la cybersécurité m'a formé à penser en termes de chaînes de dépendance, de surface d'attaque et de points de contrôle. Le protectionnisme IA opère par les mêmes mécanismes que les cybermenaces : identification des goulots d'étranglement (TSMC pour les puces, ASML pour la litho, Nvidia pour les GPU), exploitation des positions de monopole, et weaponization des interdépendances — concept formalisé par Farrell & Newman (2019) que j'applique au compute.

Mon Master en Intelligence Économique, orienté Intelligence Warfare, fournit le cadre analytique complémentaire : l'IE c'est la capacité à comprendre et anticiper les rapports de force économiques, ce qui est exactement l'objet de cette étude. La double casquette IE/cyber est en fait un avantage : peu de chercheurs croisent ces deux mondes.

PARTIE II — L'INDICE CACI

Q5. Le CACI n'existe dans aucune autre publication. N'est-ce pas un problème ?

C'est le contraire : c'est la contribution. Le CACI n'est pas sorti de nulle part — il est la formalisation d'un besoin explicitement identifié par six courants de la littérature. Hawkins et al. (Oxford, 2025) mesurent le compute par région mais ne construisent pas d'indice. La Fed Board (octobre 2025) note textuellement que « l'absence d'un indicateur unifié de capacité IA rend les comparaisons transatlantiques difficiles ». Martens/Bruegel (2024) identifie les barrières mais ne propose aucune métrique. Le FMI a un AI Preparedness Index sans composante compute.

Le CACI est la pièce manquante que tout le monde appelait. Le fait qu'il soit nouveau est sa raison d'être, pas sa faiblesse.

Q6. Comment justifiez-vous la formule multiplicative plutôt qu'additive ?

La forme $CACI(r) = [F(r) \times E(r)^{-1}] / [PIB(r) \times L(r)]$ est multiplicative parce que les facteurs sont complémentaires, non substituables. Du compute abondant sans capital humain pour l'exploiter ne produit pas de compétitivité (cas des Émirats). Du talent sans accès au compute non plus (cas de certains pays africains à fort capital humain STEM).

La validation économétrique confirme ce choix : le modèle à pondérations alternatives (combinaison linéaire F:40%, E:25%, L:20%, Reg:15%) perd sa significativité ($p = 0,40$), tandis que la formule multiplicative donne $\beta = 0,25$ significatif à 1 %. La donnée tranche le débat théorique.

Q7. Votre panel n'a que 60 observations. N'est-ce pas trop peu ?

C'est une limite assumée et explicitement discutée (§A.6). Avec 12 pays × 5 ans, les degrés de liberté sont modestes — c'est pourquoi nous utilisons des écarts-types clustered robustes et testons systématiquement la sensibilité aux exclusions d'outliers.

Mais trois arguments relativisent cette limite. D'abord, le coefficient β du CACI est significatif à 1 % dans les trois spécifications (OLS, FE, RE), ce qui est remarquable pour un panel aussi compact. Ensuite, le R^2 within atteint 0,69 — le CACI explique 70 % de la variance intra-pays, ce qui serait bon même avec 600 observations. Enfin, aucun indice de compétitivité IA n'existait avant : mieux vaut une première validation sur 60 observations qu'aucune validation du tout. L'extension à 25-30 pays et 10 ans est explicitement recommandée comme piste de recherche.

Q8. N'y a-t-il pas un problème d'endogénéité ? Les pays productifs investissent plus dans le compute.

Oui, et c'est reconnu (§A.6). La causalité inverse est le principal risque : les pays à forte productivité IA accumulent plus de compute, créant un biais ascendant sur β. C'est pourquoi le modèle à effets fixes (préféré par le test de Hausman) absorbe les caractéristiques invariantes des pays, et les effets temporels capturent les tendances communes.

Mais nous ne prétendons pas démontrer la causalité stricte — ce serait prématuré avec ce panel. Ce que nous démontrons, c'est une association robuste et statistiquement significative, cohérente avec la théorie (Bresnahan & Trajtenberg, Brynjolfsson et al.), et que le CACI a un pouvoir prédictif réel. Pour la causalité formelle, nous proposons une approche par variables instrumentales (instrumenter le CACI par la dotation en énergie nucléaire, exogène à la productivité IA) comme piste prioritaire.

Q9. Le ratio 7-12:1 US/EU est-il crédible ? Ça semble énorme.

C'est énorme — et c'est précisément le message. Mais le ratio est convergent avec quatre sources indépendantes. Le capex Big Tech US ($675B) vs investissements EU IA (~40B€) donne un ratio de ~17:1 en flux d'investissement (McKinsey, janvier 2026). Le coût du TFlop pour le training est de 0,5 $/TFlop aux US contre 1,2-1,8 $/TFlop en EU (Bruegel/Epoch AI), soit un ratio de 2,4-3,6:1 sur les seuls coûts. Le compute installé (US : 75 GW IT load, EU : ~35 GW selon CFG Europe) donne un ratio brut de ~2:1, qui monte à 7-12:1 une fois normalisé par le PIB et ajusté du coût énergétique.

Si vous pensez que 7-12:1 est trop élevé, c'est probablement parce que le décalage réel est plus grand que ce que l'intuition suggère. C'est précisément ce que le CACI rend visible.

PARTIE III — MÉTHODOLOGIE

Q10. Pourquoi des scénarios plutôt qu'une modélisation économétrique classique ?

Pour trois raisons fondamentales (détaillées au §2.1). D'abord, les variables clés sont politiques et discrétionnaires : la décision de Trump d'imposer ou non des quotas GPU à l'Europe ne se modélise pas par une régression. Ensuite, les interactions sont non-linéaires et systémiques : une restriction GPU modifie en cascade les flux d'investissement énergétique, la localisation des data centers, et la structure concurrentielle de secteurs entiers. Enfin, les données de compute sont partiellement confidentielles — il n'existe pas de base de données publique des FLOPs par pays.

La méthode des scénarios (Schwartz, 1991 ; Shell methodology) est précisément conçue pour ces situations de forte incertitude politique et technologique. Ce n'est pas un choix par défaut : c'est le choix méthodologique le plus rigoureux pour ce type de problème. L'annexe économétrique (panel CACI) fournit le complément quantitatif.

Q11. Vos données de compute sont-elles fiables ? Il n'existe pas de recensement officiel.

Pas totalement, et c'est dit explicitement (§2.4.4). Les données de FLOPs installés par pays sont des estimations, pas des recensements. Nous utilisons quatre sources croisées : Epoch AI pour le compute de training, Hawkins et al. pour l'infrastructure cloud, CFG Europe pour la capacité européenne, et le Top500 pour le HPC public. La triangulation réduit le biais mais ne l'élimine pas.

Mais le CACI est conçu pour être comparatif (ratio entre régions), pas absolu. Si nous sous-estimons le compute US de 20 % et le compute EU de 20 %, le ratio reste le même. Les erreurs systématiques s'annulent en mode comparatif. C'est un choix de design : le CACI mesure des ordres de grandeur, pas des valeurs absolues. Et même avec une marge d'erreur de 30 %, un ratio de 7-12:1 reste un gap structurel massif.

Q12. Pourquoi n'avez-vous pas mené d'entretiens ou d'enquêtes terrain ?

C'est une limite assumée, et une piste de recherche explicitement recommandée. L'étude est fondée sur des sources secondaires (rapports institutionnels, données publiques, littérature académique) et sur la construction d'un outil original (le CACI). Un terrain qualitatif (entretiens avec des décideurs d'entreprises, des responsables de politiques industrielles, des opérateurs de data centers) renforcerait considérablement la validité — en particulier pour les scénarios et les recommandations.

Le choix a été de privilégier l'intégration analytique et la construction d'outils de mesure sur une seule personne, en réservant le terrain pour une phase ultérieure ou une extension collaborative. C'est un arbitrage classique en recherche : profondeur analytique vs largeur empirique.

Q13. Les sources McKinsey, Deloitte, Accenture ne sont-elles pas biaisées ?

Oui, et c'est explicitement reconnu (§2.2.3). Les cabinets de conseil ont un biais d'optimisme systématique — ils ont intérêt à surévaluer les marchés pour justifier leurs missions de conseil. C'est pourquoi nous ne nous appuyons jamais sur une source unique. Chaque chiffre est triangulé avec au moins deux sources d'orientations différentes : IEA (institutionnel), McKinsey (industry), Bruegel (académique/think tank), BIS (réglementaire).

De plus, les biais des consultants sont plutôt conservateurs sur les risques géopolitiques (ils n'ont pas intérêt à effrayer leurs clients). Si McKinsey estime déjà un compute gap important, la réalité est probablement pire, pas meilleure.

PARTIE IV — RÉSULTATS ET SCÉNARIOS

Q14. Qu'est-ce que le « protectionnisme à trois étages » et en quoi est-ce nouveau ?

Premier étage : les export controls (hérités de Biden, transformés par Trump) qui segmentent le monde en trois tiers d'accès au compute. C'est le mécanisme le plus documenté.

Deuxième étage : les tarifs Section 232 (25 % sur les GPU IA, janvier 2026). C'est l'innovation Trump : pour la première fois, les semi-conducteurs IA sont traités comme un enjeu de sécurité nationale comparable à l'acier. L'exemption domestique crée un différentiel de coût direct entre entreprises US et non-US.

Troisième étage : la gravité capitalistique ($675B de capex annuel, investissements japonais de $550B, fonds émiratis qui convergent tous vers le sol US). Ce n'est plus du protectionnisme réglementaire — c'est un effet auto-renforçant qui concentre le compute sans intervention supplémentaire.

La nouveauté est l'empilement cumulatif : chaque étage amplifie les précédents. Les export controls limitent l'offre, les tarifs augmentent le coût, la gravité capitalistique attire les investissements. Le résultat est un système auto-renforçant.

Q15. Quel est le scénario que vous estimez le plus probable ?

Le Scénario A (« Protectionnisme graduel, réponse EU fragmentée »), estimé à ~45-50 % de probabilité. Les US maintiennent les restrictions actuelles sans durcissement majeur envers les alliés Tier 1, l'Europe réagit de manière déclaratoire sans investissements massifs, et le compute gap s'élargit progressivement jusqu'au point de basculement 2028.

C'est le scénario du « business as usual aggravé » — pas de rupture dramatique, mais une érosion structurelle de la compétitivité européenne. Le risque est précisément qu'il soit indolore à court terme et irréversible à moyen terme. C'est le scénario contre lequel les recommandations du Chapitre VII sont conçues.

Q16. Qu'est-ce que le « point de basculement 2028 » et pourquoi cette date ?

2028 est la convergence de trois contraintes simultanées, chacune documentée par des sources indépendantes. D'abord, la saturation énergétique EU : selon l'IEA et RTE, la demande électrique des data centers européens atteindra la capacité de réseau disponible vers 2028, créant des goulots d'étranglement physiques. Ensuite, le pic de demande compute : les modèles de frontière 2028-2029 nécessiteront 10× plus de compute que GPT-4, selon Epoch AI. Enfin, la montée en puissance de la robotique IA (+20-30 % de demande énergie industrielle), non encore intégrée dans les projections officielles.

Après 2028, les investissements en infrastructure deviennent non plus un avantage mais une condition de survie — et les délais de construction (3-5 ans pour un data center, 8-12 ans pour un réacteur nucléaire) signifient que les décisions prises en 2026-2027 détermineront la position de 2030.

Q17. La Chine ne construit-elle pas un écosystème alternatif viable ? Le protectionnisme US échoue-t-il ?

C'est un des paradoxes majeurs identifiés par l'étude. Oui, les restrictions accélèrent la construction d'un écosystème chinois alternatif (DeepSeek, Huawei Ascend, SMIC). DeepSeek a démontré qu'on pouvait approcher les performances frontière avec des architectures plus efficientes et des GPU moins avancés. Huawei développe une gamme complète de puces IA. SMIC progresse sur les nœuds 7nm sans EUV.

Mais cet écosystème reste en retard de 2-3 générations de GPU, et la production de masse à la frontière technologique (sub-5nm) nécessite ASML, dont la Chine est exclue. Le protectionnisme US ne « échoue » pas — il produit des effets différents de ceux annoncés : au lieu d'un monde unipolaire dominé par les US, il crée un monde fragmenté en blocs technologiques. C'est précisément ce que l'étude démontre.

Q18. Pourquoi avoir ajouté le Brésil et l'Asie alors que le focus initial était US/EU ?

Parce que le protectionnisme IA n'opère pas en vase clos. L'étude initiale (US/EU) ne capturait pas les effets de second ordre : où vont les investissements chinois rejetés des US ? Vers le Brésil (data center TikTok $38B), vers l'ASEAN, vers l'Afrique. Comment réagissent les alliés asiatiques ? Le Japon investit $550B aux US (co-financement de la suprématie), la Corée mise sur la mémoire HBM, Taïwan est le pivot de tout le système via TSMC.

Limiter l'analyse à US/EU aurait été comme analyser la Guerre Froide en ne regardant que Washington et Moscou sans s'intéresser au Tiers-Monde. Les zones Tier 2 et Tier 3 sont le terrain où la compétition se joue concrètement.

PARTIE V — RECOMMANDATIONS ET POLITIQUE

Q19. Vos recommandations sont-elles réalistes ? 200 milliards d'investissement EU, c'est énorme.

C'est énorme en valeur absolue mais modeste en proportion. Le capex Big Tech US est de $675B par an ; $200B sur 5 ans pour l'UE représente ~6 % de cet effort annuel. Le plan Draghi (septembre 2024) estimait le déficit d'investissement numérique européen à $700B/an. Nos recommandations sont en fait conservatrices par rapport au diagnostic Draghi.

Par ailleurs, les recommandations sont structurées en trois horizons temporels réalistes : court terme (2026-2027) avec des mesures immédiates (contrats GPU, Compute Zones), moyen terme (2027-2029) avec des investissements programmés (AI Factories, EPR 2), et long terme (2029-2030+) avec des transformations structurelles (SMR nucléaires, autonomie hardware). Chaque horizon a des mesures finançables et politiquement faisables.

Q20. L'option « autonomie stratégique ciblée » n'est-elle pas un compromis mou ?

Non — c'est le seul positionnement réaliste et c'est ce qui le rend fort. L'étude démontre que deux extrêmes sont inviables. L'intégration subordonnée (modèle Japon : co-financer la suprématie US) sacrifie la capacité de choix. La confrontation souverainiste (modèle Chine : tout reconstruire en autarcie) est irréaliste à horizon 2030 — l'Europe n'a ni les fonderies, ni le compute, ni le temps.

L'autonomie stratégique ciblée consiste à être souverain sur les segments d'avantage comparatif (nucléaire pour l'énergie, ASML pour la litho, Mistral pour les modèles, AI Act pour la régulation) tout en maintenant l'interopérabilité avec l'écosystème US. L'objectif n'est pas l'autarcie mais la capacité de choix — pouvoir dire non à une condition d'accès inacceptable. C'est la définition même de la souveraineté.

Q21. La France a-t-elle vraiment un avantage nucléaire pour l'IA ?

Oui, et c'est mesurable. La France produit ~70 % de son électricité à partir du nucléaire, à un coût de ~42 €/MWh (tarif ARENH 2024) contre 90-145 €/MWh en Allemagne et 55 $/MWh aux US. Pour un data center de 100 MW, ça représente une économie de 40-80 M€/an sur le seul poste énergie.

EDF a identifié quatre sites industriels totalisant 2 GW pour des data centers IA, avec l'initiative Nuclear for AI (250 MW d'ici fin 2026). Les 6 EPR 2 programmés (Penly, Bugey, 9 900 MW, construction 2027) ajouteront une capacité dédiable. Et la France est le seul pays EU avec un programme SMR actif (NUWARD, Newcleo, Stellaria) qui pourrait fournir de l'énergie dédiée aux data centers d'ici 2030-2032.

Le CACI capture cet avantage : le facteur E(r) de la France est nettement inférieur à celui de l'Allemagne ou des Pays-Bas, ce qui remonte son CACI relatif malgré un compute brut (F) inférieur.

PARTIE VI — LIMITES, CRITIQUES ET DÉFENSE

Q22. L'environnement réglementaire change très vite. Votre analyse n'est-elle pas déjà obsolète ?

L'AI Diffusion Rule de Biden a été abrogée en mai 2025. La Section 232 de Trump date de janvier 2026. La règle finale BIS sera mise à jour d'ici juillet 2026. Oui, ça bouge vite — et c'est pourquoi l'étude utilise des scénarios et non des prédictions.

Les mécanismes structurels identifiés (concentration du compute, différentiel énergétique, gravité capitalistique) sont indépendants de la réglementation spécifique du moment. Même si la Section 232 est modifiée, le capex de $675B et le compute gap restent. L'analyse est conçue pour survivre aux changements réglementaires : les scénarios couvrent un spectre allant du relâchement au durcissement. C'est la vertu de la méthode scénarielle.

Q23. DeepSeek n'invalide-t-il pas votre thèse ? On peut faire de l'IA avec moins de compute.

DeepSeek est intégré à l'analyse (Chapitre III et V). Oui, DeepSeek démontre que des gains d'efficience architecturale peuvent partiellement compenser un déficit de compute brut. Mais trois nuances essentielles s'imposent.

D'abord, DeepSeek a été entraîné sur des GPU Nvidia A100 accumulées avant les restrictions — la capacité de réplication est limitée par les stocks et les approvisionnements alternatifs (Huawei Ascend). Ensuite, l'IEA (2025) documente un effet rebond de Jevons : les gains d'efficience augmentent les usages, qui absorbent les gains et relancent la demande de compute. Enfin, DeepSeek concerne le training ; l'inférence à grande échelle (milliards de requêtes/jour) reste proportionnelle à la capacité de compute installée.

DeepSeek ne contredit pas la thèse — il l'enrichit en montrant que la compétition ne se joue pas uniquement sur le compute brut, mais aussi sur l'efficience architecturale. Ce qui ne change pas, c'est que l'accès au compute reste le facteur limitant à l'échelle systémique.

Q24. Votre étude ne manque-t-elle pas de terrain empirique pour un niveau doctoral ?

Oui — et c'est dit explicitement dans les limites (Conclusion §4). L'absence d'entretiens avec des décideurs, d'enquête terrain, et de données micro (firm-level panel) est la principale limite pour un positionnement doctoral complet. C'est pourquoi l'étude se positionne entre le M2 Recherche de haut niveau et la thèse professionnelle (DBA), avec une contribution méthodologique (le CACI) qui constitue un apport de niveau doctoral.

Le choix assumé a été de concentrer l'effort sur l'intégration analytique, la construction d'un outil original, et sa validation économétrique — ce qui est en soi une contribution substantielle. Le terrain est explicitement recommandé comme prolongement prioritaire.

Q25. Si vous aviez six mois de plus, que feriez-vous ?

Trois prolongements prioritaires. D'abord, 20-30 entretiens semi-directifs avec des décideurs (directeurs data center de CAC 40, responsables compute à la DGE, opérateurs cloud,

responsables politique industrielle UE) pour valider les scénarios et enrichir les recommandations.

Ensuite, l'extension du panel CACI à 25-30 pays sur 10 ans (2015-2024), avec des données d'entreprises (firm-level) pour tester le CACI au niveau microéconomique et adresser l'endogénéité par variables instrumentales (dotation nucléaire, proximité fonderies).

Enfin, l'extension géographique à l'Afrique — le continent absent de l'étude, où la compétition US-Chine pour le compute prend des formes spécifiques (câbles sous-marins, data centers en Afrique du Sud et au Nigeria, programme Starlink). C'est le prochain terrain de la géopolitique du compute.

« La bataille pour le compute est la bataille pour la souveraineté économique. »

— Fabrice Pizzi, AI for Americans First, 2026