

AI FOR AMERICANS FIRST

FAQ 1

Análise Geoestratégica e Econômica Integrada

part I

Fabrice Pizzi

Universidade Sorbonne

Mestrado em Inteligência Econômica — Intelligence Warfare

75% do compute IA mundial = EUA | \$675B capex EUA 2026 | razão 7–12× EUA/UE

Paris — Fevereiro 2026

7 capítulos • 4 cenários prospectivos • 3 zonas geográficas

Palavras-chave: inteligência artificial, protecionismo tecnológico, semicondutores, controles de exportação, compute soberano, geopolítica da IA, França, Estados Unidos, China

PARTE I — A TESE E SEU POSICIONAMENTO

P1. Qual é a sua tese central, em uma frase?

Os Estados Unidos construíram, sob Trump 2.0, uma arquitetura protecionista de três níveis (controles de exportação, tarifas Seção 232, gravidade capitalística) que transforma o acesso ao compute de IA em alavanca geopolítica, criando uma razão de competitividade EUA/UE de 7 a 12:1 mensurável pelo índice CACI — razão que, sem uma resposta europeia estruturada até 2028, se tornará irreversível.

P2. O que o senhor traz de novo em relação à literatura existente?

Quatro contribuições originais. Primeiro, a integração analítica de quatro dimensões tratadas separadamente na literatura: trajetórias energéticas dos data centers, mercado de semicondutores, distribuição do compute de IA e cronologia regulatória dos EUA. Nenhum estudo havia cruzado essas quatro dimensões.

Segundo, a proposição do CACI — o primeiro índice formalizado de competitividade ajustada ao compute, validado econometricamente em um painel de 12 países. Nem Hawkins et al. (Oxford), nem o FMI, nem o Fed haviam construído tal indicador.

Terceiro, a demonstração de que o protecionismo de IA produz efeitos paradoxais sistêmicos: aceleração do ecossistema alternativo chinês, pressão dos países Tier 2 em direção à China, cofinanciamento da supremacia dos EUA pelos próprios aliados.

Quarto, a análise comparativa de três zonas geográficas (Europa, América do Sul, Ásia) mostrando trajetórias de dependência estruturalmente diferentes e irreduzíveis a um modelo único.

P3. Por que este tema é pertinente em 2026?

Porque três eventos convergem entre janeiro e julho de 2026, tornando a análise urgente. Primeiro, a Seção 232 de Trump (15 de janeiro de 2026) impõe 25% de tarifas sobre GPUs de IA avançadas — é a primeira vez que semicondutores de IA são tratados como questão de segurança nacional comparável ao aço. Em seguida, o capex acumulado das Big Tech dos EUA atinge US\$ 675 bilhões em 2026, criando uma concentração sem precedentes de compute em solo americano. Por fim, a regra final do BIS deve ser atualizada até julho de 2026, com potencial de endurecimento em relação aos aliados Tier 1.

O tema não é teórico: determina concretamente se uma PME industrial francesa poderá ou não acessar o compute necessário para permanecer competitiva em 2028–2030.

P4. Por que vir da cibersegurança para tratar deste tema? Qual legitimidade?

Precisamente porque a cibersegurança me formou para pensar em termos de cadeias de dependência, superfícies de ataque e pontos de controle. O protecionismo de IA opera pelos mesmos mecanismos das ameaças cibernéticas: identificação de gargalos (TSMC para chips, ASML para litografia, Nvidia para GPUs), exploração de posições de monopólio e weaponization das interdependências — conceito formalizado por Farrell & Newman (2019) que aplico ao compute.

Meu Mestrado em Inteligência Econômica, orientado para Intelligence Warfare, fornece o quadro analítico complementar: a IE é a capacidade de compreender e antecipar as relações de força econômicas, que é exatamente o objeto deste estudo. A dupla formação IE/ciber é, na verdade, uma vantagem: poucos pesquisadores cruzam esses dois mundos.

PARTE II — O ÍNDICE CACI

P5. O CACI não existe em nenhuma outra publicação. Isso não é um problema?

É o contrário: é a contribuição. O CACI não surgiu do nada — é a formalização de uma necessidade explicitamente identificada por seis correntes da literatura. Hawkins et al. (Oxford, 2025) medem o compute por região, mas não constroem um índice. O Fed Board (outubro de 2025) observa textualmente que “a ausência de um indicador unificado de capacidade de IA torna as comparações transatlânticas difíceis”. Martens/Bruegel (2024) identifica barreiras, mas não propõe nenhuma métrica. O FMI tem um AI Preparedness Index sem componente de compute.

O CACI é a peça que faltava e que todos pediam. O fato de ser novo é sua razão de ser, não sua fraqueza.

P6. Como o senhor justifica a fórmula multiplicativa em vez de aditiva?

A forma $CACI(r) = [F(r) \times E(r)^{-1}] / [PIB(r) \times L(r)]$ é multiplicativa porque os fatores são complementares, não substituíveis. Compute abundante sem capital humano para explorá-lo não produz competitividade (caso dos Emirados). Talento sem acesso ao compute tampouco (caso de certos países africanos com alto capital humano STEM).

A validação econométrica confirma essa escolha: o modelo com ponderações alternativas (combinação linear F:40%, E:25%, L:20%, Reg:15%) perde sua significância ($p = 0,40$), enquanto a fórmula multiplicativa dá $\beta = 0,25$ significativo a 1%. Os dados encerram o debate teórico.

P7. Seu painel tem apenas 60 observações. Não é muito pouco?

É uma limitação assumida e explicitamente discutida (§A.6). Com 12 países \times 5 anos, os graus de liberdade são modestos — por isso utilizamos erros-padrão robustos clusterizados e testamos sistematicamente a sensibilidade à exclusão de outliers.

Mas três argumentos relativizam essa limitação. Primeiro, o coeficiente β do CACI é significativo a 1% nas três especificações (OLS, EF, EA), o que é notável para um painel tão compacto. Segundo, o R^2 within atinge 0,69 — o CACI explica 70% da variância intra-país, o que seria bom mesmo com 600 observações. Terceiro, nenhum índice de competitividade em IA existia antes: uma primeira validação com 60 observações é melhor do que nenhuma validação. A extensão para 25–30 países e 10 anos é explicitamente recomendada como pista de pesquisa.

P8. Não há um problema de endogeneidade? Países produtivos investem mais em compute.

Sim, e é reconhecido (§A.6). A causalidade reversa é o principal risco: países com alta produtividade em IA acumulam mais compute, criando um viés ascendente sobre β . Por isso, o modelo de efeitos fixos (preferido pelo teste de Hausman) absorve as características invariantes dos países, e os efeitos temporais capturam as tendências comuns.

Mas não pretendemos demonstrar causalidade estrita — seria prematuro com este painel. O que demonstramos é uma associação robusta e estatisticamente significativa, coerente com a teoria (Bresnahan & Trajtenberg, Brynjolfsson et al.), e que o CACI tem poder

preditivo real. Para a causalidade formal, propomos uma abordagem por variáveis instrumentais (instrumentar o CACI pela dotação em energia nuclear, exógena à produtividade de IA) como pista prioritária.

P9. A razão 7–12:1 EUA/UE é credível? Parece enorme.

É enorme — e essa é precisamente a mensagem. Mas a razão é convergente com quatro fontes independentes. O capex das Big Tech dos EUA (US\$ 675 bilhões) vs investimentos da UE em IA (~€40 bilhões) dá uma razão de ~17:1 em fluxos de investimento (McKinsey, janeiro de 2026). O custo por TFlop para treinamento é de US\$ 0,5/TFlop nos EUA contra US\$ 1,2–1,8/TFlop na UE (Bruegel/Epoch AI), uma razão de 2,4–3,6:1 apenas nos custos. O compute instalado (EUA: 75 GW de carga TI, UE: ~35 GW segundo CFG Europe) dá uma razão bruta de ~2:1, que sobe para 7–12:1 após normalização pelo PIB e ajuste pelo custo energético.

Se você acha que 7–12:1 é alto demais, provavelmente é porque a defasagem real é maior do que a intuição sugere. É precisamente isso que o CACI torna visível.

PARTE III — METODOLOGIA

P10. Por que cenários em vez de modelagem econométrica clássica?

Por três razões fundamentais (detalhadas no §2.1). Primeiro, as variáveis-chave são políticas e discricionárias: a decisão de Trump de impor ou não cotas de GPU à Europa não pode ser modelada por regressão. Segundo, as interações são não lineares e sistêmicas: uma restrição de GPU modifica em cascata os fluxos de investimento energético, a localização dos data centers e a estrutura competitiva de setores inteiros. Terceiro, os dados de compute são parcialmente confidenciais — não existe banco de dados público de FLOPs por país.

O método de cenários (Schwartz, 1991; metodologia Shell) é precisamente projetado para essas situações de alta incerteza política e tecnológica. Não é uma escolha por falta de opção: é a escolha metodológica mais rigorosa para esse tipo de problema. O apêndice econométrico (painel CACI) fornece o complemento quantitativo.

P11. Seus dados de compute são confiáveis? Não existe censo oficial.

Não totalmente, e isso é dito explicitamente (§2.4.4). Os dados de FLOPs instalados por país são estimativas, não censos. Utilizamos quatro fontes cruzadas: Epoch AI para compute de treinamento, Hawkins et al. para infraestrutura de nuvem, CFG Europe para capacidade europeia e Top500 para HPC público. A triangulação reduz o viés, mas não o elimina.

Mas o CACI é projetado para ser comparativo (razão entre regiões), não absoluto. Se subestimamos o compute dos EUA em 20% e o compute da UE em 20%, a razão permanece a mesma. Erros sistemáticos se anulam no modo comparativo. É uma escolha de design: o CACI mede ordens de grandeza, não valores absolutos. E mesmo com uma margem de erro de 30%, uma razão de 7–12:1 permanece um gap estrutural massivo.

P12. Por que não realizou entrevistas ou pesquisas de campo?

É uma limitação assumida e uma pista de pesquisa explicitamente recomendada. O estudo é baseado em fontes secundárias (relatórios institucionais, dados públicos, literatura acadêmica) e na construção de uma ferramenta original (o CACI). Um trabalho de campo qualitativo (entrevistas com decisores empresariais, responsáveis por políticas industriais, operadores de data centers) reforçaria consideravelmente a validade — particularmente para os cenários e recomendações.

A escolha foi priorizar a integração analítica e a construção de ferramentas de medição por um único pesquisador, reservando o campo para uma fase posterior ou uma extensão colaborativa. É um trade-off clássico em pesquisa: profundidade analítica vs. amplitude empírica.

P13. As fontes McKinsey, Deloitte, Accenture não são enviesadas?

Sim, e isso é explicitamente reconhecido (§2.2.3). As consultorias têm um viés sistêmico de otimismo — têm interesse em supervalorizar mercados para justificar seus contratos de consultoria. Por isso, nunca nos apoiamos em uma única fonte. Cada dado é triangulado com pelo menos duas fontes de orientações diferentes: IEA (institucional), McKinsey (indústria), Bruegel (acadêmico/think tank), BIS (regulatório).

Além disso, os vieses dos consultores tendem a ser conservadores sobre riscos geopolíticos (não têm interesse em assustar seus clientes). Se a McKinsey já estima um gap de compute significativo, a realidade provavelmente é pior, não melhor.

PARTE IV — RESULTADOS E CENÁRIOS

P14. O que é o “protecionismo de três níveis” e o que há de novo?

Primeiro nível: controles de exportação (herdados de Biden, transformados por Trump) que segmentam o mundo em três níveis de acesso ao compute. É o mecanismo mais documentado.

Segundo nível: tarifas Seção 232 (25% sobre GPUs de IA, janeiro de 2026). É a inovação Trump: pela primeira vez, semicondutores de IA são tratados como questão de segurança nacional comparável ao aço. A isenção doméstica cria um diferencial de custo direto entre empresas americanas e não americanas.

Terceiro nível: gravidade capitalística (US\$ 675 bilhões de capex anual, investimentos japoneses de US\$ 550 bilhões, fundos dos Emirados convergindo para solo americano). Não é mais protecionismo regulatório — é um efeito autorreforçante que concentra o compute sem intervenção adicional.

A novidade é o empilhamento cumulativo: cada nível amplifica os anteriores. Controles de exportação limitam a oferta, tarifas aumentam o custo, gravidade capitalística atrai investimentos. O resultado é um sistema autorreforçante.

P15. Qual cenário o senhor estima como mais provável?

O Cenário A (“Protecionismo gradual, resposta fragmentada da UE”), estimado em ~45–50% de probabilidade. Os EUA mantêm as restrições atuais sem endurecimento major em relação aos aliados Tier 1, a Europa reage de forma declaratória sem investimentos massivos, e o gap de compute se amplia progressivamente até o ponto de virada de 2028.

É o cenário do “business as usual agravado” — sem ruptura dramática, mas uma erosão estrutural da competitividade europeia. O risco é precisamente que seja indolor no curto prazo e irreversível no médio prazo. É o cenário contra o qual as recomendações do Capítulo VII são desenhadas.

P16. O que é o “ponto de virada 2028” e por que essa data?

2028 é a convergência de três restrições simultâneas, cada uma documentada por fontes independentes. Primeiro, saturação energética da UE: segundo a IEA e RTE, a demanda de eletricidade dos data centers europeus atingirá a capacidade de rede disponível por volta de 2028, criando gargalos físicos. Segundo, pico de demanda de compute: modelos de fronteira de 2028–2029 necessitarão 10× mais compute que o GPT-4, segundo Epoch AI. Terceiro, a escala da robótica de IA (+20–30% de demanda energética industrial), ainda não integrada nas projeções oficiais.

Após 2028, investimentos em infraestrutura se tornam não mais uma vantagem, mas uma condição de sobrevivência — e os prazos de construção (3–5 anos para um data center, 8–12 anos para um reator nuclear) significam que as decisões tomadas em 2026–2027 determinarão a posição de 2030.

P17. A China não está construindo um ecossistema alternativo viável? O protecionismo dos EUA está falhando?

É um dos maiores paradoxos identificados pelo estudo. Sim, as restrições aceleram a construção de um ecossistema alternativo chinês (DeepSeek, Huawei Ascend, SMIC). O DeepSeek demonstrou que era possível se aproximar do desempenho de fronteira com arquiteturas mais eficientes e GPUs menos avançadas. A Huawei desenvolve uma gama completa de chips de IA. A SMIC progride nos nós de 7nm sem EUV.

Mas esse ecossistema permanece 2–3 gerações de GPU atrás, e a produção em massa na fronteira tecnológica (sub-5nm) requer ASML, da qual a China está excluída. O protecionismo dos EUA não está “falhando” — produz efeitos diferentes dos anunciados: em vez de um mundo unipolar dominado pelos EUA, cria um mundo fragmentado em blocos tecnológicos. É precisamente o que o estudo demonstra.

P18. Por que adicionar o Brasil e a Ásia quando o foco inicial era EUA/UE?

Porque o protecionismo de IA não opera em vaso comunicante. O estudo inicial (EUA/UE) não capturava os efeitos de segunda ordem: para onde vão os investimentos chineses rejeitados dos EUA? Para o Brasil (data center TikTok US\$ 38 bilhões), para a ASEAN, para a África. Como reagem os aliados asiáticos? O Japão investe US\$ 550 bilhões nos EUA (cofinanciamento da supremacia), a Coreia aposta na memória HBM, Taiwan é o pivô de todo o sistema via TSMC.

Limitar a análise a EUA/UE seria como analisar a Guerra Fria olhando apenas para Washington e Moscou sem considerar o Terceiro Mundo. As zonas Tier 2 e Tier 3 são o terreno onde a competição se desenrola concretamente.

PARTE V — RECOMENDAÇÕES E POLÍTICA

P19. Suas recomendações são realistas? €200 bilhões de investimento da UE é enorme.

É enorme em valor absoluto, mas modesto em proporção. O capex das Big Tech dos EUA é de US\$ 675 bilhões por ano; US\$ 200 bilhões em 5 anos para a UE representa ~6% desse esforço anual. O Plano Draghi (setembro de 2024) estimava o déficit de investimento digital europeu em US\$ 700 bilhões/ano. Nossas recomendações são, na verdade, conservadoras em comparação com o diagnóstico Draghi.

Além disso, as recomendações são estruturadas em três horizontes temporais realistas: curto prazo (2026–2027) com medidas imediatas (contratos de GPU, Compute Zones), médio prazo (2027–2029) com investimentos programados (AI Factories, EPR 2), e longo prazo (2029–2030+) com transformações estruturais (SMRs nucleares, autonomia em hardware). Cada horizonte tem medidas financeáveis e politicamente viáveis.

P20. A opção “autonomia estratégica direcionada” não é um compromisso fraco?

Não — é o único posicionamento realista, e é isso que o torna forte. O estudo demonstra que dois extremos são inviáveis. A integração subordinada (modelo Japão: cofinanciar a supremacia dos EUA) sacrifica a capacidade de escolha. O confronto soberanista (modelo China: reconstruir tudo em autarquia) é irrealista até 2030 — a Europa não tem nem as fundições, nem o compute, nem o tempo.

A autonomia estratégica direcionada consiste em ser soberano nos segmentos de vantagem comparativa (nuclear para energia, ASML para litografia, Mistral para modelos, AI Act para regulação) mantendo a interoperabilidade com o ecossistema dos EUA. O objetivo não é autarquia, mas capacidade de escolha — poder dizer não a uma condição de acesso inaceitável. Essa é a própria definição de soberania.

P21. A França realmente tem uma vantagem nuclear para a IA?

Sim, e é mensurável. A França produz ~70% de sua eletricidade a partir de energia nuclear, a um custo de ~€42/MWh (tarifa ARENH 2024) contra €90–145/MWh na Alemanha e US\$ 55/MWh nos EUA. Para um data center de 100 MW, isso representa uma economia de €40–80 milhões/ano apenas no custo de energia.

A EDF identificou quatro sítios industriais totalizando 2 GW para data centers de IA, com a iniciativa Nuclear for AI (250 MW até o final de 2026). Os 6 reatores EPR 2 planejados (Penly, Bugey, 9.900 MW, construção em 2027) adicionarão capacidade dedicável. E a França é o único país da UE com um programa SMR ativo (NUWARD, Newcleo, Stellaris) que poderia fornecer energia dedicada a data centers até 2030–2032.

O CACI captura essa vantagem: o fator E(r) da França é significativamente inferior ao da Alemanha ou dos Países Baixos, o que eleva seu CACI relativo apesar de um compute bruto (F) inferior.

PARTE VI — LIMITAÇÕES, CRÍTICAS E DEFESA

P22. O ambiente regulatório muda muito rápido. Sua análise já não está obsoleta?

A AI Diffusion Rule de Biden foi revogada em maio de 2025. A Seção 232 de Trump data de janeiro de 2026. A regra final do BIS será atualizada até julho de 2026. Sim, as coisas mudam rápido — e é por isso que o estudo usa cenários e não previsões.

Os mecanismos estruturais identificados (concentração de compute, diferencial energético, gravidade capitalística) são independentes da regulamentação específica do momento.

Mesmo que a Seção 232 seja modificada, o capex de US\$ 675 bilhões e o gap de compute permanecem. A análise é projetada para sobreviver às mudanças regulatórias: os cenários cobrem um espectro que vai do relaxamento ao endurecimento. É a virtude do método cenearial.

P23. O DeepSeek não invalida sua tese? É possível fazer IA com menos compute.

O DeepSeek está integrado à análise (Capítulos III e V). Sim, o DeepSeek demonstra que ganhos de eficiência arquitetural podem compensar parcialmente um déficit de compute bruto. Mas três nuances essenciais se impõem.

Primeiro, o DeepSeek foi treinado com GPUs Nvidia A100 acumuladas antes das restrições — a capacidade de replicação é limitada pelos estoques e suprimentos alternativos (Huawei Ascend). Segundo, a IEA (2025) documenta um efeito rebote de Jevons: ganhos de eficiência aumentam os usos, que absorvem os ganhos e relançam a demanda de compute. Terceiro, o DeepSeek diz respeito ao treinamento; a inferência em grande escala (bilhões de requisições/dia) permanece proporcional à capacidade de compute instalada.

O DeepSeek não contradiz a tese — a enriquece mostrando que a competição não se joga unicamente no compute bruto, mas também na eficiência arquitetural. O que não muda é que o acesso ao compute permanece o fator limitante na escala sistêmica.

P24. Seu estudo não carece de campo empírico para um nível doutoral?

Sim — e está dito explicitamente nas limitações (Conclusão §4). A ausência de entrevistas com decisores, pesquisa de campo e dados micro (painel de empresas) é a principal limitação para um posicionamento doutoral completo. Por isso, o estudo se posiciona entre o Mestrado em Pesquisa de alto nível e a tese profissional (DBA), com uma contribuição metodológica (o CACI) que constitui um aporte de nível doutoral.

A escolha deliberada foi concentrar o esforço na integração analítica, na construção de uma ferramenta original e na validação econométrica — o que é, por si só, uma contribuição substancial. O trabalho de campo é explicitamente recomendado como extensão prioritária.

P25. Se tivesse mais seis meses, o que faria?

Três extensões prioritárias. Primeiro, 20–30 entrevistas semiestruturadas com decisores (diretores de data center do CAC 40, funcionários de compute da DGE, operadores de nuvem, responsáveis por política industrial da UE) para validar cenários e enriquecer recomendações.

Segundo, extensão do painel CACI para 25–30 países em 10 anos (2015–2024), com dados de empresas (nível de firma) para testar o CACI no nível microeconômico e abordar a endogeneidade por variáveis instrumentais (dotação nuclear, proximidade de fundições).

Terceiro, extensão geográfica à África — o continente ausente do estudo, onde a competição EUA-China pelo compute assume formas específicas (cabos submarinos, data centers na África do Sul e Nigéria, programa Starlink). É o próximo terreno da geopolítica do compute.

“A batalha pelo compute é a batalha pela soberania econômica.”

— Fabrice Pizzi, *AI for Americans First*, 2026