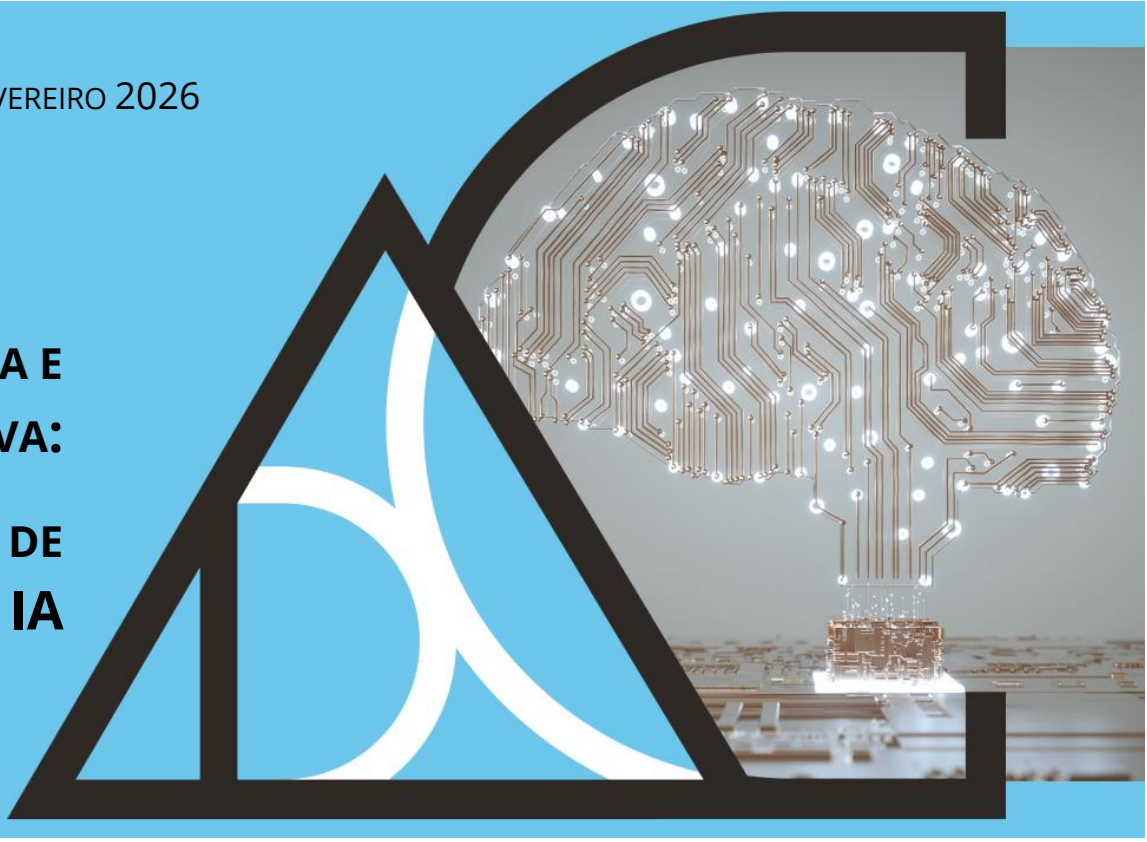


Concorrência e IA Generativa: Acesso a *Chips* de IA

A Autoridade da Concorrência (AdC) tem vindo a acompanhar os desenvolvimentos no setor de inteligência artificial (IA) generativa desde o final de 2022. A AdC publicou, em novembro de 2023, um *issues paper* sobre inteligência artificial[1] e, em 2024, iniciou uma série de *short papers* sobre o mesmo tema[2].

Este é o quarto *short paper* da série e examina questões de concorrência relacionadas com o acesso a *chips* utilizados no treino e na execução de modelos de IA.

I. Introdução

A capacidade de computação[3] e de memória são requisitos essenciais para a IA generativa, tanto na fase de treino como de inferência. Os principais modelos de IA competem em termos de desempenho, o que abrange simultaneamente a qualidade dos resultados gerados e o tempo necessário para os produzir. Para melhorar a qualidade dos resultados, os fornecedores de IA têm procurado escalar os seus modelos de diversas formas. Tal pode passar por aumentar a dimensão dos modelos, usar um maior volume de dados de treino, ampliar as janelas de contexto ou aumentar o tempo de treino ou de inferência[4]. Os

[1] Disponível aqui.

[2] Até ao momento foram publicados três *short papers*. O primeiro, de setembro de 2024, é sobre o acesso e a utilização de dados em IA generativa e as suas implicações para a concorrência (aqui). O segundo *short paper*, de dezembro de 2024, aborda questões relacionadas com o acesso a modelos de IA por parte de fornecedores de IA terceiros a jusante, e o papel da abertura de modelos de IA em promover a concorrência e a inovação (aqui). O terceiro *short paper*, de julho de 2025, incide sobre questões de concorrência relacionadas com o acesso a talento no setor de IA generativa (aqui).

[3] Entende-se por capacidade de computação a capacidade de processamento necessária para executar tarefas de IA, incluindo as operações matemáticas (como sejam multiplicações de matrizes em larga escala e funções associadas) que sustentam as fases de treino e de inferência dos modelos de IA. Abrange tanto os recursos de *hardware* (por exemplo, GPUs e ASICs) como a disponibilidade efetiva desses recursos ao longo do tempo.

[4] Como os ganhos decorrentes do aumento da escala na fase de pré-treino parecem estar a diminuir (ver, por exemplo, aqui e aqui), os fornecedores de IA têm recentemente intensificado a utilização de capacidade

fornecedores de IA têm procurado igualmente melhorar a eficiência dos sistemas de IA, nomeadamente inovando ao nível da arquitetura dos modelos[5]. A aposta em escalar os modelos e o aumento acentuado da procura por serviços de IA nos últimos anos traduziram-se num crescimento significativo da necessidade de recursos de computação e de memória de elevado desempenho.

O acesso a capacidade de computação depende de *hardware* especializado, designadamente de *chips* aceleradores de IA (*chips* de IA). Estes *chips* são otimizados para lidar com as operações essenciais dos modelos de IA[6], sendo significativamente mais eficientes do que os processadores tradicionais.

Os *chips* de IA mais amplamente utilizados são os GPUs[7], que são semicondutores especializados em processamento paralelo e essenciais para a execução dos modelos de IA. Atualmente, os fornecedores de IA fazem uso de centenas de milhares de GPUs de alto desempenho, maioritariamente fornecidas pela NVIDIA, para o treino e a execução dos seus modelos (*cf.* Caixa 1)[8]. Além dos GPUs, os fornecedores de IA podem recorrer também a ASICs[9], *chips* concebidos de forma personalizada para tarefas e aplicações específicas, em particular para inferência. Estes *chips* apresentam maior eficiência nessas tarefas (por exemplo, maior rapidez, menor custo ou menor consumo energético), embora sejam menos flexíveis[10]. Os TPUs (*Tensor Processing Units*) da Google, por exemplo, são ASICs utilizados para

computacional durante a fase de inferência (*inference-time compute*). Um exemplo são os chamados modelos de raciocínio (*reasoning models*). Em vez de produzirem diretamente a resposta final, estes modelos geram etapas intermediárias de raciocínio como parte do resultado. Este processo utiliza mais recursos computacionais durante a fase da inferência, mas tende a produzir respostas de maior qualidade. Este tipo de processamento exige igualmente maior capacidade de memória, o que pode criar oportunidades para *chips* alternativos aos existentes, nomeadamente os da NVIDIA, otimizados para executar grandes volumes de operações semelhantes em paralelo (Financial Times, "*How 'inference' is driving competition to Nvidia's AI chip dominance*", março de 2025, aqui).

[5] Um exemplo é a técnica *Mixture of Experts*, que ativa dinamicamente apenas as partes do modelo mais relevantes para processar um pedido específico, reduzindo a computação total necessária e os respetivos custos.

[6] Os sistemas de IA modernos são dominados por operações de álgebra linear paralelas (principalmente multiplicação de matrizes), complementadas por funções não lineares.

[7] *Graphics Processing Unit* – vulgo "placa gráfica".

[8] Ver, por exemplo, o *State of AI Report Compute Index*, que estima que os maiores fornecedores de IA podem estar a utilizar atualmente centenas de milhares de GPUs de alto desempenho, maioritariamente da NVIDIA (aqui). Outros relatórios afirmam que a xAI pode estar a utilizar 200.000 GPUs da NVIDIA, ou que a OpenAI planeia alcançar 1 milhão de GPUs até ao final de 2025 (aqui). Adicionalmente, dados da Epoch AI sobre *clusters* de GPUs revelam que empresas como a Meta AI, Oracle, DataVolt, OpenAI/Microsoft, Sesterce, xAI e Reliable Industries planeiam deter mais de 1 milhão de GPUs equivalentes ao NVIDIA H100, nos próximos 5 anos. Estes dados incluem os *clusters* de GPUs existentes e planeados, confirmados e prováveis, não duplicados e privados, de acordo com a Epoch AI. Ver mais em Pilz, Rahman, Sanders & Heim (2025), "*Data on GPU Clusters*" (aqui). Acedido a 12 de fevereiro de 2026.

[9] *Application-Specific Integrated Circuits*.

[10] Existe um *trade-off* entre generalização e especialização no *design* de *chips*. Num extremo do espectro, os processadores de uso geral, como os CPUs, oferecem maior flexibilidade e conseguem executar tarefas muito diversificadas, mas são menos eficientes em tarefas específicas. No outro extremo, os ASICs proporcionam melhor desempenho e eficiência para um conjunto restrito de tarefas, mas são menos flexíveis. Os GPUs equilibram este *trade-off*, uma vez que oferecem elevado desempenho em computação paralela e nas operações específicas exigidas pela IA, mantendo ainda um grau considerável de flexibilidade.

treinar e executar modelos como os que suportam o Gemini. De acordo com a informação disponível, os GPUs representaram cerca de 72% do mercado total de *chips* de IA em 2023, enquanto os ASICs representaram aproximadamente 22%[11].

A capacidade de computação também está fortemente concentrada. De acordo com uma base de dados da Epoch AI, as empresas com a maior capacidade de computação são a xAI, a Meta AI, a Google, a Microsoft, a Tesla e a NVIDIA, representando juntas cerca de 90% da capacidade de computação conhecida[12]. Adicionalmente, cerca de 75% da capacidade conhecida disponível é detida por empresas privadas[13], e o peso das entidades públicas tem vindo a diminuir nos últimos anos[14].

O desempenho dos modelos de IA também depende criticamente de acesso a capacidade de memória. Durante o treino e, especialmente, durante a inferência, os modelos necessitam de processar grandes volumes de dados de forma rápida e eficiente. Atrasos no acesso à memória podem degradar a velocidade dos modelos e comprometer aplicações em tempo real[15]. Para mitigar este impacto, os *chips* de IA integram memória de alta largura de banda (HBM) diretamente no pacote do chip. A HBM é um tipo avançado de memória que permite velocidades de acesso e de transferência de dados substancialmente superiores às da memória convencional utilizada em equipamentos de consumo.

As empresas acedem tipicamente a capacidade de computação e de memória através de fornecedores de serviços de *cloud*, que desempenham um papel central na cadeia de valor da IA. O custo de construção de uma infraestrutura computacional adequada ao desenvolvimento de IA é proibitivo para a maioria dos fornecedores de IA. Os fornecedores de serviços de *cloud* oferecem acesso escalável a recursos avançados de computação e memória, permitindo que os fornecedores de IA treinem e implementem modelos sem incorrer nos elevados custos de capital associados à aquisição e manutenção de *hardware* dedicado.

---

[11] De acordo com um relatório da "*The Information Network*", analista da indústria de semicondutores, de maio de 2025, conforme citado pelo *eeNews Europe*. As projeções para 2025 preveem que o peso dos ASICs no mercado aumente, sobretudo em detrimento dos GPUs.

[12] Informação disponível na base de dados da Epoch AI sobre *clusters* de GPUs. Estes dados incluem os *clusters* de GPUs existentes, confirmados, não duplicados e privados, de acordo com a Epoch AI. Incluindo os *clusters* de GPU planeados e confirmados, pela Epoch AI, as maiores empresas privadas em termos de capacidade de computação seriam a xAI, a Meta AI, a Google, a Amazon, a Microsoft e a NVIDIA. A capacidade de computação é baseada no máximo teórico de desempenho em 16-bit FLOP/s. Ver mais em Pilz et al. (2026).

[13] De acordo com a base de dados da Epoch AI sobre *clusters* de GPUs, cerca de 75% da capacidade de computação é detida por empresas privadas, 12% por entidades públicas e 13% por parcerias público-privadas. Incluindo *clusters* de GPU planeados, no entanto, 57% da capacidade de computação é detida por empresas privadas, 43% por parcerias público-privadas e menos de 1% por entidades públicas. Estes dados incluem os *clusters* de GPUs, existentes e planeados, confirmados e prováveis, não duplicados, de acordo com a Epoch AI. A capacidade de computação é baseada no máximo teórico de desempenho em 16-bit FLOP/s. Ver mais em Pilz et al. (2026).

[14] Ver Pilz, Sanders, Rahman & Heim (2025). "*Trends in AI Supercomputers*". Disponível aqui.

[15] O desempenho da memória não tem acompanhado os avanços na capacidade de computação. Segundo uma publicação especializada da indústria de IA, desde 2023, a capacidade de computação em IA aumentou cerca de 750% e as velocidades de processamento triplicaram, enquanto a largura de banda da memória cresceu apenas 1,6 vezes.

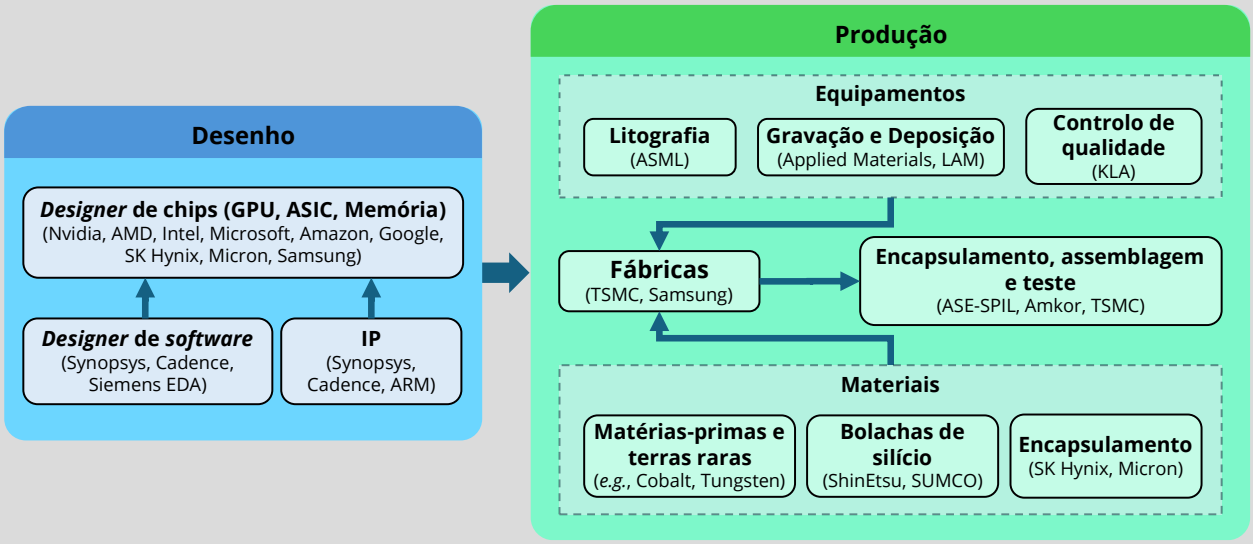
Além do *hardware*, o *software* desempenha um papel fundamental na otimização do desempenho dos *chips* de IA. As plataformas de *software*, incluindo *drivers*, compiladores, *runtimes* e *frameworks* de desenvolvimento, fazem a interligação entre os modelos de IA e o *hardware* subjacente.

Atingir o estado da arte em sistemas de IA, em termos de desempenho, depende não só da capacidade de computação, mas também da interação entre os recursos de computação, de memória e de comunicação[16]. O acesso contínuo e sustentado a esses *inputs* é determinante para o desempenho dos modelos e é influenciado por mercados concentrados de tecnologias avançadas de *chips*. Este cenário levanta questões sobre a dinâmica competitiva, as dependências estruturais e os riscos potenciais para a concorrência ao longo da cadeia de valor da IA. Estas questões são analisadas nas secções seguintes.

II. Estrangulamentos ao longo da cadeia de valor de *chips* de IA

A produção de *chips* de IA assenta numa complexa cadeia de valor global, que envolve diversas etapas interligadas, desde o *design* até ao fabrico. A Caixa 1 apresenta esta cadeia de valor, detalhando cada uma das principais fases, os componentes envolvidos e os principais intervenientes ao longo do processo.

Caixa 1 – Cadeia de valor dos *chips* de IA e principais intervenientes



Desenho de chips

O desenho de *chips* cria o plano técnico do *chip*, que define os seus componentes e o modo como estes operam. Este processo é complexo, intensivo em I&D, e visa assegurar que os *chips*

[16] A subutilização dos *chips* é frequentemente atribuída a estrangulamentos de memória, à sobrecarga de sincronização e a ineficiências de *software*, mais do que à simples insuficiência de computação bruta (aqui e aqui).

apresentam elevados níveis de desempenho, eficiência energética, fiabilidade e custo-benefício.

A NVIDIA é o principal interveniente no desenho de GPUs de alto desempenho[17] utilizados em IA, detendo mais de 90% da quota de mercado de GPUs discretos para uso em *datacenters*[18]. Outros concorrentes nesta fase incluem a AMD e a Intel. O segmento de ASICs específicos para IA, inclui *startups* como a Cerebras Systems e a Groq, bem como fornecedores de serviços de *cloud* como a Google (TPUs) e Amazon (Inferentia e Trainium).

A maioria dos *chips* de IA não é concebida completamente de raiz[19]. Em vez disso, os *designers* de *chips* integram módulos pré-desenhados licenciados, conhecidos como *IP cores*[20], desenvolvidos por empresas como a Synopsys, a Cadence e a Arm, em arquiteturas personalizadas, de forma a reduzir a complexidade e acelerar a entrada no mercado. Estes intervenientes, juntamente com a Siemens EDA, fornecem igualmente *software* de desenho especializado (*electronic design automation* – EDA), essencial para a verificação de esquemas de circuitos, a otimização do *design* físico e a simulação de desempenho antes do fabrico físico.

Os *chips* de memória, e em particular a memória HBM, são igualmente críticos para escalar e acelerar o treino e a inferência nos aceleradores de IA. Os módulos de HBM são fisicamente integrados com o *chip* acelerador (como os GPUs ou os TPUs), criando uma unidade integrada em que o bloco de processamento (*compute die*) fornece a capacidade de processamento e a HBM assegura a largura de banda de memória necessária. A oferta global de HBM encontra-se concentrada, sendo a SK Hynix, a Samsung e a Micron os principais fornecedores[21].

Produção de chips

Os planos de *design* dos *chips* são enviados para as fábricas (*foundries* ou *fabs*), que são contratadas para fabricar os *chips* físicos de acordo com as especificações do *designer*. Existe igualmente concentração nesta fase, sendo a TSMC o principal fabricante de *chips* de IA de alto desempenho, particularmente em nós de processo (*nodes*) de 5 nanometros ou abaixo[22].

A montante, o processo de fabrico depende de materiais especializados, incluindo bolachas (*wafer*) de silício, elementos de terras raras e produtos químicos, obtidos através de uma complexa rede global de abastecimento.

---

[17] Os principais *chips* de IA da NVIDIA incluem os modelos A100 e o H100, bem como os mais recentes *chips* da geração Blackwell. A NVIDIA desenvolveu igualmente os *chips* Hopper (H100/H200) e recentemente apresentou os *chips* Blackwell (B200/B300) (aqui).

[18] Quota de mercado utilizada na decisão da Comissão Europeia no processo M.11766 – NVIDIA / RUN:AI (ver decisão aqui). Dados da Epoch AI sobre a *clusters* de GPUs apresentam números semelhantes. A NVIDIA é o principal fornecedor de GPUs de cerca de 88% dos *clusters* com fornecedores conhecidos, ao passo que a AMD é o segundo com 7% e a Google é o terceiro com 4%. Estes dados incluem apenas *clusters* de GPUs existentes, confirmados, não-duplicados e privados, de acordo com a Epoch AI. Ver mais em Pilz et al. (2026).

[19] Embedded, "*Silicon Integrated Circuit Intellectual Properties Design*", outubro de 2024 (aqui).

[20] Os *IP cores* são blocos de circuitos padronizados e reutilizáveis que fornecem funcionalidades essenciais, como processadores, controladores de memória ou interfaces de conectividade.

[21] The Economist, "*Memory chips could be the next bottleneck for AI*", outubro de 2024 (aqui).

[22] CNN Business, "*Why this key chip technology is crucial to the AI race between the US and China*", junho de 2025 (aqui).

Após o fabrico, os *chips* são enviados para a fase de encapsulamento, assemblagem e teste (ATP - *Assembly, Testing and Packaging*), onde são encapsulados, testados e ligados a sistemas de memória e de interconexão. O mercado global de ATP é liderado pela ASE e pela Amkor, enquanto a TSMC também fornece serviços de encapsulamento avançado internamente[23].

Tanto a fase de *design* como a fase de fabrico da cadeia de valor de *chips* de IA apresentam um elevado grau de concentração, nomeadamente devido a fatores estruturais. Em ambas as fases existem economias de escala significativas. Do lado do *design*, o desenvolvimento de *chips* de última geração implica investimentos significativos em I&D, ciclos extensos de verificação e o recurso a ferramentas altamente sofisticadas. Do lado do fabrico, a produção de *chips* avançados exige investimentos avultados em unidades de fabrico[24], bem como acesso a equipamento altamente especializado, como sistemas de litografia ultravioleta extrema.

As características estruturais da cadeia de valor de *chips* de IA podem também contribuir para uma rigidez da oferta. Custos fixos elevados, prazos longos de desenvolvimento e de entrada no mercado, e a capacidade limitada dos fabricantes, restringem a possibilidade de aumentar rapidamente a produção. À medida que a procura por *chips* mais avançados acelera[25], estas restrições intensificam-se ao longo da cadeia de valor, contribuindo para a escassez persistente de *chips* que afeta fornecedores de IA e prestadores de serviços de *cloud*.

Com efeito, a procura por GPUs de elevado desempenho continua a exceder a oferta disponível[26]. Uma situação semelhante verifica-se no segmento de memória, com relatórios a indicarem que os *chips* HBM se encontram praticamente esgotados para 2025 e grande parte de 2026[27]. O acesso ao fabrico de *chips* de IA avançados permanece limitado[28], e tem sido dada prioridade a clientes com maior capacidade

[23] Ver relatórios aqui e aqui.
[24] Estimativas indicam que novas unidades de fabrico podem custar entre 15 e 20 mil milhões de dólares (aqui), com prazos de construção e de instalação de equipamento a variar entre 18 meses e mais de quatro anos (aqui, aqui e aqui).
[25] Comparando os *clusters* de GPUs planeados e existentes na base de dados da Epoch AI, a capacidade de computação total dos sistemas planeados nos próximos ~5 anos é aproximadamente 37 vezes maior que a capacidade de computação existente. Estes dados incluem os *clusters* de GPUs confirmados e prováveis, existentes e planeados, não duplicados, de acordo com a Epoch AI. A capacidade de computação é baseada no máximo teórico de desempenho em 16-bit FLOP/s. Ver mais em Pilz et al. (2026).
[26] Apesar dos aumentos de produção esperados para os *chips* da geração Blackwell, a oferta da NVIDIA ainda não conseguiu acompanhar a procura, segundo a conferência de resultados do 3.º trimestre de 2025 da NVIDIA.
[27] Yahoo Finance, "*Micron Sells Out 2025 HBM Supply: Can It Meet Soaring Demand in 2026?*", junho de 2025 (aqui).
SAM Mobile, "*SK Hynix will soon sell out 2026 HBM chips while Samsung struggles*", março de 2025 (aqui).
[28] Mais de 50% das empresas de IA generativa relatam a escassez de GPUs como um obstáculo significativo à expansão das suas operações. Vertu, "*Unveiling Trends in the Global AI Chip Market*", maio de 2025 (aqui).

financeira, segundo a informação disponível[29]. As restrições na capacidade de encapsulamento causam igualmente atrasos, em particular porque as fábricas de encapsulamento avançado são intensivas em capital e não podem ser escaladas rapidamente[30].

O ciclo de vida dos *chips* de IA tem também encurtado, impulsionado pelo aumento das exigências computacionais por parte dos fornecedores de IA e pelo ritmo acelerado da inovação tecnológica no setor[31]. Esta dinâmica intensifica ainda mais as barreiras à entrada e à expansão ao longo da cadeia de abastecimento de IA, uma vez que as empresas têm de se adaptar a ciclos de vida mais curtos enquanto continuam a operar dentro das limitações de capacidade e de investimento[32].

Ciclos de vida mais curtos dos *chips* traduzem-se numa depreciação mais rápida do *hardware* de IA, aumentando o risco financeiro para as empresas que detê m e alugam capacidade de computação. Este aspeto é particularmente relevante para intermediários de *cloud* de menor dimensão, cujo modelo de negócio depende da manutenção de taxas elevadas de utilização ao longo de períodos mais longos, de modo a amortizar os investimentos em GPUs. Em contraste, os principais fornecedores de serviços de *cloud* beneficiam de acesso a capital mais barato e de uma procura interna sustentada, o que lhes permite absorver uma depreciação mais rápida, suavizar a utilização ao longo do tempo e renovar o *hardware* com maior frequência.

As pressões de natureza estrutural e do lado da procura, bem como a escassez daí resultante, conferiram poder de mercado aos principais operadores no setor de IA e contribuíram para aumentos sustentados do preço dos *chips* de IA[33].

Este contexto cria barreiras à entrada e à expansão para os fornecedores de IA, em virtude do elevado custo de acesso à

[29] Em meados de 2024, a TSMC revelou que a sua capacidade de encapsulamento avançado para 2024 e 2025 estava totalmente reservada por apenas dois clientes – NVIDIA e AMD – que competiam entre si para assegurar capacidade para *chips* de IA.

[30] Embora o CEO da NVIDIA, Jensen Huang, tenha destacado que a indústria tinha "*provavelmente quatro vezes*" (tradução AdC) mais capacidade de encapsulamento avançado em 2025 do que em 2023, a procura era tão elevada que este continuava a ser o principal fator limitativo (aqui). No final de 2024, a NVIDIA vendia os seus mais recentes *chips* de IA Blackwell tão rapidamente quanto a TSMC os conseguia montar, permanecendo os serviços de encapsulamento "*um estrangulamento devido às restrições de capacidade*" (tradução AdC), segundo o presidente da TSMC, Mark Liu.

[31] Um relatório conclui que o tempo médio entre o lançamento dos principais *chips* de IA e a publicação do modelo de ponta treinado com esses *chips* é de aproximadamente quatro anos. Esta tendência é igualmente reconhecida pelos fornecedores de serviços de *cloud*. A CoreWeave divulgou na sua documentação de Oferta Pública Inicial (OPI) que o lançamento do Blackwell desvalorizou imediatamente os seus sistemas baseados na arquitetura Hopper (geração anterior de *chips* da NVIDIA), exigindo a contabilização de uma maior depreciação. A AWS e outros prestadores encurtaram igualmente os prazos de depreciação dos seus servidores (aqui).

[32] A Microsoft, por exemplo, enfrenta atrasos no desenvolvimento e na implementação dos seus *chips* de IA, tendo dificuldades em igualar o desempenho e o ritmo de comercialização de novos *chips* da NVIDIA (aqui).

[33] Uma das consequências tem sido o aumento acentuado do custo de treino de modelos de IA, que terá crescido mais de 300% em apenas alguns anos, impulsionado pelo aumento dos preços dos GPUs. Vertu, "*Unveiling Trends in the Global AI Chip Market*", maio de 2025 (aqui). De acordo com a *Cloud*zero, na Google *Cloud*, uma única instância de GPU A100 pode custar mais de 15 vezes o preço de uma instância padrão de CPU.

capacidade de computação necessária para treinar e implementar modelos avançados. A maioria dos fornecedores de IA, em particular *startups*, não consegue construir infraestruturas própria de computação devido aos elevados custos iniciais. Embora o recurso à computação em *cloud* possa atenuar parcialmente estes efeitos de escala, o elevado custo marginal de acesso à capacidade computacional pode limitar a sua capacidade de entrada no mercado, de inovação e de concorrência ao longo do ecossistema de IA[34].

Infraestrutura pública de computação de alto desempenho (*High-Performance Computing* – HPC) pode mitigar em parte este estrangulamento ao alargar o acesso a capacidade de computação em larga escala[35]. Estes supercomputadores podem apoiar ser usados em tarefas de IA intensivas em computação agregando milhares de *chips* avançados ligados entre si.

Ainda que a sua escala permaneça limitada relativamente à procura global de IA por computação[36]**, está a ser construída mais infraestrutura pública de HPC para ser usada em projetos indústrias e comerciais de IA**[37]. Ao nível europeu, foram recentemente lançadas as Fábricas de IA[38] e estão planeadas as Gigafábricas de IA[39], que têm como objetivo aumentar a escala da infraestrutura de computação pública para apoiar o desenvolvimento e a implementação de soluções de IA.

O potencial pró-concorrencial da infraestrutura pública de HPC dependerá da forma como for gerida e das condições de acesso. Dada a escassez de capacidade de computação e de memória para fornecedores de IA, a forma como o acesso a estes supercomputadores for concretizado determinará as condições de concorrência em mercados de IA a jusante. Por este motivo, os critérios de acesso a esta infraestrutura devem

---

[34] Uma pesquisa realizada pela Civo, um fornecedor de serviços de *cloud*, revelou que 85% das organizações reportaram atrasos nos seus projetos de IA devido à disponibilidade limitada de GPUs, com mais de um terço a citar atrasos entre três e seis meses. Ver aqui o artigo que reporta os resultados desse estudo.

[35] Ver OECD Roundtables on Competition Policy Papers, No. 330, "*Competition in artificial intelligence infrastructure*", novembro de 2025 (aqui).

[36] De acordo com a base de dados da Epoch AI sobre *clusters* de GPUs, a capacidade de computação total existente ao abrigo do programa EuroHPC é comparável à de empresas privadas como a Meta (~1,59 vezes a capacidade do EuroHPC) ou a Google (~0,93 vezes). Estes dados incluem os *clusters* de GPUs confirmados, existentes, não duplicados, de acordo com a Epoch AI. A capacidade de computação é baseada no máximo teórico de desempenho em 16-bit FLOP/s. Ver mais em Pilz et al. (2026).

[37] Por exemplo, algumas iniciativas em Portugal, como o InovIA (programa de *vouchers* para inovação que permite o acesso ágil a supercomputadores), destinam-se a Pequenas e Médias Empresas e a *start-ups*, e pretendem dar-lhes acesso a infraestrutura de HPC para fins de I&D e para experimentação.

[38] No âmbito da EuroHPC, a União Europeia pretende instalar 19 Fábricas de IA: ecossistemas de HPC otimizados para IA, que oferece capacidade de computação e serviços de apoio à indústria europeia e a investigadores, para o desenvolvimento de modelos de IA de grande dimensão (ver mais aqui e aqui).

[39] A Comissão anunciou que as Gigafábricas de IA serão capazes de treinar os modelos de IA mais avançados e que terão uma capacidade de computação pelo menos quatro vezes superior à das Fábricas de IA. Os investimentos serão realizados através de parcerias público-privadas, com uma contribuição de 20 mil milhões de euros da Comissão a partir da iniciativa InvestAI (aqui).

permanecer transparentes, objetivos e não discriminatórios.

Estrangulamentos estruturais na cadeia de valor de *chips* de IA podem comprometer o acesso ao mercado

As restrições persistentes do lado da oferta, a elevada concentração ao longo da cadeia de valor e o aumento da procura por *hardware* de ponta criaram estrangulamentos na produção e entrega de *chips* de IA, limitando a capacidade dos fornecedores de IA de menor dimensão de aceder à computação necessária para concorrer no desenvolvimento de IA avançada.

Iniciativas públicas de HPC podem expandir o acesso a computação para IA, ainda que a sua escala ainda seja limitada

A infraestrutura pública de HPC pode alargar o acesso a capacidade de computação em larga escala e mitigar, em parte, restrições do lado da oferta. Embora ainda limitadas em escala, têm existido iniciativas europeias para expandir o papel desta infraestrutura para desenvolvimento de IA. O seu impacto na concorrência dependerá de critérios de acesso transparentes, objetivos e não discriminatórios.

III. Fornecedores de serviços de *cloud* expandem ao longo da cadeia de valor de IA

Os principais fornecedores de serviços de *cloud* tornaram-se centrais no desenvolvimento de IA. Estes – frequentemente designados por *hyperscalers* – incluem a Amazon Web Services (AWS), a Google *Cloud* e a Microsoft Azure. Mesmo fornecedores de IA de referência, com modelos de última geração e uma base muito alargada de utilizadores, como a OpenAI, dependem desta computação na *cloud* para treinar, fazer inferência e implementar os seus modelos de IA.

Para reduzir a dependência dos fornecedores de *chips*, os *hyperscalers* têm investido cada vez mais no *design* de *chips* de IA proprietários, como o Trainium e o Inferentia da Amazon[40], ou os TPUs da Google[41].

Embora estes *chips* tenham sido inicialmente concebidos para utilização interna, a maioria dos fornecedores de serviços de *cloud* passou, entretanto, a integrá-los nos seus próprios canais de distribuição de serviços de *cloud*[42]. Estes *chips* permanecem integrados na infraestrutura de *cloud* dos respetivos prestadores e, em regra, não se encontram disponíveis para aquisição no mercado aberto (*i.e.*, os fornecedores de IA acedem a estes *chips* através das plataformas de *cloud* correspondentes). Nestes ambientes, os fornecedores de serviços de *cloud* podem oferecer acesso a uma combinação de *chips*

[40] AWS Tranium (aqui); AWS Inferentia (aqui).

[41] Os TPUs da Google são especificamente otimizados para as tarefas de IA da própria Google, demonstrando eficiência tanto em tarefas de inferência como de treino (ver mais aqui e aqui). Mais de metade das tarefas de IA na Google *Cloud* dependem agora de *chips* TPU (aqui).

[42] A OpenAI, por exemplo, encontra-se atualmente a alugar TPUs da Google, com o objetivo de diversificar a sua capacidade computacional em IA face aos *chips* da NVIDIA e aos serviços de *cloud* da Azure (ver mais aqui). A AWS disponibiliza igualmente os seus próprios *chips* de IA (ver aqui, aqui e aqui).

proprietários e de GPUs de terceiros, como os da NVIDIA ou da AMD.

Alguns fornecedores de serviços de *cloud* poderão vir a fornecer diretamente os seus próprios *chips* a fornecedores de IA. A título exemplificativo, de acordo com informação pública, os TPUs da Google poderão vir a ser utilizados, no futuro, pela Meta[43].

A entrada dos *hyperscalers* no *design* de *chips* pode ter efeitos pró-concorrenciais. Estes desenvolvimentos podem contribuir para a diversificação da oferta e para a redução da dependência relativamente a um conjunto restrito de intervenientes, ajudando a atenuar os estrangulamentos atualmente existentes.

No entanto, esta evolução pode igualmente suscitar riscos concorrenciais, em particular o risco de estratégias de exclusão. A integração vertical pode conferir aos *hyperscalers* a capacidade e, em determinadas circunstâncias, o incentivo para alavancar o seu duplo papel enquanto fornecedores de serviços de *cloud* e *designers* de *chips*, reforçando a sua posição competitiva ao longo da cadeia de valor de IA. Por um lado, controlam os ambientes de *cloud* que constituem o principal canal através do qual os fornecedores de IA acedem a recursos de computação. Por outro, concebem e implementam *chips* proprietários, integrados nesses mesmos ambientes de *cloud*. Em tais cenários, os *hyperscalers* podem ter incentivos para favorecer os seus próprios *chips* através de estruturas de preços diferenciadas, critérios de elegibilidade para descontos, alocação de capacidade ou uma integração mais estreita com o seu próprio *software*[44].

À medida que os fornecedores de serviços de *cloud* se expandem para o *design* de *chips*, as empresas de *chips* estão, por sua vez, a ir além do fornecimento de *hardware*, avançando para serviços de computação em *cloud* orientados para IA. Exemplos desta evolução incluem iniciativas recentes como a Tiber AI *Cloud* da Intel[45], a Developer *Cloud* da AMD[46] e o DGX *Cloud* da NVIDIA[47].

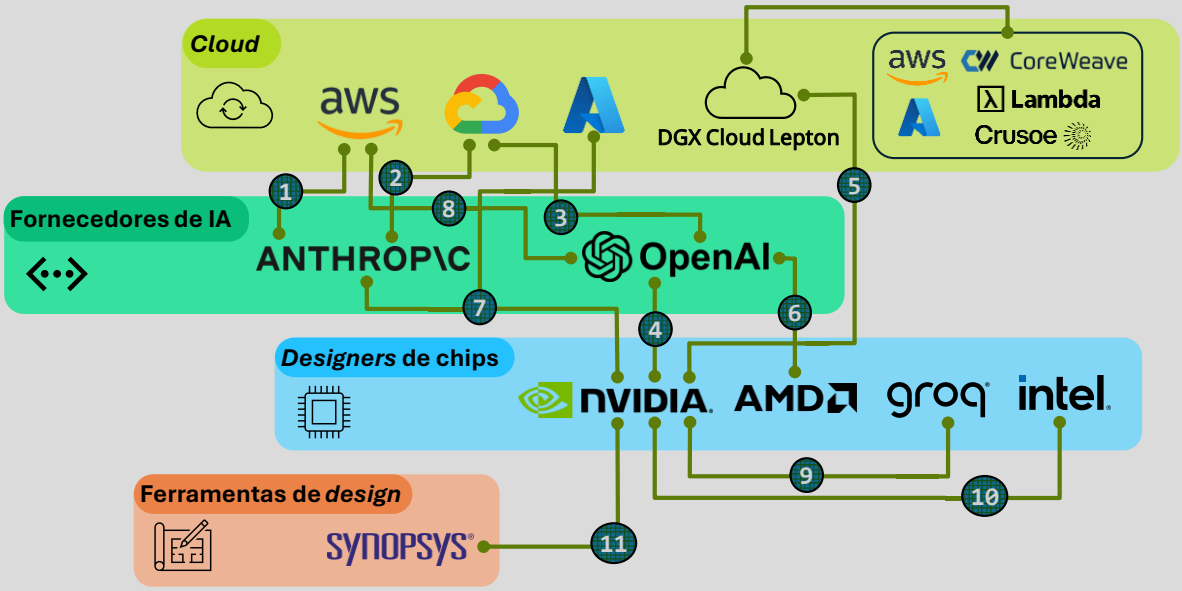
Até agora, estas plataformas parecem estar a ser usadas principalmente para mostrar as capacidades do *hardware* subjacente e dos ecossistemas a empresas que procuram investir em IA. Com o tempo, no entanto, podem assumir um papel mais estratégico, reduzindo a dependência dos *hyperscalers*, enquanto garantem que os produtos proprietários estejam disponíveis através de ambientes hospedados adaptados às suas características de

[43] The Wall Street Journal, "*Meta Is in Talks to Use Google's Chips in Challenge to Nvidia*", novembro de 2025 (aqui).
[44] Por exemplo, a Google *Cloud* exclui várias famílias de GPUs da NVIDIA – incluindo A100, H100, H200, B200 (Blackwell) – das reduções automáticas de preço ao abrigo dos *Sustained Use Discounts* (aqui), excluindo igualmente as instâncias A4 (baseadas em arquitetura Blackwell) dos *Committed-Use Discounts* (aqui). As instâncias TPU, em contraste, continuam elegíveis para esses descontos.
[45] A *Tiber AI Cloud* da Intel foi lançada em outubro de 2024 como uma evolução da anterior *Intel Tiber Developer Cloud*, sendo direcionada para tarefas de IA em grande escala, e não apenas testes de desenvolvimento (aqui).
[46] A *Developer Cloud* da AMD foi lançada em junho de 2025 como uma plataforma orientada para fornecedores de IA, com o objetivo de aumentar o acesso às GPUs Instinct da AMD, com *software* ROCm pré-configurado (aqui).
[47] Em junho de 2025, a NVIDIA expandiu o seu *DGX Cloud* com o *DGX Cloud Lepton*, uma plataforma de agregação de GPUs que reúne capacidade excedentária dos seus parceiros de *cloud* (ver Caixa 2). Esta solução permite que os fornecedores de IA acedam dinamicamente a recursos de GPU não utilizados, sem ficarem vinculados a um único prestador, facilitando a transição entre diferentes plataformas de *cloud* (ver mais aqui e aqui).

desempenho, e reforçando o controlo sobre a interface *hardware–software*.

As parcerias desempenham um papel central na operacionalização destas estratégias ao longo da cadeia de valor de IA. Os *hyperscalers* estabeleceram diversas parcerias com fornecedores de IA, em particular envolvendo o acesso a infraestruturas de *cloud* e promovendo simultaneamente a utilização dos seus *chips*. Os *designers* de *chips*, por sua vez, estabeleceram parcerias com fornecedores de serviços de *cloud* de menor dimensão que procuram expandir no mercado de IA. A Caixa 2 apresenta um conjunto de parcerias divulgadas publicamente.

Caixa 2 – Parcerias em IA ao longo da cadeia de valor



As parcerias tornaram-se uma característica central do ecossistema de IA, moldando a forma como os recursos de computação são desenvolvidos e distribuídos. Estas parcerias refletem frequentemente a dupla tendência de os *hyperscalers* se expandirem para o *design* de *chips* e de os *designers* de *chips* entrarem nos serviços de *cloud*, procurando ambos atrair fornecedores de IA para os seus ecossistemas. Os seguintes exemplos ilustram estes desenvolvimentos, embora a lista não seja exaustiva.

1. **AWS e Anthropic**: Logo após a parceria com a Google, a Anthropic iniciou a sua relação com a AWS, estabelecendo a AWS como o seu principal parceiro de *cloud* e de treino. Este acordo seguiu-se a um investimento substancial da Amazon, de cerca de 1,25 mil milhões de dólares em setembro de 2023, atingindo os 8 mil milhões de dólares em 2024. Esta colaboração inclui explicitamente o trabalho conjunto no desenvolvimento do Trainium, o *chip* proprietário da AWS concebido para o treino de modelos de IA[48].

[48] Anthropic, "*Powering the next generation of AI development with AWS*", novembro de 2024 (aqui).

2. **Google *Cloud* e Anthropic**: A Anthropic colaborou inicialmente com a Google *Cloud*, anunciando a sua parceria em fevereiro de 2023. No âmbito desta colaboração, a Anthropic usou os *clusters* de GPUs e TPUs da Google para o desenvolvimento dos seus modelos de IA[49].
3. **Google *Cloud* e OpenAI**: A OpenAI estabeleceu uma parceria com a Google *Cloud* a partir do início de 2025, com o objetivo de aceder a recursos adicionais de computação, incluindo TPUs da Google[50].
4. **NVIDIA e OpenAI**: Em setembro de 2025, a NVIDIA e a OpenAI celebraram uma parceria para a implementação de até 10 gigawatts de sistemas NVIDIA nos próximos *datacenters* da OpenAI. Ao abrigo deste acordo, a NVIDIA irá investir até 100 mil milhões de dólares, em tranches de 10 mil milhões por gigawatt de capacidade, estando o primeiro gigawatt previsto para o segundo semestre de 2026. Por cada 10 mil milhões de dólares investidos pela NVIDIA, a OpenAI adquirirá cerca de 35 mil milhões de dólares em GPUs da NVIDIA[51]. Contudo, notícias recentes sugerem que as negociações poderão estar estagnadas, não tendo sido ainda formalizado um acordo definitivo[52].
5. **DGX *Cloud* Lepton da NVIDIA**: trata-se num *marketplace* de GPUs em que a NVIDIA agrega capacidade de múltiplos fornecedores de serviços de *cloud* num único catálogo. Os participantes confirmados nesta plataforma incluem *hyperscalers*, como a AWS e a Microsoft Azure, e outros fornecedores de serviços de *cloud* como a CoreWeave, a Crusoe, a Nebius, a Lambda Labs, entre outros[53]. Alguns dos termos comerciais tornaram-se públicos em setembro de 2025. Por exemplo, a NVIDIA terá assinado um contrato de 1,5 mil milhões de dólares, por um período de quatro anos, para alugar cerca de 18.000 GPUs que havia anteriormente vendido à Lambda[54]. Adicionalmente, a NVIDIA terá assinado um contrato de 6,3 mil milhões de dólares com a CoreWeave, ao abrigo do qual a NVIDIA comprará qualquer capacidade de *cloud* não utilizada pelos clientes[55]. Em janeiro de 2026, a NVIDIA terá ainda investido 2 mil milhões de dólares na CoreWeave, aumentando a sua participação acionista nesta empresa[56].
6. **AMD e OpenAI**: Em outubro de 2025, a AMD celebrou um acordo para fornecer 6 gigawatts de capacidade de computação, utilizando os seus futuros *chips* de IA MI450, nos próximos *datacenters* da OpenAI. As primeiras entregas previstas para o segundo semestre de 2026. O acordo inclui a opção de a OpenAI adquirir até 10% das ações da AMD a um preço nominal de 0,01 dólares por ação[57].
7. **Microsoft, NVIDIA e Anthropic**: Em novembro de 2025, a Microsoft e a NVIDIA ter-se-ão comprometido a investir até 5 mil milhões e 10 mil milhões de dólares, respetivamente, na

[49] Anthropic, "*Anthropic Partners with Google Cloud*", fevereiro de 2023 (aqui).
[50] Revolgy, "*Google Cloud signs deal with OpenAI, ending Microsoft's exclusive role*", junho de 2025 (aqui).
[51] The Economist, "*Nvidia's $100bn bet on OpenAI raises plenty of questions*", setembro de 2025 (aqui).
[52] CNBC, "*Nvidia, OpenAI appear stalled on their mega deal. But the AI giants still need each other*", fevereiro de 2026 (aqui).
[53] Anunciado pela NVIDIA em junho de 2025. *Google Cloud* não lista entre os participantes desta iniciativa (aqui).
[54] Tom's *Hardware*, "*Nvidia signs $1.5 billion deal with cloud startup Lambda to rent back its own AI chips*", setembro de 2025 (aqui).
[55] Reuters, "*CoreWeave, Nvidia sign $6.3 billion cloud computing capacity order*", setembro de 2025 (aqui).
[56] Reuters, "*Nvidia invests $2 billion in CoreWeave to boost data center build-out*", Janeiro de 2026 (aqui).
[57] Reuters, "*AMD signs AI chip-supply deal with OpenAI, gives it option to take a 10% stake*", outubro de 2025 (aqui).

Anthropic. Em paralelo, a Anthropic adquirirá cerca de 30 mil milhões de dólares em capacidade de computação à Microsoft, cujos *datacenters* recorrem a *chips* de IA da NVIDIA[58].

8. **AWS e OpenAI**: Em novembro de 2025, a AWS e a OpenAI terão celebrado um contrato de sete anos no valor de 38 mil milhões de dólares, que permite à OpenAI aceder a um grande número de GPUs NVIDIA GB200 e GB300 (Blackwell) alojados na infraestrutura da AWS[59].
9. ***Acquihire* da Groq pela NVIDIA**: Em dezembro de 2025, a NVIDIA celebrou um acordo de licenciamento de tecnologia não exclusivo com a Groq, uma empresa especializada em *chips* de IA orientados para inferência. Em paralelo, a NVIDIA contratou vários executivos e trabalhadores da Groq, incluindo o seu fundador, mantendo-se a Groq como empresa independente[60].
10. **Intel e NVIDIA**: Em dezembro de 2025, a NVIDIA terá adquirido uma participação acionista de cerca de 5 mil milhões de dólares (cerca de 4%) na Intel, um fabricante de *chips* concorrente[61].
11. **NVIDIA e Synopsys**: Em dezembro de 2025, a NVIDIA investiu 2 mil milhões de dólares (cerca de 2,6% da participação) na Synopsys, fornecedora de *software* de *design* de *chips*, no âmbito de uma parceria estratégica destinada a acelerar soluções de engenharia e de *design* em computação e inteligência artificial[62].

Algumas destas parcerias e investimentos suscitam preocupações concorrenciais, na medida em que combinam compromissos de aquisição de produtos e serviços, dependências tecnológicas e transferências de ativos estratégicos de forma suscetível de atenuar a rivalidade entre concorrentes atuais ou potenciais.

Em particular, o recurso crescente a *reverse acquihires* – conforme discutido no estudo da AdC sobre os mercados laborais em IA generativa – **pode conduzir à concentração de talento e de *know-how* especializado**. **Uma *reverse acquihire* pode ainda consubstanciar uma operação de concentração de acordo com o Regulamento das Concentrações Comunitárias ("*EU Merger Regulation*") e a Lei da Concorrência**[63].

Adicionalmente, algumas parcerias e investimentos poderão conceder às partes envolvidas acesso privilegiado a informação técnica e comercial sensível não disponível aos concorrentes[64]. Em determinadas circunstâncias, tais acordos poderão enquadrar-se no âmbito do artigo 101.º do Tratado sobre o Funcionamento da União Europeia (TFUE),

[58] Financial Times, "*Microsoft and Nvidia to invest up to $15bn in OpenAI rival Anthropic*", novembro de 2025 (aqui).
[59] OpenAI, "*AWS and OpenAI announce multi-year strategic partnership*", novembro de 2025 (aqui).
[60] Groq, "*Groq and Nvidia Enter Non-Exclusive Inference Technology Licensing Agreement to Accelerate AI Inference at Global Scale*", dezembro de 2025 (aqui).
[61] Reuters, "*Nvidia takes $5 billion stake in Intel under September agreement*", dezembro de 2025 (aqui).
[62] NVIDIA, "*NVIDIA and Synopsys Announce Strategic Partnership to Revolutionize Engineering and Design*", dezembro de 2025 (aqui).
[63] Ver o *Short Paper* da AdC "Concorrência e IA Generativa: Mercados Laborais", aqui.
[64] Ver U.S. Federal Trade Commission Staff Report, "*Partnerships Between Cloud Service Providers and AI Developers*", Janeiro de 2025 (aqui) e Comissão Europeia, *Policy brief* "*Competition in Generative AI and Virtual Worlds*", setembro de 2024 (aqui).

designadamente se a troca de informação contribuir reduzir a incerteza estratégica ou sustentar um comportamento colusivo[65].

Acresce que as parcerias entre empresas ativas em diferentes níveis da cadeia de valor da IA podem alterar os incentivos concorrenciais em mercados adjacentes. Uma empresa que detenha uma posição significativa num determinado segmento (por exemplo, serviços de *cloud* ou fornecimento de *chips*) poderá ter a capacidade e o incentivo para influenciar ou restringir o acesso a modelos de IA ou a *chips* de IA. Tal poderá afetar a concorrência tanto a montante (no desenvolvimento de modelos e no fornecimento de *hardware*) como a jusante (na distribuição de modelos de IA e de serviços baseados em IA, através de serviços de *cloud*)[66].

Os investimentos circulares, nos quais grandes empresas investem em fornecedores de IA que, subsequentemente, passam a depender dessas mesmas empresas para o fornecimento de *chips* ou de serviços de *cloud*, podem igualmente reforçar a posição dos incumbentes. Mesmo sem exclusividade explícita, este tipo de investimentos pode limitar a exposição a fornecedores alternativos, alinhar incentivos ao longo do ecossistema e acelerar processos de *tipping* em mercados de infraestruturas de IA caracterizados por efeitos de escala e concentração de *inputs*. Esta dinâmica é particularmente relevante em mercados em que os *designers* de *chips* dispõem de ecossistemas de *software* extensos, *frameworks* de desenvolvimento proprietários ou fortes efeitos de rede. Neste contexto, as dependências tecnológicas e o alinhamento de incentivos podem reforçar posições incumbentes, reduzir a contestabilidade e limitar a escolha efetiva dos fornecedores de IA. Estes riscos são analisados com maior detalhe na secção seguinte.

[65] Ver Comissão Europeia, Orientações sobre a aplicação do artigo 101.º do Tratado sobre o Funcionamento da União Europeia aos acordos de cooperação horizontal (revisão de 2023), julho de 2023 (aqui). Em particular, as Orientações explicam que a troca de informação pode ser apreciada ao abrigo do artigo 101.º quando for além do que é objetivamente necessário ou proporcional para a execução do acordo. Intercâmbio de informações adicional entre concorrentes pode reduzir ainda mais a incerteza estratégica, e mesmo uma divulgação unilateral por um concorrente pode, em determinadas circunstâncias, constituir uma prática concertada quando esse concorrente atua com base nessa divulgação e desde que exista um nexo de causa e efeito entre a divulgação e o comportamento posterior do concorrente no mercado.

[66] Ver decisão da CMA no âmbito do inquérito à concentração relativa à parceria Microsoft/OpenAI, março de 2025 (aqui). Nesta decisão, a CMA analisou se um eventual aumento da influência da Microsoft sobre a OpenAI poderia criar incentivos para restringir o acesso de concorrentes da Microsoft aos modelos fundacionais da OpenAI, nomeadamente nos mercados de capacidade computacional em *cloud* e nos mercados a jusante de *software* de produtividade.

A expansão da atividade ao longo da cadeia de valor de IA pode aumentar a concentração de mercado

À medida que os fornecedores de serviços de *cloud* se expandem para o *design* de *chips* e os fornecedores de *chips* entram nos serviços de *cloud*, o controlo sobre o *hardware* e a infraestrutura poderá tornar-se mais concentrado. Tal poderá reforçar incentivos de ecossistema que retêm os fornecedores de IA, suscitando preocupações ao nível da interoperabilidade, de *self-preferencing* e de riscos de exclusão.

IV. Integração *hardware-software* reforça os efeitos de *lock-in*

Os *chips* de IA combinam componentes de *hardware* e *software* para permitir a computação paralela complexa. Um elemento central da cadeia de valor de IA é a plataforma de *software* que faz a ponte entre os modelos de IA e o *hardware* subjacente, tanto ao nível de um único chip como de múltiplos *chips* a operar de forma conjunta.

Entre as principais plataformas de *software* para o desenvolvimento de IA[67], o CUDA, juntamente com as respetivas bibliotecas[68], tornou-se a plataforma mais utilizada, sendo apenas compatível com *hardware* da NVIDIA. De acordo com a empresa, esta plataforma é utilizada por mais de cinco milhões de programadores e mais de 3.700 aplicações[69]. Adicionalmente, a maioria das ferramentas de desenvolvimento de IA e das bibliotecas de programação foi desenvolvida inicialmente para o CUDA e está sobretudo otimizada para essa plataforma, ainda que o suporte a arquiteturas alternativas esteja a crescer[70].

Em virtude da interdependência entre *software* e *hardware*, a posição da NVIDIA no CUDA reforçou a sua posição no mercado de *chips* de IA. A adoção generalizada do CUDA gera fortes efeitos de rede em torno dessa plataforma e dos *chips* da NVIDIA[71]. À medida que mais fornecedores de IA recorrem ao CUDA, mais ferramentas e bibliotecas de programação para IA são desenvolvidas e otimizadas para essa plataforma, tornando os *chips* da NVIDIA mais versáteis e fáceis de utilizar. Este processo, por

[67] Para efeitos do presente *short paper*, entende-se por plataforma de *software* para o desenvolvimento de IA o conjunto integrado de ferramentas, linguagens, bibliotecas, compiladores e *runtimes* que permitem aos fornecedores de IA e programadores em geral programar e otimizar as tarefas para *chips* de IA específicos. Este conceito abrange igualmente extensões de linguagem (por exemplo, CUDA C++, HIP C++) e os respetivos *kits* de desenvolvimento de *software* (SDKs) disponibilizados pelos *designers* de *chips*.

[68] A plataforma CUDA (*Compute Unified Device Architecture*) é o modelo de programação e a *software stack* da NVIDIA para a programação dos seus GPUs. Inclui extensões de linguagem, compiladores, bibliotecas e ferramentas que permitem aos fornecedores de IA, e programadores em geral, explorar o processamento paralelo dos *chips*. Ao permitir uma computação paralela eficiente, a CUDA acelera significativamente tarefas computacionais complexas, especialmente aquelas associadas a IA, *machine learning, deep learning*, simulações científicas e cálculos intensivos de dados.

[69] Ver aqui.

[70] Embora o CUDA continue a ser a principal base de otimização para muitas tarefas de IA, a Google, por exemplo, tem vindo a melhorar o suporte para os seus TPUs no *PyTorch* (aqui e aqui).

[71] Ver *e.g.*, Hagiu & Wright (2025), disponível aqui.

sua vez, atrai mais fornecedores de IA para o CUDA e expande as ferramentas e bibliotecas de terceiros para o desenvolvimento de IA, reforçando ainda mais a posição da NVIDIA na cadeia de valor de IA.

Embora o ecossistema de ferramentas em torno do CUDA traga benefícios aos fornecedores de IA, pode também retê-los nos *chips* da NVIDIA. Para os fornecedores de IA, a mudança para *chips* alternativos implica abandonar o CUDA e adotar outras plataformas de *software*, implicando a perda de muitas das ferramentas otimizadas que utilizavam. Esta transição pode exigir a reescrita de código de baixo nível e a requalificação das equipas técnicas, o que poderá estar apenas ao alcance dos maiores fornecedores de IA[72].

Ainda que as principais ferramentas de programação se tenham tornado progressivamente mais independentes do tipo de *chip* utilizado[73], esta abstração apenas mitiga parcialmente os custos de mudança. Ferramentas como o *PyTorch* ou o *TensorFlow* permitem executar aplicações em diferentes plataformas de *hardware* sem recurso direto a *APIs* específicas de cada *chip*. No entanto, essa portabilidade revela-se sobretudo eficaz em casos de uso genéricos e nas fases iniciais de desenvolvimento. Em contextos de otimização de desempenho e de implementação dos modelos de IA em larga escala, as ferramentas continuam a depender de *hardware* específico.

Assim, os efeitos de *lock-in* manifestam-se de forma desigual entre diferentes categorias de fornecedores de IA. Fornecedores de IA de menor dimensão, com menores requisitos de desempenho, podem tolerar ineficiências e revelar menor sensibilidade à escolha do *hardware*. Por outro lado, grandes empresas tecnológicas dispõem frequentemente da escala e dos recursos necessários para desenvolver ferramentas próprias e operar simultaneamente em múltiplos ecossistemas de *software*. Entre estes dois extremos encontra-se um segmento de empresas de média dimensão, que não dispõe nem de margem para acomodar ineficiências, nem de capacidade para personalizações profundas. Para este grupo, os custos de mudança e a dependência de plataformas dominantes podem ser particularmente limitativos.

Mesmo quando existem ferramentas de tradução[74], estas dificilmente asseguram plena compatibilidade com o ecossistema do

---

[72] Por exemplo, esta questão tem sido reportada no contexto dos TPUs da Google, uma vez que, historicamente, os fornecedores de IA eram obrigados a recorrer ao *JAX*, em vez do *PyTorch*, para desenvolver os seus modelos. Estes custos de mudança podem limitar a atratividade dos TPUs da Google. Para mitigar este problema, a Google e a Meta terão estabelecido uma parceria para reduzir esses custos de mudança, reforçando a compatibilidade dos TPUs com o *PyTorch* e facilitando a transição dos fornecedores de IA dos *chips* da NVIDIA para os da Google (ver nota de rodapé 70 e aqui).

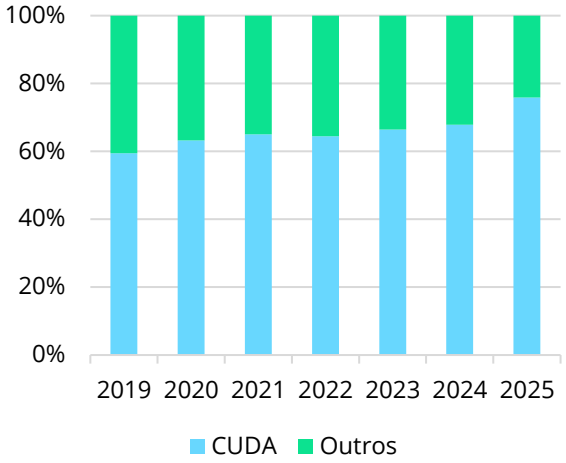
[73] Os programadores tipicamente interagem com ferramentas de programação dedicadas a IA, e usam as funcionalidades padrão no seu código (*e.g.*, importam e usam bibliotecas de programação como o PyTorch ou o TensorFlow), em vez de programarem diretamente em ferramentas específicas a determinado *hardware*, como o CUDA.

[74] Exemplos de ferramentas de tradução incluem o HIPIFY, o ZLUDA, o SCALE ou o SYCLomatic da Intel.

CUDA, levando as equipas a sacrificar desempenho ou funcionalidade[75].

Existem plataformas de *software* alternativas, mas a sua adoção permanece limitada. Embora soluções como o ROCm da AMD, o oneAPI/SYCL da Intel, ou *standards* abertos como o OpenCL possam, em alguns casos, oferecer desempenho equiparável ou facilitar a portabilidade, os seus ecossistemas são, em geral, menos maduros do que o do CUDA[76]. A figura *infra* ilustra este padrão. O CUDA é consistentemente a plataforma com a maioria das referências em repositórios do *GitHub* quando comparado com as alternativas, e a sua quota aumentou de forma contínua entre 2019 e 2025[77]. Ainda que baseados numa análise por palavras-chave e, por isso, indicativa, estes dados sugerem um reforço continuado da importância relativa do CUDA no desenvolvimento de IA.

Instâncias de repositórios *GitHub* que referenciam o CUDA face a outras plataformas de *software*



Fonte: Dados recolhidos pela AdC em janeiro de 2026 com base numa análise automatizada de repositórios públicos do *GitHub*. As percentagens reportam o rácio entre novos repositórios criados mensalmente que referenciam o CUDA e os que referenciam plataformas de *software* alternativas (ROCm/HIP, oneAPI/SYCL, OpenCL, Metal, Vulkan, XLA, Neuron).

As diferenças no grau de maturidade das plataformas refletem-se igualmente nos desfasamentos temporais com que novas ferramentas se tornam disponíveis em plataformas não-CUDA. Por exemplo, o *FlashAttention* – um mecanismo que reduz simultaneamente a utilização de memória e o tempo de execução dos modelos de IA – passou a estar disponível na plataforma CUDA em novembro de 2022. Contudo, apenas ficou

[75] Ver *e.g.*, Zahid et al. (2024), disponível aqui. Os autores identificam que a utilização do HIPIFY, um tradutor de CUDA para HIP, introduziu discrepâncias numéricas adicionais entre código CUDA (em *hardware* NVIDIA) e código HIP (em *hardware* AMD). Adicionalmente, segundo a documentação da AMD, a tradução de código CUDA para HIP pode gerar avisos, exigir correções manuais e envolver iterações de inspeção e intervenção antes de o código funcionar corretamente (ver aqui).

[76] HPCwire, "*Spelunking the HPC and AI GPU Software Stacks*", junho de 2024 (aqui).

[77] O *GitHub* é a maior plataforma mundial para hospedar e partilhar código de *software*, amplamente utilizada por fornecedores de IA e programadores para colaboração em projetos *open-source*, incluindo para fins comerciais. Por captar uma parte significativa da atividade dos programadores, a análise de referências nos repositórios do GitHub constitui um *proxy* útil para medir a prevalência relativa das diferentes plataformas de *software* no universo da IA.

disponível em maio de 2024 na plataforma ROCm de forma comparável, o que corresponde a uma diferença de aproximadamente 18 meses. Da mesma forma, o *bitsandbytes*, uma biblioteca para inferência de baixa precisão que reduz o consumo de memória e acelera tarefas de IA com modelos de grande dimensão, foi introduzido na plataforma CUDA em agosto de 2022, e apenas na plataforma ROCm em dezembro de 2024, um desfasamento de cerca de 28 meses[78].

Os *hyperscalers* estão a promover igualmente ferramentas de *software* associadas ao seu próprio *hardware*[79], que não dependem do CUDA. No entanto, estas iniciativas permanecem relativamente circunscritas, sendo viáveis sobretudo porque os *hyperscalers* se expandiram para o *design* de *chips*, conforme discutido na secção anterior. Neste caso, **subsiste o risco de os fornecedores de IA não escaparem aos efeitos de *lock-in*, limitando-se a transitar de um ecossistema fechado para outro**.

Esta integração entre as plataformas de *software* e o *hardware* subjacente pode contribuir para reforçar barreiras à entrada e à expansão no *design* de *chips*. As empresas concorrentes têm de, simultaneamente, igualar o desempenho do *hardware*, investir no desenvolvimento de um ambiente de *software* comparável e ultrapassar efeitos de rede, convencendo programadores e fornecedores de IA a criar e otimizar ferramentas para as suas plataformas.

Os custos de mudança do ecossistema CUDA podem reforçar efeitos de *lock-in* e as barreiras à entrada

A posição do CUDA enquanto plataforma dominante para o desenvolvimento de IA, conjugada com a sua falta de compatibilidade com *hardware* alternativo, pode implicar custos de mudança elevados para os fornecedores de IA e levantar barreiras à entrada ou à expansão para fornecedores concorrentes de *chips* no mercado de *hardware* de IA.

V. Do desenho de *chips* à expansão para infraestruturas de rede para IA

À medida que as tarefas de IA excedem a capacidade de *chips* individuais, o desempenho passa a depender da eficiência da comunicação entre *chips*. O treino e a inferência em grande escala exigem a troca frequente de dados entre múltiplos *chips*, tornando a latência, a largura de banda e a sincronização fatores críticos para o desempenho global do sistema. Neste contexto, as tecnologias de rede tornam-se um componente fundamental na computação de IA.

As interligações de alta largura de banda que ligam múltiplos *chips* dentro de um servidor são, geralmente, componentes proprietários do ecossistema de cada *designer* de *chips*.

[78] Estas técnicas de otimização foram selecionadas com base num guia da Hugging Face ("*Optimizing LLMs for Speed and Memory*", disponível aqui), que identifica o FlashAttention e a inferência de baixa precisão (através de bibliotecas como o bitsandbytes) como métodos eficazes para a implementação eficiente de LLMs. As datas de lançamento foram obtidas do PyPI, utilizando as primeiras versões de pacotes publicamente disponíveis correspondentes ao suporte CUDA e as primeiras versões documentando ou permitindo o suporte não CUDA comparável (por exemplo, ROCm).
[79] O XLA/JAX da Google para TPUs e o SDK Neuron da Amazon para *chips* Trainium e Inferentia.

Exemplos incluem o *NVLink* para a NVIDIA e o *Infinity Fabric* para a AMD. Como estas tecnologias funcionam apenas dentro dos seus próprios ecossistemas, limitam a possibilidade de combinar diferentes *chips* de IA[80].

Apesar da existência de algumas iniciativas para alargar o acesso a estas tecnologias, estes projetos continuam a depender de arquiteturas proprietárias e de condições técnicas específicas[81]. Em paralelo, padrões abertos como o *UALink* têm vindo a ser desenvolvidos com o objetivo de criar uma camada neutra para a interoperabilidade entre aceleradores[82].

Quando as tarefas de IA excedem a capacidade de computação e de memória de um único servidor, estas têm de ser distribuídas por *clusters* de máquinas interligadas. Estes *clusters* dependem de redes de alta velocidade para assegurar uma comunicação eficiente entre servidores.

Ao contrário do que sucede à escala intra-servidor, as interligações entre servidores recorrem, em maior medida, a padrões abertos. O *Ethernet* é a tecnologia predominante[83], amplamente utilizada em *datacenters*[84] e acessível a todos os fornecedores.

Uma alternativa ao *Ethernet* é o *InfiniBand*, uma tecnologia de interconexão proprietária da NVIDIA, particularmente usada em computação de alto desempenho[85]. O *InfiniBand* oferece menor latência e maior largura de banda do que o *Ethernet*, mas geralmente a um custo superior[86]. A NVIDIA posiciona o *InfiniBand* como referência para o treino de modelos em grande escala, promovendo simultaneamente as suas próprias soluções baseadas em *Ethernet*[87].

À medida que os *designers* de *chips* se expandem para o fornecimento de infraestruturas de rede, poderá surgir o risco de alavancagem do poder de mercado detido ao nível dos *chips* de IA para esse segmento. Efeitos deste tipo podem surgir não apenas através da integração técnica, mas também através de acordos contratuais, incluindo cláusulas de exclusividade, práticas de *bundling*[88]

---

[80] Em termos gerais, a combinação de *chips* de diferentes *designers* exige que a comunicação recorra à tecnologia *PCIe*, um padrão aberto, que apresenta menor largura de banda e maior latência.

[81] Um exemplo é o *NVLink Fusion* da NVIDIA, anunciado em maio de 2025 como um programa de licenciamento que permite aos parceiros selecionados integrar portas *NVLink* nos seus próprios CPUs ou aceleradores personalizados. A iniciativa alarga o acesso além dos próprios GPUs da NVIDIA, mas a participação está condicionada, uma vez que os dispositivos ainda têm de se conectar através dos produtos da NVIDIA, como o *NVSwitch*, e depender do *software* da NVIDIA para gerir a comunicação entre os dispositivos (ver mais aqui).

[82] Ver página na *internet* do consórcio *UALink* aqui.

[83] Lightwave, "*Ethernet maintains a lead over InfiniBand in the AI race*", setembro de 2025 (aqui).

[84] Como ilustram projetos recentes da Azure, da Oracle ou da Meta (aqui).

[85] A lista TOP500 (aqui), que classifica os 500 supercomputadores mais poderosos do mundo com base em *benchmarks* de desempenho padronizados, indica que os interconectores *InfiniBand* são utilizados em 57% dos sistemas listados em novembro de 2025. A prevalência desta tecnologia da NVIDIA é ainda mais elevada entre as instalações recentes, atingindo 65% dos sistemas implementados nos últimos três anos.

[86] Ver mais aqui e aqui.

[87] Fierce Electronics, "*Nvidia envisions 'one gigantic GPU' by linking data centers*", agosto de 2025 (aqui).

[88] O Departamento de Justiça dos EUA terá iniciado uma investigação à NVIDIA relativa a um eventual abuso de posição dominante no fornecimento de *chips* de IA, incluindo possíveis práticas de condições privilegiadas no

ou descontos de fidelização, que podem reduzir os incentivos para os fornecedores de IA adotarem *chips* alternativos[89].

A expansão para o fornecimento de infraestruturas de rede pode reforçar o poder de mercado ao longo da cadeia de valor de IA

À medida que os *designers* de *chips* alargam a sua atividade ao fornecimento de infraestruturas de rede para IA, podem reforçar efeitos de *lock-in*, quer através da integração técnica entre *hardware* e tecnologias de interligação, quer através de práticas contratuais. Tal dinâmica poderá aumentar o risco de alavancagem do poder de mercado detido ao nível dos *chips* para segmentos adjacentes da cadeia de valor da IA.

fornecimento e preços destes *chips*, que favorecem consumidores que compram os sistemas da NVIDIA como um todo (incluindo componentes de *hardware* e de interligação). Ver The Information, "Nvidia Faces DOJ Antitrust Probe Over Complaints From Rivals", de Agosto de 2024 (aqui).

[89] A este respeito, preocupações semelhantes foram identificadas noutros documentos por outras autoridades da concorrência e organizações internacionais, incluindo: (i) OECD Roundtables on Competition Policy Papers, No. 330, "*Competition in artificial intelligence infrastructure*", de novembro de 2025 (aqui); (ii) Autorité de la Concurrence, "*Opinion 24-A-05 on the competitive functioning of the generative artificial intelligence sector*", de junho de 2025 (aqui); e (iii) Comissão Europeia, Policy brief "*Competition in Generative AI and Virtual Worlds*", de setembro de 2024 (aqui).

CONCORRÊNCIA E IA GENERATIVA: ACESSO A *CHIPS* DE IA – PRINCIPAIS MENSAGENS



Estrangulamentos estruturais na cadeia de valor de *chips* de IA podem comprometer o acesso ao mercado

As restrições persistentes do lado da oferta, a elevada concentração ao longo da cadeia de valor e o aumento da procura por hardware de ponta criaram estrangulamentos na produção e entrega de chips de IA, limitando a capacidade dos fornecedores de IA de menor dimensão de aceder à computação necessária para concorrer no desenvolvimento de IA avançada.



Iniciativas públicas de HPC podem expandir o acesso a computação para IA, ainda que a sua escala ainda seja limitada

A infraestrutura pública de HPC pode alargar o acesso a capacidade de computação em larga escala e mitigar, em parte, restrições do lado da oferta. Embora ainda limitadas em escala, têm existido iniciativas europeias para expandir o papel desta infraestrutura para desenvolvimento de IA. O seu impacto na concorrência dependerá de critérios de acesso transparentes, objetivos e não discriminatórios.



A expansão da atividade ao longo da cadeia de valor de IA pode aumentar a concentração de mercado

À medida que os fornecedores de serviços de *cloud* se expandem para o *design* de *chips* e os fornecedores de chips entram nos serviços de *cloud*, o controlo sobre o *hardware* e a infraestrutura poderá tornar-se mais concentrado. Tal poderá reforçar incentivos de ecossistema que retêm os fornecedores de IA, suscitando preocupações ao nível da interoperabilidade, de *self-preferencing* e de riscos de exclusão.

Os custos de mudança do ecossistema CUDA podem reforçar efeitos de *lock-in* e as barreiras à entrada

A posição do CUDA enquanto plataforma dominante para o desenvolvimento de IA, conjugada com a sua falta de compatibilidade com *hardware* alternativo, pode implicar custos de mudança elevados para os fornecedores de IA e levantar barreiras à entrada ou à expansão para fornecedores concorrentes de *chips* no mercado de *hardware* de IA.



A expansão para o fornecimento de infraestruturas de rede pode reforçar o poder de mercado ao longo da cadeia de valor de IA

À medida que os *designers* de *chips* alargam a sua atividade ao fornecimento de infraestruturas de rede para IA, podem reforçar efeitos de *lock-in*, quer através da integração técnica entre *hardware* e tecnologias de interligação, quer através de práticas contratuais. Tal dinâmica poderá aumentar o risco de alavancagem do poder de mercado detido ao nível dos *chips* para segmentos adjacentes da cadeia de valor da IA.