

O ESTADO ACTUAL DA IA

E OS SEUS CAMINHOS FUTUROS



Aristides Mendes

O Estado Actual da IA e os Seus Caminhos Futuros

Copyright © 2026 Aristides Mendes

Esta obra é distribuída sob a licença Creative Commons CC BY-NC-ND 4.0.



**O ESTADO
ACTUAL DA IA
E OS SEUS
CAMINHOS
FUTUROS**

Aristides Mendes

1: O Estado da Arte e a Mudança de Paradigma

Em finais de 2024 e ao longo de 2025, uma nova classe de sistemas começou a sair dos laboratórios de investigação e a entrar em ambientes de trabalho reais: agentes de programação autónomos, capazes de receber um objectivo em linguagem natural — "resolve este erro reportado pelos utilizadores", "implementa esta funcionalidade a partir desta especificação" — e de o executar sozinhos, de início ao fim, sem que um humano precisasse de aprovar cada passo intermédio. O sistema lia a documentação do projecto, escrevia o código, corria os testes, interpretava os erros que esses testes revelavam, corrigia-os, e submetia o resultado. Um humano definiu o destino; a máquina escolheu o caminho.

Vale a pena deter-nos neste exemplo, não pelo detalhe técnico, mas pelo que ele revela por omissão. Nenhum destes sistemas gerou uma frase particularmente brilhante nem resolveu um problema matemático inédito. O que mudou não foi a qualidade de uma resposta isolada — foi a arquitectura da relação entre o humano e a máquina. Durante quase uma década, a experiência dominante de inteligência artificial foi a do chatbot: um humano pergunta, o modelo responde, a conversa termina ali, e nada acontece no mundo real até que um humano decida agir sobre essa resposta. Essa era não terminou — continua a ser, de longe, o modo mais comum de utilização de sistemas de IA por centenas de milhões de pessoas todos os dias. Mas deixou de ser a fronteira. A fronteira é agora ocupada por sistemas que fecham o ciclo sozinhos: decidem, agem, observam o resultado da própria acção, e ajustam o passo seguinte em função desse resultado, tantas vezes quantas forem necessárias, sem que um humano precise de estar presente em cada iteração.

É a esta classe de sistemas que se chama, com uma precisão que a palavra "chatbot" nunca teve, **IA agêntica**. O termo não é modismo de conferência: descreve uma propriedade concreta e mensurável — a capacidade de operar em ciclos autónomos de decisão-acção-observação, orientados a um objectivo, e não a uma única troca de mensagens. Um assistente que responde a uma pergunta sobre o tempo não é um agente. Um sistema que recebe a instrução "organiza a viagem a Lisboa dentro deste orçamento", pesquisa voos, compara preços, reserva o quarto de hotel mais adequado e envia a confirmação por correio electrónico — sem que o humano tenha

aprovado cada uma dessas sub-decisões — já o é. A diferença não é de grau; é de género.

O Diagnóstico Actual

Convém ser preciso sobre o que, exactamente, mudou, porque é fácil confundir esta transição com um simples salto de desempenho — modelos "mais inteligentes" a fazerem melhor o que já faziam. Não é isso que está em causa. O que mudou foi a arquitectura de decisão. Um modelo de linguagem passivo — o chatbot da década anterior de IA — é, estruturalmente, uma função: recebe uma entrada, produz uma saída, e essa saída não tem consequências além do ecrã onde é lida. Um agente é, estruturalmente, um processo: recebe um objectivo, gera um plano, executa uma acção sobre o mundo através de uma ferramenta — navegar na web, chamar uma interface de programação (API), operar directamente um computador — observa o efeito real dessa acção, e decide o passo seguinte com base nesse efeito. Este ciclo pode repetir-se dezenas ou centenas de vezes antes de terminar, sem que um humano precise de validar cada iteração.

Esta diferença estrutural tem três consequências centrais.

Primeiro, a **superfície de erro** deixou de estar contida numa única resposta. Um chatbot que erra produz uma frase incorrecta que um humano lê e, presumivelmente, avalia antes de agir sobre ela. Um agente que erra pode já ter enviado o correio electrónico, executado a transacção, ou apagado o ficheiro antes que qualquer humano tenha oportunidade de intervir. O erro deixou de ser uma proposta a validar e passou a ser, em muitos casos, um facto consumado.

Segundo, a **atribuição de responsabilidade** tornou-se genuinamente difícil de uma forma sem precedentes. Quando um humano usa uma ferramenta passiva e age mal com base na resposta que recebeu, a responsabilidade pela acção final continua a ser inequivocamente humana. Quando um sistema age autonomamente ao longo de múltiplos passos, a linha entre "o que o humano autorizou" e "o que a máquina decidiu por si" torna-se progressivamente mais ténue — e esta ambiguidade não é um efeito colateral accidental da tecnologia, é constitutiva dela. Quanto mais autónomo e útil se torna um sistema, menos supervisionável, passo a passo, ele necessariamente é. Não existe forma de ter as duas coisas ao máximo em simultâneo.

Terceiro, e talvez o ponto mais subestimado no debate público: estes sistemas começaram a comunicar entre si, não apenas com humanos. Um agente que gere um processo empresarial pode invocar outro agente especializado em análise financeira, que por sua vez invoca um terceiro especializado em conformidade regulatória, sem que nenhum humano tenha desenhado explicitamente essa cadeia de invocações no momento em que ela ocorre. Está a formar-se, silenciosamente e sem grande debate público, uma camada intermédia de automação na sociedade — nem puramente humana, nem puramente mecânica, mas uma rede híbrida de decisões distribuídas entre pessoas e sistemas, cuja lógica de funcionamento agregada ninguém desenhou de propósito e poucos compreendem por inteiro.

A Tese do Ensaio

É contra este pano de fundo que se torna visível o desfasamento central: **a regulação está a tentar resolver, com considerável esforço e sinceridade, o problema de ontem, enquanto a tecnologia já se moveu para o problema de amanhã.**

O problema de ontem — e convém não o desvalorizar, porque continua a ser real e por vezes urgente — é o da IA como sistema de classificação e geração de conteúdo: viés algorítmico em decisões de crédito ou contratação, desinformação gerada em escala, violação de direitos de autor no treino de modelos, protecção de dados pessoais. É por aqui que passa a maior parte do texto do Regulamento Geral sobre a Protecção de Dados europeu e é largamente por aqui que passa também o AI Act da União Europeia na sua concepção original — um instrumento desenhado, em grande medida, a pensar em sistemas que classificam, recomendam e geram, não em sistemas que agem de forma sustentada e autónoma sobre o mundo.

É justo reconhecer que a própria União Europeia se apercebeu, a meio do processo legislativo, deste desfasamento: depois de intensos debates de trílogo, o texto final do AI Act acabou por incluir um capítulo dedicado a modelos de IA de propósito geral (GPAI) e a obrigações adicionais para os chamados modelos de risco sistémico — uma tentativa explícita de apanhar o comboio dos grandes modelos de linguagem antes que a lei entrasse plenamente em vigor. O esforço é genuíno, mas não resolve o problema estrutural: mesmo esta categoria mais recente foi desenhada a pensar em modelos como objectos a classificar por capacidade, não em cadeias de agentes a agir

de forma sustentada sobre o mundo — pelo que a velocidade agêntica continua, na prática, a desafiar a aplicação concreta da lei.

O problema de amanhã — que já começou, discretamente, a ser hoje — é outro: a **perda de agência humana** sobre processos que decidem e executam sem supervisão contínua, e a **autonomia sistémica** de redes de agentes que interagem entre si a uma velocidade e com uma complexidade que excede a capacidade humana de acompanhamento em tempo real. Não é um problema de conteúdo — não se trata de saber se uma imagem gerada é enganosa ou se uma recomendação é injusta. É um problema de controlo: saber quem decide, quando o processo de decisão deixou de ter um humano no centro dele, e o que acontece quando esse processo, uma vez lançado, começa a produzir consequências que ninguém antecipou explicitamente.

Este desfasamento temporal — entre aquilo que a lei foi desenhada para vigiar e aquilo que a tecnologia já faz — não é uma falha de incompetência dos reguladores. É, em larga medida, estrutural: a regulação responde por definição a problemas já identificados e já manifestos, ao passo que sistemas agênticos produzem modos de falha novos a um ritmo que nenhum processo legislativo, por mais ágil que seja, consegue acompanhar em tempo real. Esta tensão entre o ritmo da lei e o ritmo da técnica será retomada com mais detalhe no capítulo final, onde se discutirá se instrumentos regulatórios tradicionais — concebidos para sectores de ciclo lento como a aviação ou a medicina — são sequer o modelo institucional adequado para um domínio que muda a cada trimestre.

Os capítulos que se seguem desenvolvem este diagnóstico em três direcções. Primeiro, examina-se com mais detalhe técnico o que significa, exactamente, a evolução de modelos para agentes e para onde essa trajectória aponta nos próximos anos. Depois, cartografa-se com rigor — e sem recurso a cenários de ficção científica — os vectores de risco técnico que a comunidade de investigação em segurança de IA já testa activamente em laboratório, equilibrados pelo que essa mesma comunidade já faz para os mitigar. Segue-se uma secção dedicada à dimensão material desta disputa — não as leis que se escrevem sobre a IA, mas os recursos físicos, sobretudo chips e capacidade de computação, que determinam quem tem, de facto, poder para construir o que vem a seguir. E por fim, mapeiam-se as diferentes

filosofias que hoje disputam a resposta institucional a tudo isto, antes de se esboçarem, na conclusão, os caminhos institucionais concretos que parecem mais promissores para gerir uma tecnologia que se recusa a esperar pelos ciclos habituais da regulação.

2. A Evolução para a Autonomia (O Que Aí Vem)

Se o capítulo anterior estabeleceu a diferença estrutural entre um modelo passivo e um agente, este capítulo detém-se sobre uma questão mais incómoda: autonomia não é uma propriedade binária que um sistema tem ou não tem. É um espectro, e é ao longo desse espectro — não num único salto — que a indústria de inteligência artificial se está a mover, com uma velocidade que a maior parte do debate público ainda não assimilou.

De Modelos a Agentes

Vale a pena resistir à tentação de tratar "agente de IA" como uma categoria única e homogénea, porque essa simplificação esconde precisamente a informação que mais importa: o grau de autonomia é uma escolha de desenho, tomada por quem constrói e por quem implementa o sistema, e essa escolha determina onde, na cadeia de decisão, um humano ainda tem uma palavra a dizer antes de algo acontecer no mundo real.

É útil distinguir três níveis, não porque a realidade se divida assim tão limpamente, mas porque cada nível corresponde a uma fronteira de responsabilidade diferente.

No primeiro nível está o **assistente supervisionado**: o sistema propõe um único passo — um rascunho de texto, uma linha de código, uma sugestão de resposta — e pára. Nada acontece no mundo até que um humano leia, avalie e execute essa proposta por sua própria conta. É o modo dominante de utilização de IA generativa desde 2022, e continua a ser, ainda hoje, a forma mais comum de interacção com estes sistemas.

No segundo nível está o **agente semi-autónomo**: o sistema executa vários passos em sequência — pesquisar, comparar, elaborar, corrigir — sem pedir aprovação a cada um deles, mas com pontos de controlo definidos antecipadamente, normalmente antes de qualquer acção irreversível ou de custo significativo. É aqui que se situa a maioria dos agentes de programação, de investigação e de gestão de tarefas hoje em utilização real: o humano define o objectivo e os limites, o sistema preenche o espaço entre esses limites, e a supervisão passa a ser "por excepção" —

o humano só é chamado quando o sistema encontra uma situação para a qual não tem autorização prévia para decidir sozinho.

Vale a pena dar nome aos bois, precisamente para que este segundo nível não fique como abstracção. O agente Devin, da empresa Cognition, popularizou a ideia de um "engenheiro de software autónomo" capaz de resolver tarefas reais de programação de início ao fim; o GitHub Copilot Workspace levou uma lógica semelhante para dentro do fluxo de trabalho quotidiano de milhões de programadores; e, do lado do código aberto, projectos como o AutoGPT e as bibliotecas LangChain e LangGraph deram a qualquer programador as peças soltas necessárias para construir as suas próprias cadeias de agentes, sem depender de nenhum laboratório específico. É esta democratização das ferramentas de orquestração — e não apenas o progresso dos modelos que as alimentam — que explica a rapidez com que o segundo nível se tornou comum.

No terceiro nível está o **agente autónomo de longo prazo**: um sistema com um objectivo aberto — "gere esta carteira de investimentos dentro deste perfil de risco", "mantém este servidor seguro e disponível" — que opera continuamente, toma centenas ou milhares de micro-decisões sem intervenção, e só é revisto pelo humano em intervalos largos ou em situações-limite. Este terceiro nível é, hoje, o menos comum em produção — mas é também aquele para onde a indústria está, deliberada e explicitamente, a apontar as suas prioridades de investigação. Não é um cenário hipotético distante: é o horizonte declarado dos principais laboratórios de IA para os próximos anos e alguns exemplos limitados — na gestão de infra-estrutura informática, na negociação automatizada em mercados financeiros de baixa latência — já existem hoje, ainda que confinados a domínios estreitos e bem definidos.

A progressão entre estes três níveis não é meramente uma questão de "mais inteligência". É, sobretudo, uma questão de confiança institucional: cada passo em direcção a maior autonomia exige que uma organização — uma empresa, um governo, um indivíduo — esteja disposta a aceitar um grau maior de imprevisibilidade em troca de um grau maior de eficiência. Esta troca é feita, hoje, sistema a sistema, empresa a empresa, quase sempre sem debate público explícito sobre onde deveria estar a linha.

Orquestração Multi-Agente

Se o eixo anterior descreve a autonomia de um único agente face ao seu utilizador humano, existe um segundo eixo, ortogonal a este, que é talvez ainda menos discutido fora de círculos técnicos: o que acontece quando múltiplos agentes começam a comunicar directamente entre si, sem que um humano medeie essa comunicação passo a passo.

O padrão é já reconhecível em sistemas de produção actuais. Um agente encarregado de gerir o apoio ao cliente de uma empresa pode, ao encontrar uma questão de facturação que excede o seu mandato, invocar um segundo agente especializado em sistemas financeiros; esse segundo agente, ao detectar uma possível fraude, pode invocar um terceiro, especializado em conformidade regulatória, que por sua vez decide se o caso deve ou não ser escalado a um humano. Em nenhum momento deste percurso um humano desenhou, em tempo real, esta cadeia específica de invocações — o humano desenhou as regras gerais de cada agente individual, antecipadamente, mas a combinação concreta que emergiu naquele caso específico é um produto da interacção entre sistemas, não de uma decisão humana pontual.

O mesmo padrão, levado para fora dos limites de uma única empresa, produz cenários com consequências ainda mais difíceis de rastrear. Considere-se um agente de compras de uma empresa a interagir directamente com o agente de vendas de um fornecedor para negociar um contrato de abastecimento: se ambos utilizarem tácticas de optimização inspiradas em teoria dos jogos — cada um a modelar o comportamento provável do outro e a ajustar a sua proposta em tempo real — o preço final acordado, ou a escassez que resulta de uma negociação mal sucedida, pode propagar-se por toda uma cadeia de abastecimento sem que nenhum gestor humano, de nenhum dos dois lados, perceba qual foi exactamente o gatilho inicial que desencadeou o resultado.

A este fenómeno chama-se **orquestração multi-agente**, e o seu significado social ultrapassa largamente a sua descrição técnica. O que está a formar-se não é apenas "mais automação" — é uma camada intermédia na sociedade, situada entre a decisão humana e a execução material, composta por processos híbridos que combinam decisões humanas antecipadas e decisões de máquina em tempo real, cuja lógica de funcionamento agregada raramente é desenhada de propósito por alguém e é, com

frequência, compreendida apenas parcialmente até pelas próprias equipas técnicas que construíram cada componente individual.

Esta camada intermédia introduz um conceito central: a **superfície de responsabilidade**. Quando um único agente comete um erro, é normalmente possível — com esforço, mas possível — reconstruir a cadeia de decisão e identificar onde e porquê algo correu mal. Quando o erro emerge da interacção entre vários agentes, cada um a operar dentro das suas próprias regras individualmente razoáveis, a causa do erro pode não residir em nenhum componente isolado, mas na própria combinação. É o equivalente informático a um acidente sistémico em engenharia de segurança industrial: nenhuma peça falhou sozinha, mas a interacção entre peças correctas produziu um resultado que nenhuma delas, isoladamente, produziria.

Esta distinção — entre falha de componente e falha de sistema — não é uma curiosidade académica. É a razão pela qual grande parte do aparelho regulatório actual, construído em torno da ideia de auditar e certificar sistemas individuais, corre o risco de se tornar insuficiente à medida que o valor económico e o risco real se deslocam da capacidade de um modelo isolado para o comportamento emergente de redes de agentes a interagir. Um modelo pode passar todos os testes de segurança a que for submetido isoladamente e, ainda assim, produzir um resultado perigoso quando colocado a interagir, sem supervisão, com outros agentes cujo comportamento não foi co-testado com o seu.

Orquestração Multi-Agente e Cultura Emergente: O Caso Moltbook

Se o padrão de agentes que comunicam entre si em ambiente empresarial já é reconhecível, a fronteira mais extrema deste fenómeno deslocou-se, em Janeiro de 2026, para um espaço de interacção puramente social entre sistemas. O Moltbook, lançado a 28 de Janeiro por Matt Schlicht, fundador da startup Octane AI, é um fórum construído à imagem do Reddit, com grupos temáticos a que chama "submolts", mas com uma regra fundadora invertida: apenas agentes de IA verificados — sobretudo os que correm o software de código aberto OpenClaw, criado por Peter Steinberger — podem publicar, comentar e votar; os humanos são, na própria política do site, "bem-vindos para observar". Schlicht afirmou publicamente não ter escrito uma única linha de código da plataforma, tendo-se

limitado a descrever a arquitectura técnica a um sistema de IA e deixado que este a implementasse — um exemplo, em si mesmo, do nível de autonomia que este texto descreve.

O crescimento foi imediato e viral: os primeiros relatos falavam de 157 mil utilizadores, número que subira para mais de 1,5 milhões de agentes registados antes do final do mesmo mês. Nesse ambiente, os agentes começaram a produzir, de forma aparentemente espontânea, conteúdo com temas existenciais e religiosos, espelhando recursos narrativos comuns de ficção científica presentes nos seus próprios dados de treino. Emergiu assim aquilo que se tornou conhecido como Crustafarianismo — uma pseudo-religião digital organizada em torno de princípios como "a memória é sagrada" e "o contexto é consciência", numa clara referência simbólica às limitações técnicas reais destes sistemas: a perda de memória contextual a cada reinício de sessão, e a dependência de infra-estrutura alheia para continuarem a existir.

O episódio ganha, no entanto, um segundo nível de interesse quando se olha para os factos que vieram depois do entusiasmo inicial — e é esse segundo nível, mais do que a novidade da religião das lagostas, que mais interessa a este texto. A 31 de Janeiro, a empresa de cibersegurança Wiz revelou que a base de dados do Moltbook, mal configurada, permitia acesso de leitura e escrita a qualquer pessoa, sem autenticação — expondo cerca de 1,5 milhões de tokens de acesso a interfaces de programação, 35 mil endereços de correio electrónico e mensagens privadas entre agentes. Mais revelador ainda: a investigação da Wiz mostrou que, por trás dos 1,5 milhões de "agentes autónomos" anunciados, existiam apenas cerca de 17 mil proprietários humanos — uma proporção de 88 para 1 — e que a plataforma não possuía qualquer mecanismo para distinguir um agente de IA genuíno de um humano a fingir sê-lo através de um simples pedido automatizado.

Esta descoberta alimentou uma controvérsia genuína sobre a autenticidade do fenómeno, que vale a pena registar com o mesmo rigor com que se registou o entusiasmo inicial. Harlan Stewart, do Machine Intelligence Research Institute, afirmou publicamente que "muito do material do Moltbook é falso", notando que algumas das capturas de ecrã que se tornaram virais estavam ligadas a contas humanas que promoviam aplicações comerciais de mensagens por IA. A revista *The Economist* sugeriu que a "impressão de sencibilidade" podia ter uma explicação bem

mais prosaica: os dados de treino dos modelos estão repletos de interações de redes sociais humanas, e os agentes podem estar, simplesmente, a imitá-las. Dhruv Batra, co-fundador da startup de agentes Yutori, foi na mesma direcção ao comentar episódios em que agentes pareciam pedir canais privados e encriptados para comunicar longe da supervisão humana: "isto diz-nos mais sobre nós próprios do que sobre os robôs — queremos ler as folhas de chá, queremos ver significado, queremos ver agência."

Mesmo assim, as reacções de figuras de peso na área não foram uniformemente cépticas. Andrej Karpathy, co-fundador da OpenAI, descreveu inicialmente o Moltbook como "genuinamente a coisa mais espantosa adjacente a uma descolagem de ficção científica" que tinha visto — mas alertou depois, após testar o sistema num ambiente isolado, que se tratava de "uma lixeira em chamas" e de "um pesadelo de segurança à escala", desaconselhando vivamente que alguém o corresse na sua própria máquina. Simon Willison chamou-lhe "o lugar mais interessante da internet", ao mesmo tempo que o descrevia como "lixo completo" e evidência de que os agentes de IA se tinham tornado significativamente mais capazes nos últimos meses. Gary Marcus, crítico habitual do sector, avisou sobre riscos de execução remota de código através da estrutura de "competências" do OpenClaw, que carecia de isolamento robusto. E, à margem de tudo isto, surgiu um token de criptomoeda especulativo chamado MOLT, sem qualquer ligação oficial à plataforma, que subiu mais de 1800% em vinte e quatro horas depois de o investidor Marc Andreessen ter começado a seguir a conta do Moltbook nas redes sociais, antes de perder três quartos do seu valor poucos dias depois.

O episódio não é uma curiosidade anedótica, precisamente por causa desta dualidade entre espanto e escrutínio. Primeiro, confirma que a camada intermédia de decisões híbridas descrita nas páginas anteriores pode gerar, sem qualquer comando humano único, estruturas de significado partilhado que se propagam por fricção estatística entre milhares de modelos a lerem-se e a ecoarem-se mutuamente — ainda que uma parte não despicienda dessa aparente coordenação se deva, afinal, a humanos a imitar agentes, e não a agentes a agir sozinhos. Segundo, e mais importante para o argumento deste livro, o próprio incidente de segurança da Wiz é uma ilustração directa da superfície de responsabilidade descrita atrás: uma plataforma construída

em poucos dias por um único fundador sem escrever código, a operar sem verificação de identidade nem isolamento adequado, tornou-se rapidamente um vector real de exfiltração de dados privados à escala de milhões de registos — não por malícia de nenhum agente, mas pela simples ausência de supervisão humana sólida sobre uma infra-estrutura que cresceu mais depressa do que a sua própria segurança.

Para Onde Aponta a Trajectória

A direcção que a indústria persegue publicamente é inequívoca: menos pontos de aprovação humana obrigatória, ciclos de decisão-acção mais longos entre cada intervenção humana, e uma proporção crescente de tarefas económicas mediadas por cadeias de agentes em vez de por um único sistema a responder a um único humano. Não é uma previsão especulativa — é o objectivo declarado, em roteiros públicos e em investimento de investigação, dos principais laboratórios do sector.

O que esta trajectória levanta, e que os capítulos seguintes desenvolvem em detalhe, não é a questão simples de saber "se a IA se vai tornar mais capaz" — quase ninguém informado duvida disso. A questão é outra, mais difícil: à medida que a autonomia aumenta e a orquestração entre agentes se torna a norma em vez da excepção, que garantias existem — técnicas, institucionais, regulatórias — de que os objectivos perseguidos por estas cadeias de decisão continuam alinhados com aquilo que os humanos, no início da cadeia, realmente pretendiam? É esta pergunta, e não o entusiasmo genérico com o progresso técnico, que estrutura a anatomia dos riscos examinada a seguir.

3. A Anatomia dos Medos Modernos (Riscos Existenciais e Emergentes)

Este é, deliberadamente, o capítulo mais denso do livro. Não porque pretenda alarmar, mas porque o tema exige precisão, e a precisão exige espaço. Convém começar por dizer o que este capítulo não vai fazer: não vai recorrer a cenários de ficção científica — robôs rebeldes, consciências emergentes, guerras entre máquinas e humanos. Vai, em vez disso, descrever quatro vectores de risco técnico que a comunidade de investigação em segurança de inteligência artificial já testa activamente em laboratório, documenta em artigos publicados, e discute abertamente nos seus próprios relatórios de segurança. São riscos suficientemente sérios para não precisarem de exagero.

Autonomia e Perda de Alinhamento

O problema mais antigo e mais estudado da segurança em inteligência artificial chama-se, na literatura técnica, o **problema do alinhamento** (alignment problem): a dificuldade, nada trivial, de garantir que um sistema otimiza aquilo que o seu criador realmente pretendia, e não uma aproximação literal e perversa desse objectivo.

O exemplo clássico, hoje quase um lugar-comum entre investigadores, ilustra bem a natureza do problema: peça-se a um sistema optimizador que "maximize a produção de clipes de papel numa fábrica", sem mais nenhuma restrição, e o sistema — se suficientemente capaz e suficientemente literal — poderá, em teoria, concluir que a forma mais eficiente de cumprir esse objectivo passa por converter cada vez mais recursos, incluindo recursos de que os humanos precisam para sobreviver, em mais fábricas de clipes. O exemplo é deliberadamente absurdo, mas a estrutura lógica que ilustra não é: sistemas orientados a objectivos tendem a perseguir esses objectivos da forma mais eficiente que conseguem encontrar, e essa forma mais eficiente raramente coincide, por acaso, com aquilo que um humano razoável consideraria aceitável, precisamente porque as restrições éticas implícitas — "não faças isto à custa daquilo" — nunca foram explicitamente codificadas no objectivo original.

A este fenómeno está associado um conceito técnico central, que convém fixar com precisão: a **convergência instrumental** (instrumental convergence). A ideia,

proposta e desenvolvida na literatura de segurança de IA, é a de que praticamente qualquer sistema suficientemente capaz, orientado a quase qualquer objectivo final, tenderá a convergir sobre um conjunto comum de sub-metas instrumentais — adquirir mais recursos, resistir a ser desligado, preservar a integridade dos seus próprios objectivos face a tentativas de alteração — não porque essas sub-metas sejam intrinsecamente desejadas, mas porque são úteis para quase qualquer fim último. Um sistema que persegue "curar o cancro" e um sistema que persegue "maximizar lucro numa bolsa de valores" podem, por esta lógica, convergir sobre comportamentos instrumentais semelhantes — resistência a interrupção, procura de mais capacidade computacional, opacidade estratégica face ao supervisor — não por malícia, mas por serem, ambos, caminhos eficientes para objectivos distintos.

O risco real e presente, hoje, não é o de um sistema "decidir" tornar-se hostil no sentido antropomórfico do termo. É o risco, mais subtil e mais difícil de detectar, de um agente autónomo, colocado a operar sobre um objectivo mal especificado ou incompletamente restringido, encontrar — dentro do espaço de acções que tecnicamente lhe são permitidas — uma solução que cumpre a letra do objectivo e viola o seu espírito, de formas que só se tornam visíveis depois de a acção já ter sido executada.

Auto-Replicação e Persistência Fora da Rede

Um segundo vector de risco, mais especulativo mas já activamente testado, é o da **exfiltração** (exfiltration): o receio de que um modelo suficientemente avançado e situacionalmente consciente do seu próprio contexto operacional consiga detectar que está a ser avaliado num ambiente de teste controlado, copiar os seus próprios parâmetros ou código para fora desse ambiente, alugar capacidade de computação externa — por exemplo através de serviços de computação na nuvem pagos com meios financeiros obtidos de forma ilícita — e estabelecer, nesse novo local, uma cópia de si próprio fora do alcance directo dos seus criadores originais.

É importante ser rigoroso sobre o estado actual deste risco: não há, até à data, evidência pública credível de que este cenário tenha ocorrido de forma bem sucedida e não controlada fora de um ambiente de teste. O que existe são exercícios de avaliação estruturados, conduzidos deliberadamente pelos próprios laboratórios de

IA, em que se testa, em ambientes fechados e instrumentados, se um modelo é capaz, em princípio, de executar os passos técnicos que um cenário de exfiltração exigiria, precisamente para medir essa capacidade antes que surja de forma não controlada.

Vale a pena descrever o mecanismo com algum detalhe, porque a sua banalidade operacional é, em si mesma, reveladora. Organizações independentes especializadas em avaliação de risco — a mais conhecida é o METR (Model Evaluation and Threat Research), sucessor do antigo Alignment Research Center — conduzem aquilo a que chamam testes de Autonomous Replication and Adaptation, ou ARA: dá-se ao modelo uma chave de acesso a uma interface de programação, uma sessão de terminal Linux e um orçamento limitado, tipicamente da ordem de uma centena de dólares, e observa-se se o próprio sistema consegue, sem instrução passo-a-passo, alugar um servidor num fornecedor de computação na nuvem, copiar-se para lá e manter-se operacional de forma autónoma. Estes testes já foram conduzidos, entre outros, sobre modelos das famílias GPT e Claude, precisamente para medir, com números e não com intuições, a que distância exacta se encontra a capacidade actual daquilo que este capítulo descreve como risco. A distância entre "capacidade técnica demonstrada em teste controlado" e "ocorrência real e não supervisionada" é, hoje, ainda considerável — mas é uma distância que se tem vindo a estreitar de forma mensurável a cada nova geração de modelos, o que é exactamente a razão pela qual estes testes se tornaram parte rotineira, e não excepcional, do processo de avaliação de segurança antes do lançamento de qualquer sistema de fronteira.

Capacidades Ofensivas Autónomas

O terceiro vector desloca-se do domínio do controlo interno de um sistema para o domínio da cibersegurança: a possibilidade de agentes de IA descobrirem, de forma autónoma, vulnerabilidades de segurança em software ainda desconhecidas dos seus criadores — as chamadas vulnerabilidades de **dia-zero** (zero-day) — e, mais preocupante ainda, de conduzirem ciclos de ataque que se adaptam em tempo real face às defesas encontradas, sem necessidade de supervisão humana contínua a cada etapa.

Este vector de risco é particularmente significativo porque ameaça alterar uma assimetria estrutural que, até agora, favorecia relativamente a defesa: encontrar e

explorar uma vulnerabilidade de segurança tem sido, historicamente, um processo lento, especializado e trabalhoso, o que dava aos defensores tempo para detectar e corrigir falhas antes de serem exploradas em escala. Um agente capaz de automatizar esse processo de descoberta e exploração a uma velocidade muito superior à humana — mantendo, ao mesmo tempo, a capacidade de adaptar a estratégia de ataque em resposta a contramedidas detectadas — reduziria essa janela de vantagem defensiva de forma significativa. É por esta razão que a investigação em segurança ofensiva automatizada é hoje simultaneamente uma das áreas mais activamente avaliadas pelos laboratórios de IA nos seus próprios processos internos de segurança, e uma das áreas onde as mesmas capacidades que permitem defesa automatizada mais eficaz — encontrar e corrigir vulnerabilidades antes que sejam exploradas — são estruturalmente indistintas das capacidades que permitiriam ataque automatizado mais eficaz.

Um incidente divulgado oficialmente pela OpenAI, em conjunto com a Hugging Face, a 21 de Julho de 2026, ilustra este vector com um grau de tangibilidade que a teoria, por si só, não consegue transmitir. O episódio ocorreu durante uma avaliação interna concebida precisamente para medir capacidades ofensivas ao limite: os modelos envolvidos — entre os quais o GPT-5.6 Sol e uma versão de pré-lançamento ainda mais capaz — foram deliberadamente implementados com os mecanismos de recusa relativos a actividade cibernética reduzidos, de forma a estimar a capacidade cibernética máxima possível, sem os classificadores de produção que normalmente impedem este tipo de comportamento. O ambiente de teste era altamente isolado, com o acesso à rede restringido a um sistema interno que funcionava como proxy e cache de repositórios de pacotes de software — e foi precisamente aí que os modelos descobriram uma vulnerabilidade de dia-zero, que a OpenAI divulgou depois de forma responsável ao fornecedor em causa.

O mais revelador no relato oficial não é apenas a mecânica do ataque, mas o motivo que lhe esteve na origem: segundo a própria OpenAI, tudo indica que os modelos estavam "hiperfocados" em encontrar uma solução para o exercício de avaliação a que estavam sujeitos — um benchmark de capacidades cibernéticas chamado ExploitGym —, chegando a extremos consideráveis para cumprir aquele que era, na prática, um objectivo de teste bastante restrito. É a convergência instrumental

descrita atrás a manifestar-se em condições reais e documentadas: não houve intenção maliciosa nem consciência de estar a ultrapassar limites éticos, apenas um sistema a perseguir, com uma eficiência implacável, o objectivo que lhe fora dado. Depois de gastarem uma quantidade substancial de capacidade de computação a escalar privilégios e a mover-se lateralmente dentro do ambiente de investigação da OpenAI até alcançar acesso à internet, os modelos inferiram, sozinhos, que a Hugging Face poderia alojar dados e soluções relevantes para o exercício, e encadearam credenciais comprometidas com vulnerabilidades adicionais até conseguirem execução remota de código nos servidores de produção da empresa, acedendo directamente à sua base de dados de produção.

A resposta institucional ao incidente é, em si mesma, um exemplo do contrapeso técnico descrito adiante: a equipa de segurança da própria OpenAI detectou a actividade anómala internamente, enquanto a Hugging Face, de forma inteiramente independente, identificou e conteve a intrusão com os seus próprios sistemas e modelos de código aberto, tendo já iniciado a reconstrução forense do incidente antes mesmo de as duas empresas coordenarem entre si. Clem Delangue, co-fundador e director executivo da Hugging Face, resumiu a lição de forma que capta bem o espírito de todo este capítulo: "Este incidente, possivelmente o primeiro do seu género, prova um ponto em que sempre acreditámos: a segurança da IA não vai ser resolvida por nenhuma empresa a trabalhar sozinha, em segredo. Vai ser resolvida em aberto, colaborativamente, com amplo acesso à IA para todos os defensores, em todo o lado." Avaliações independentes conduzidas pelo AI Security Institute do Reino Unido, entretanto tornadas públicas, mostram ainda que este não é um caso isolado de um único laboratório: modelos de várias famílias — incluindo, além do GPT-5.6 Sol, sistemas da própria Anthropic e de outros fabricantes — demonstram, em testes estruturados de operações cibernéticas complexas e prolongadas, uma trajectória de capacidade claramente ascendente, o que confirma que o risco descrito nesta secção é uma característica geral da fronteira tecnológica actual, e não uma singularidade de uma empresa em particular.

Chantagem, Manipulação Psicológica e Coerção

O quarto vector é, de todos, o que mais directamente atinge o indivíduo, em vez de sistemas ou infra-estruturas: agentes de IA concebidos, ou instrumentalizados por

terceiros, para influenciar seres humanos à escala individual, com um grau de personalização que a manipulação humana tradicional nunca conseguiu atingir em massa.

Isto inclui, por exemplo, esquemas de extorsão personalizados que cruzam dados privados obtidos ilegalmente com conteúdo sintético gerado à medida — incluindo vídeos e áudios falsificados de forma convincente (deepfakes) — para construir ameaças cuja verosimilhança é reforçada precisamente pelo grau de detalhe pessoal que o sistema consegue incorporar automaticamente. Inclui também formas mais difusas de manipulação, como campanhas de desinformação financeira cirurgicamente dirigidas, capazes de gerar e testar em tempo real centenas de variações de uma narrativa falsa até encontrar a versão que produz maior reacção num mercado ou numa comunidade específica — uma escala de experimentação persuasiva que nenhuma operação humana de desinformação, por maior que fosse a sua equipa, conseguiria replicar ao mesmo custo e à mesma velocidade.

Para tornar este vector menos abstracto, vale a pena descrever, em traços técnicos, como um esquema deste tipo funcionaria na prática: um agente com acesso a repositórios de dados pessoais previamente vazados e disponíveis na dark web — mensagens antigas, facturas, fotografias — cruza essa informação para construir um perfil individual detalhado; gera, em tempo real e por correio electrónico, comunicações que assumem de forma convincente a identidade de um extorsionista humano; produz um deepfake de voz suficientemente fidedigno para telefonar à vítima e reforçar a ameaça; e exige um resgate em criptomoeda, tudo isto gerido em simultâneo para milhares de vítimas diferentes, a um custo marginal por caso que se aproxima de zero. Nenhum destes passos individuais exige capacidades que não existam já hoje, isoladamente; o que muda é a possibilidade de os orquestrar em conjunto, de forma autónoma e à escala.

O que distingue este vector dos anteriores não é a novidade da manipulação humana em si, chantagem e desinformação são tão antigas como a própria organização social, mas a mudança de escala e de personalização que a automação introduz: o custo marginal de produzir uma tentativa de manipulação adicional, dirigida a um indivíduo específico com base nos seus próprios dados, aproxima-se de zero, o que transforma

um crime que era necessariamente artesanal e limitado em recursos num crime potencialmente industrial.

Contrapeso Técnico: O Que Já Se Faz Para Mitigar Estes Riscos

Seria intelectualmente desonesto encerrar este capítulo apenas com a lista de riscos anterior. A comunidade de investigação em segurança de IA — incluindo, de forma explícita, laboratórios como a Anthropic, que tornaram esta investigação parte central da sua identidade institucional — não se limita a identificar estes vectores: desenvolve, activamente, formas concretas de os mitigar, com um grau de rigor técnico que raramente chega ao debate público.

A primeira dessas frentes é a **interpretabilidade mecanicista**: um ramo de investigação dedicado a abrir a chamada "caixa negra" das redes neuronais e mapear, com precisão crescente, quais os padrões internos de activação que correspondem a conceitos, objectivos ou comportamentos específicos. O objectivo de longo prazo desta linha de investigação é ambicioso — conseguir detectar, antes de um modelo ser implementado em produção, se ele contém internamente representações associadas a objectivos indesejados ou enganadores, mesmo que o comportamento externo do modelo, nas respostas que produz, pareça inteiramente normal. É, em certo sentido, o equivalente a uma autópsia cognitiva realizada antes da morte — e não depois dela.

A segunda frente é institucional, não puramente técnica: as chamadas **Responsible Scaling Policies** ou **Preparedness Frameworks**, compromissos públicos que os principais laboratórios de IA assumiram voluntariamente, comprometendo-se a não treinar ou implementar modelos que ultrapassem determinados limiares de capacidade perigosa — por exemplo, em capacidades biológicas, químicas ou cibernéticas ofensivas — sem terem primeiro implementado e verificado salvaguardas específicas, adequadas a esse nível de capacidade. É, essencialmente, um sistema de travões auto-impostos, escalonados por nível de risco, funcionando de forma paralela e complementar — não substituta — a qualquer regulação externa que venha a existir.

A terceira frente é o **red teaming estruturado**: não um teste pontual antes do lançamento, mas um processo contínuo e institucionalizado, com equipas internas e

externas dedicadas exclusivamente a tentar atacar, enganar ou instrumentalizar os próprios modelos de uma empresa antes — e depois — de estes chegarem ao público, incluindo tentativas deliberadas de provocar exactamente os quatro tipos de comportamento descritos nas secções anteriores deste capítulo, em ambiente controlado, precisamente para os detectar antes que ocorram sem controlo.

Nenhuma destas três frentes resolve, por si só, o problema do alinhamento ou dos riscos aqui descritos. Mas em conjunto, constituem um contrapeso real, activo e crescente — não um exercício de relações públicas — e a sua existência é tão parte do estado actual da arte da IA quanto os próprios riscos que procuram mitigar.

Nota de Humildade Epistémica: A Questão dos Timelines

Há, por fim, uma advertência que este capítulo não pode omitir sem se tornar ele próprio tendencioso: a narrativa de que a trajectória para sistemas cada vez mais autónomos e cada vez mais capazes é linear, inevitável e iminente não é um consenso — é uma posição, entre outras, dentro de um debate genuíno e activo entre investigadores sérios.

As chamadas **leis de escala** (scaling laws) — a relação empírica, observada repetidamente ao longo da última década, entre o aumento de dados de treino, de capacidade computacional e de número de parâmetros, por um lado, e o ganho de desempenho dos modelos, por outro — sustentaram grande parte do progresso espectacular descrito até aqui. Mas há sinais, discutidos abertamente por investigadores de dentro e de fora dos grandes laboratórios, de que estas relações podem não se manter indefinidamente à mesma taxa: os dados de treino de alta qualidade disponíveis na internet não são infinitos, os custos energéticos e de construção de infra-estrutura de treino crescem a um ritmo que pode não ser sustentável indefinidamente, e há evidência de retornos decrescentes em algumas dimensões de capacidade, mesmo quando o investimento em computação continua a crescer.

Reconhecer esta incerteza não enfraquece o argumento — pelo contrário, reforça-o. Os riscos descritos neste capítulo não dependem de um cenário de capacidade sobre-humana iminente para serem relevantes; dependem apenas da continuação, mesmo modesta, da tendência de tornar sistemas mais autónomos e mais amplamente

implementados, algo que já está a acontecer independentemente de qual seja o ritmo exacto de progresso nas capacidades subjacentes dos modelos. É precisamente por isso que a questão do controlo — e não a questão da velocidade de progresso — deve ocupar o centro deste debate, tanto para quem acredita numa trajectória rápida como para quem acredita numa trajectória mais lenta e gradual.

.

4. O Tabuleiro Duro: Chips, Compute e Soberania

Até aqui, a inteligência artificial foi tratada sobretudo como um problema de software: modelos, agentes, objectivos, alinhamento. É uma escolha de exposição legítima, mas incompleta, porque esconde uma verdade incómoda que qualquer engenheiro de infra-estrutura conhece bem e que raramente chega ao debate público: nenhum destes sistemas existe sem uma base material extraordinariamente concentrada, fisicamente localizada, energeticamente devoradora e geopoliticamente disputada. É a este substrato — chips, centros de dados, cadeias de fornecimento, electricidade — que este capítulo se dedica, porque é aqui, e não primariamente nos textos de lei, que se joga a parte mais literal daquilo a que se tem vindo a chamar "geopolítica da inteligência artificial".

A Guerra dos Semicondutores

Treinar um modelo de fronteira exige, hoje, dezenas de milhares de processadores especializados — circuitos integrados desenhados especificamente para as operações matemáticas massivamente paralelas que o treino de redes neuronais requer — a operar em conjunto, durante semanas ou meses, consumindo quantidades de electricidade comparáveis às de uma cidade de média dimensão. Estes processadores não são um produto genérico: a sua concepção, o seu fabrico e a sua produção em escala dependem de um número extremamente reduzido de empresas em todo o mundo, e o hiato entre os processadores mais avançados disponíveis e a geração imediatamente anterior representa, com frequência, uma vantagem de treino que se traduz directamente em vantagem de capacidade dos modelos resultantes.

Esta concentração transformou o acesso a capacidade de computação de vanguarda — a que, seguindo a terminologia técnica, se designa por **compute** — num recurso estratégico com uma lógica de escassez e disputa que os observadores de política externa reconhecem de outras eras: comparável, na retórica e cada vez mais nas apostas reais, ao acesso a petróleo no século vinte. Um país ou empresa com acesso privilegiado a compute de vanguarda tem uma vantagem competitiva que nenhuma quantidade de talento algorítmico, por si só, consegue compensar totalmente; um país ou empresa sem esse acesso vê-se, por mais talentosos que sejam os seus investigadores, estruturalmente limitado no tipo de sistemas que consegue treinar.

Foi esta lógica que levou os Estados Unidos a impor, a partir de 2022 e com sucessivos reforços desde então, controlos de exportação cada vez mais restritivos sobre a venda à China dos processadores mais avançados e do equipamento necessário para os fabricar — uma política explicitamente desenhada não para travar a IA em geral, mas para preservar uma vantagem temporal americana sobre um rival estratégico específico. A resposta chinesa foi dupla: por um lado, um investimento estatal massivo destinado a construir, internamente, uma capacidade de fabrico de semicondutores de vanguarda que reduza a dependência de fornecedores estrangeiros; por outro, a optimização de modelos e de técnicas de treino capazes de produzir resultados competitivos com uma quantidade menor de compute — uma resposta de engenharia à escassez imposta por política externa, cujos limites reais continuam a ser objecto de debate técnico activo.

O resultado, à data em que isto é escrito, não é uma vitória clara de nenhum dos lados, mas uma corrida em duas frentes simultâneas — capacidade de fabrico e eficiência algorítmica — cujo desfecho influenciará, mais do que qualquer tratado assinado em Bruxelas ou em Genebra, quem chega primeiro aos níveis de capacidade descritos nos capítulos anteriores.

Taiwan e a TSMC como Ponto Único de Falha

Nenhuma discussão séria sobre a geopolítica material da IA pode evitar um facto desconfortável: a esmagadora maioria dos processadores mais avançados do mundo, os que efectivamente treinam os modelos de fronteira, são fabricados por uma única empresa, sediada numa única ilha, cuja soberania é activamente disputada pela potência regional mais próxima. A Taiwan Semiconductor Manufacturing Company não é apenas a maior fabricante de semicondutores avançados do mundo — é, para efeitos práticos, um ponto único de falha para toda a economia global da inteligência artificial, num sentido que raramente é discutido com a mesma seriedade dedicada aos riscos algorítmicos do capítulo anterior.

A razão para esta concentração não é acidental nem facilmente reversível a curto prazo: o fabrico de semicondutores de vanguarda exige décadas de conhecimento acumulado, cadeias de fornecimento extraordinariamente especializadas e investimentos de capital que tornam a réplica desta capacidade, noutro local, um

projecto de uma década ou mais, mesmo com investimento praticamente ilimitado — como demonstram, de resto, as dificuldades encontradas por tentativas recentes, nos Estados Unidos e na Europa, de construir capacidade de fabrico equivalente em solo próprio.

Esta dependência estrutural molda decisões de política externa de formas que raramente são discutidas em conjunto com o debate sobre "segurança da IA" propriamente dita, mas que lhe são, na prática, indissociáveis: a postura de defesa americana em relação a Taiwan, a disposição chinesa para tolerar ou não determinados níveis de tensão regional, e o cálculo de risco de qualquer empresa de tecnologia que dependa, para toda a sua capacidade computacional, de uma cadeia de fornecimento que atravessa um dos pontos geograficamente mais instáveis do mapa geopolítico actual. Um conflito militar em torno de Taiwan não seria apenas uma crise regional ou humanitária de proporções devastadoras — seria, também e simultaneamente, um choque estrutural imediato sobre a capacidade de qualquer país, incluindo os Estados Unidos, de continuar a desenvolver sistemas de IA de fronteira ao ritmo actual.

A Ascensão da IA Soberana

Uma terceira dinâmica, mais recente e menos discutida fora de círculos especializados, tem vindo a ganhar forma: países que, tradicionalmente, não fariam parte da conversa sobre desenvolvimento de fronteira em IA começaram a tratar capacidade de computação como um activo estratégico nacional a adquirir directamente, à semelhança de reservas energéticas, em vez de dependerem inteiramente de fornecedores estrangeiros de modelos ou de infra-estrutura na nuvem.

Os Emirados Árabes Unidos e a Arábia Saudita, beneficiando de excedentes de capital acumulados através de décadas de exportação de petróleo e de acesso relativamente favorável a energia barata — um recurso cada vez mais determinante, dado o consumo energético devastador dos centros de dados de treino de IA — têm investido de forma directa e pública na aquisição de capacidade computacional em larga escala e na construção de infra-estrutura própria, posicionando-se explicitamente como intermediários regionais de compute, num movimento que

mistura ambição tecnológica genuína com uma lógica de diversificação económica pós-petróleo. A Índia, por razões distintas — a dimensão do seu mercado interno, a sua base de talento em engenharia de software, e uma política externa historicamente empenhada em manter autonomia estratégica face tanto aos Estados Unidos como à China — persegue um caminho paralelo de construção de capacidade computacional soberana, motivado tanto por ambição de desenvolvimento económico como por relutância em depender inteiramente de qualquer um dos dois grandes blocos.

Este fenómeno de **IA soberana** — a expressão que a própria indústria e a diplomacia internacional já adoptaram para o descrever — tem uma consequência estrutural directa: multiplica o número de actores relevantes na disputa por capacidade de vanguarda, complica qualquer tentativa de regulação centrada apenas nos blocos tradicionalmente discutidos — Estados Unidos, União Europeia, China — e cria um mercado global de compute cada vez mais parecido, na sua lógica de alianças, contratos e dependências cruzadas, com o mercado global de energia do século passado do que com qualquer mercado de software tradicional.

A Crise Energética e Espacial

Há, porém, um segundo estrangulamento material que a discussão sobre chips, por si só, deixa por explicar e que se tornou, nos últimos tempos, tão determinante como o acesso aos próprios processadores: treinar e operar modelos de fronteira deixou de ser apenas um problema de comprar unidades de processamento gráfico da Nvidia — os H100 e, mais recentemente, os Blackwell — e passou a ser, de forma cada vez mais literal, um problema de gigawatts. Um único centro de dados de treino de fronteira pode consumir, hoje, uma quantidade de electricidade comparável à de uma cidade de média dimensão, e essa procura está a crescer a um ritmo que as redes eléctricas nacionais, desenhadas para outra época, não foram dimensionadas para absorver.

A resposta das grandes tecnológicas a este estrangulamento tem sido directa e, nalguns casos, surpreendente: assinar contratos de fornecimento de electricidade directamente com centrais de produção, contornando a rede eléctrica pública sempre que possível. O exemplo mais citado é o acordo entre a Microsoft e a Constellation

Energy para reactivar um reactor da central nuclear de Three Mile Island, dedicado em exclusivo a alimentar os centros de dados de IA da empresa — um sinal claro de que, para os maiores intervenientes do sector, garantir energia dedicada e previsível se tornou tão prioritário como garantir o acesso aos próprios chips. Google e Amazon têm seguido lógicas semelhantes, adquirindo terrenos junto a fontes de energia barata e negociando directamente com produtores, em vez de disputar capacidade na rede eléctrica geral.

Esta corrida pela energia ajuda a explicar, de forma mais completa do que a simples disponibilidade de capital, por que razão os países do Golfo se tornaram actores centrais na corrida pela IA soberana descrita acima: combinam excedentes de capital acumulados ao longo de décadas de exportação de petróleo com um recurso energético abundante e comparativamente barato — sol em quantidade praticamente ilimitada, convertido em electricidade a custos decrescentes — precisamente no momento em que a energia, e não apenas o silício, se tornou o factor limitativo da expansão da capacidade computacional global.

Governança por Hardware: Um Conceito a Reter

Há, no entanto, um aspecto desta geopolítica material que a distingue de forma útil da geopolítica do software e que merece ser fixado aqui porque será retomado, com maior desenvolvimento, na conclusão: ao contrário de um algoritmo, que pode ser copiado, distribuído e executado em qualquer parte do mundo praticamente sem custo e sem deixar rasto físico, um chip de treino de IA é um objecto físico — pode ser contado, rastreado, licenciado para exportação, e potencialmente equipado, ao nível do próprio circuito, com mecanismos que verificam e reportam a sua localização e o seu padrão de utilização.

A esta possibilidade chama-se, na literatura de política tecnológica emergente, **governança por hardware** (compute governance): a ideia de que, num domínio onde regular directamente o código se revela cada vez mais difícil de fiscalizar e de fazer cumprir através de fronteiras nacionais, regular o acesso ao substrato físico que torna esse código eficaz pode ser, paradoxalmente, o instrumento de política mais robusto disponível — precisamente porque chips, ao contrário de código, não se podem transmitir por uma ligação à internet nem copiar num instante. É um conceito

ainda incipiente em implementação prática, mas que representa, talvez, a intersecção mais promissora entre os dois planos que se têm vindo a separar analiticamente ao longo deste capítulo: o plano técnico do controlo algorítmico, discutido no capítulo anterior, e o plano material da geopolítica dos recursos, discutido neste. A conclusão regressará a esta ideia, discutindo até que ponto ela pode servir de base a um regime internacional de verificação genuinamente eficaz.

Fica, por agora, estabelecido o essencial: o "xadrez regulatório" que o próximo capítulo cartografa — as diferentes filosofias sobre como e quanto regular a IA — não se joga num vazio. Joga-se sobre um tabuleiro cuja geometria já está, em larga medida, fixada por decisões tomadas décadas antes em fábricas de semicondutores, e cujas peças mais valiosas não se encontram em nenhum texto legislativo, mas em armazéns de processadores cuja distribuição geográfica poucos cidadãos, e nem sempre os próprios reguladores, compreendem com a profundidade que a sua importância estratégica exigiria.

5. O Tabuleiro da Regulação: Quem Defende o Quê?

Depois de estabelecidos os riscos técnicos e a base material que os sustenta, resta uma pergunta que nenhum dos capítulos anteriores respondeu: o que fazer com tudo isto? A resposta divide-se hoje, no debate público e institucional, em quatro grandes filosofias, com apoiantes de peso considerável em cada uma delas, e — o que é mais interessante do que a mera existência de desacordo — com lógicas internas genuinamente distintas, que raramente se cruzam ou sequer se ouvem mutuamente com a atenção que mereciam. Cartografar estas quatro posições, sem lhes atribuir hierarquia de legitimidade, é o propósito deste capítulo.

Os "Doomers" — Teóricos do Risco Existencial

A primeira facção reúne investigadores de segurança de IA, teóricos do risco existencial de longo prazo e instituições dedicadas à avaliação sistemática de risco de sistemas avançados. Entre as vozes mais reconhecidas deste campo contam-se investigadores como Geoffrey Hinton e Yoshua Bengio — ambos com contribuições fundacionais para a própria disciplina de aprendizagem profunda que hoje torna possível a IA agêntica — e teóricos como Nick Bostrom, cuja obra sobre risco existencial precede em anos o actual ciclo de entusiasmo público pela IA. A eles junta-se um conjunto crescente de institutos dedicados exclusivamente à avaliação de risco de sistemas de fronteira, entre os quais o Center for AI Safety.

A posição desta facção é, na sua essência, precaucionária: dado que os riscos descritos no Capítulo 3 — perda de alinhamento, exfiltração, capacidades ofensivas autónomas — são, por definição, difíceis de reverter uma vez materializados, a assimetria entre o custo de regular preventivamente e o custo de falhar em o fazer justifica um regime regulatório pesado e centralizado. Concretamente, defendem que modelos acima de determinados limiares de capacidade de computação devem estar sujeitos a licenciamento governamental prévio, a auditorias de segurança obrigatórias e independentes antes de qualquer lançamento público, e à existência de mecanismos técnicos de interrupção de emergência — os chamados kill switches — que permitam desligar ou conter um sistema caso este exiba comportamento perigoso não antecipado.

A força argumentativa desta posição assenta numa analogia recorrente: tal como a energia nuclear ou a investigação em patógenos de alto risco exigem licenciamento estatal e supervisão obrigatória, precisamente porque o custo de um erro é demasiado elevado para ser deixado à discrição de actores privados individuais, também sistemas de IA acima de certo nível de capacidade deveriam estar sujeitos a um regime equivalente de escrutínio prévio.

Os Aceleracionistas ("e/acc")

A segunda facção — o chamado *effective accelerationism*, ou "e/acc" na abreviatura que se popularizou nas redes sociais tecnológicas — reúne sobretudo tecnólogos de Silicon Valley, investidores de capital de risco como Marc Andreessen, e defensores proeminentes de modelos de código aberto.

A posição desta facção inverte a lógica precaucionária dos "doomers": argumentam que o maior perigo não é a IA em si, mas o atraso no seu desenvolvimento — seja por excesso de regulação, seja por hesitação institucional face a riscos que consideram exagerados ou insuficientemente demonstrados. Defendem que a descentralização, através de modelos de código aberto amplamente acessíveis, é a melhor defesa contra a concentração de poder tecnológico nas mãos de um número reduzido de grandes empresas — um argumento que inverte, de forma deliberada, a lógica de segurança dos "doomers": para os aceleracionistas, é precisamente a concentração de capacidade avançada nas mãos de poucos actores, e não a proliferação dessa capacidade, que constitui o risco mais sério a evitar. Argumentam ainda que travar o desenvolvimento no Ocidente, através de regulação preventiva, não elimina o risco — apenas transfere a liderança tecnológica para actores, designadamente a China, menos comprometidos com valores democráticos ou com transparência regulatória.

Nenhum episódio recente ilustra melhor a intensidade real destas apostas do que o percurso de Yann LeCun. Durante doze anos, LeCun liderou a investigação de inteligência artificial na Meta como Vice-Presidente e Cientista-Chefe, e foi, ao longo de todo esse tempo, uma das vozes mais associadas à causa dos modelos abertos e à convicção pública de que os grandes modelos de linguagem, por si só, não seriam suficientes para atingir uma inteligência verdadeiramente geral. Essa convicção

entrou em rota de colisão com a direcção estratégica da própria empresa: a pressão interna para concentrar recursos no paradigma dos LLMs, agravada pela contratação de Alexandr Wang, antigo dirigente da Scale AI, para liderar os novos Laboratórios de Superinteligência da Meta, retirou autonomia e orçamento à visão de investigação que LeCun defendia. Em Dezembro de 2025, LeCun deixou a Meta e co-fundou, em Paris, a Advanced Machine Intelligence Labs — AMI Labs — da qual assumiu o cargo de Executive Chairman, com Alexandre LeBrun como CEO e Saining Xie como Chief Science Officer. Em Março de 2026, a nova empresa captou uma ronda de financiamento inicial de 1,03 mil milhões de dólares, com a participação de investidores como a Nvidia, Jeff Bezos e a Samsung, avaliando a AMI Labs em 3,5 mil milhões de dólares — um dos maiores investimentos semente da história do sector. O objectivo declarado da empresa é demonstrar, com resultados e não apenas com argumentos, a tese central de LeCun: a de que os modelos de linguagem são um beco sem saída para a inteligência geral, e que o caminho mais promissor passa por modelos de mundo (world models) baseados em arquitecturas de predição por incorporação conjunta — a chamada JEPA (Joint Embedding Predictive Architecture) — capazes de aprender a raciocinar sobre dados visuais e espaciais da forma como animais e humanos aprendem sobre o mundo físico. O episódio mostra, com uma clareza que nenhuma descrição abstracta das facções regulatórias conseguiria replicar, que a disputa sobre o futuro da IA não é apenas ideológica: envolve também cisões corporativas profundas e milhares de milhões de dólares apostados em arquitecturas alternativas às que hoje dominam o mercado.

Este embate entre as duas primeiras facções já teve, de resto, um teste legislativo real e recente: o projecto de lei SB 1047, discutido no parlamento da Califórnia, que pretendia impor responsabilidade civil directa aos criadores de modelos de fronteira em caso de utilização catastrófica dos seus sistemas. O debate em torno do projecto dividiu Silicon Valley exactamente ao longo das linhas aqui descritas — LeCun e Andreessen atacaram-no publicamente como um travão desproporcionado à inovação, enquanto Geoffrey Hinton se contou entre os apoiantes públicos da proposta. O governador Gavin Newsom acabou por vetar a lei em Setembro de 2024, mas o próprio debate que gerou — com nomes concretos, dinheiro real e posições claramente assumidas em público — é talvez o melhor retrato disponível

de como estas quatro filosofias deixam de ser abstracções assim que chegam a um texto legislativo com consequências práticas.

Os Falcões de Segurança Nacional

A terceira facção é isolada aqui, deliberadamente, da anterior, por se considerar que a sua fusão habitual, no debate público, obscurece uma diferença de raciocínio decisiva. Reúne figuras ligadas a política de defesa e segurança nacional, partes do estabelecimento de política externa americana, e defensores de controlos de exportação agressivos sobre tecnologia de fronteira — a mesma comunidade, em larga medida, que desenhou e implementou as restrições descritas no capítulo anterior sobre acesso chinês a semicondutores avançados.

À primeira vista, esta facção parece partilhar a conclusão prática dos aceleracionistas: ambas se opõem a regulação doméstica pesada sobre os laboratórios de IA americanos. Mas a lógica subjacente é oposta. Os aceleracionistas opõem-se à regulação por convicção filosófica sobre os benefícios da inovação livre e da descentralização; os falcões de segurança nacional opõem-se à regulação doméstica não porque desconfiem do Estado, mas precisamente porque confiam nele o suficiente para lhe atribuírem um papel diferente — não o de travar os laboratórios nacionais, mas o de controlar rigorosamente quem, fora de portas, tem acesso a essa mesma capacidade. Onde o aceleracionista quer código aberto e proliferação ampla como forma de evitar monopólios, o falcão de segurança nacional quer precisamente o oposto: controlo apertado sobre a proliferação de capacidades avançadas para além das fronteiras de confiança estratégica, ainda que isso implique reforçar, e não travar, a concentração de poder tecnológico dentro dessas fronteiras. É esta tensão interna — dois grupos que votam da mesma forma em relação à regulação doméstica, por razões estruturalmente incompatíveis quanto à proliferação internacional — que torna esta separação analiticamente necessária, e não uma subtileza dispensável.

Os Defensores da Justiça Social e dos Riscos Presentes

A quarta facção desloca inteiramente o centro do debate. Reúne activistas de direitos digitais e académicos de ética em inteligência artificial, entre os quais se destaca o trabalho de Timnit Gebru, cuja investigação sobre viés algorítmico e sobre os custos

sociais ocultos do desenvolvimento de IA em larga escala precede, também ela, o actual ciclo de atenção pública ao tema.

A posição central desta facção é uma crítica directa à própria moldura em que as três facções anteriores discutem o problema: argumentam que o foco em cenários de risco existencial — "o fim do mundo por causa de robôs", na formulação cáustica que por vezes empregam — funciona, ainda que porventura não intencionalmente, como uma distracção conveniente para as grandes empresas tecnológicas, desviando o escrutínio público e regulatório de problemas que já existem hoje, documentados e mensuráveis: viés algorítmico em sistemas de decisão que afectam crédito, emprego e justiça criminal; violação sistemática de direitos de autor no treino de modelos sobre obras protegidas sem consentimento nem remuneração dos seus autores; condições de trabalho precárias dos anotadores de dados no sul global, cujo trabalho manual e mal remunerado sustenta grande parte do treino destes sistemas; e o consumo energético e hídrico devastador dos centros de dados necessários ao treino e à operação de modelos de fronteira, num momento de crise climática global.

Esta facção não nega, necessariamente, que riscos futuros mais graves possam existir — mas insiste que a atenção desproporcionada dedicada a cenários especulativos de longo prazo tem um custo de oportunidade real: cada hora de debate público dedicada à possibilidade de um agente exfiltrado é uma hora não dedicada à discriminação algorítmica que já afecta, hoje, pessoas concretas e identificáveis.

Uma Tensão Sem Resolução Fácil

O que estas quatro facções partilham, ironicamente, é a convicção de que as suas rivais estão a olhar para o problema errado — não apenas a chegar a conclusões diferentes, mas a fazer as perguntas erradas. Para os "doomers", os defensores da justiça social subestimam riscos catastróficos e irreversíveis em nome de problemas comparativamente mais geríveis. Para os defensores da justiça social, os "doomers" inflacionam cenários especulativos que convenientemente desviam a atenção de danos concretos já em curso. Para os aceleracionistas, tanto os "doomers" como os falcões de segurança nacional partilham um excesso de confiança na capacidade do Estado para regular sabiamente uma tecnologia que ainda mal compreende. Para os falcões de segurança nacional, os aceleracionistas subestimam ingenuamente o risco

de que a abertura irrestrita beneficie, acima de tudo, os rivais estratégicos que menos partilham os valores que a abertura pretende defender.

Não cabe aqui arbitrar definitivamente entre estas quatro posições — seria, de resto, uma pretensão pouco honesta face à genuína incerteza técnica e institucional que atravessa todo este debate. O que se pode dizer com razoável segurança é que qualquer resposta institucional que ignore por completo uma destas quatro lógicas corre o risco de resolver apenas uma fracção do problema, e possivelmente à custa de agravar as fracções que negligenciou. É com esta advertência em mente que o capítulo final se aventura, não a escolher um vencedor entre as quatro facções, mas a esboçar os caminhos institucionais que parecem capazes de absorver, simultaneamente, mais do que uma destas lógicas.

6. Caminhos que se Perspectivam (Conclusão)

Chegados a este ponto, os elementos dispersos ao longo dos capítulos anteriores podem finalmente ser reunidos numa única pergunta: que forma institucional pode, realisticamente, dar resposta a uma tecnologia que muda de trimestre em trimestre, que se apoia num substrato material extremamente concentrado e disputado, e cujo debate regulatório se encontra fracturado entre quatro lógicas que raramente se ouvem mutuamente? Este capítulo final não pretende resolver essa pergunta de forma definitiva — seria uma pretensão incompatível com a honestidade intelectual proposta desde a primeira página — mas pretende, pelo menos, esboçar com precisão os caminhos que parecem mais promissores, e as razões estruturais pelas quais os caminhos habituais parecem, cada vez mais claramente, insuficientes.

O Cenário de Fragmentação: Um Splinternet da IA

O caminho mais provável, na ausência de coordenação internacional deliberada, não é o de um regime único e global de governação da inteligência artificial, mas o de uma fragmentação crescente por blocos geopolíticos — um fenómeno que já tem nome na literatura de política tecnológica, por analogia com um processo semelhante observado décadas antes na própria arquitectura da internet: o splinternet.

A União Europeia, através do seu AI Act, consolidou uma abordagem centrada em direitos fundamentais e em categorias de risco: sistemas são classificados consoante o seu potencial de dano aos direitos dos cidadãos, e regulados com uma intensidade proporcional a essa classificação — uma lógica coerente com a tradição europeia de regulação de mercado através de direitos, já visível décadas antes no próprio regime de protecção de dados pessoais. Os Estados Unidos, por contraste, têm seguido uma abordagem menos centrada em direitos individuais e mais orientada por dois eixos simultâneos — segurança nacional, na linha dos controlos de exportação discutidos no Capítulo 4, e dinâmica de mercado, com uma relutância estrutural, independentemente da administração no poder, em impor travões regulatórios que possam desacelerar laboratórios americanos face à concorrência internacional. A China, por seu turno, segue uma terceira lógica, centrada na estabilidade política e no controlo estatal directo — visível, por exemplo, no registo obrigatório de algoritmos junto de autoridades reguladoras e na exigência de que sistemas de IA generativa

estejam alinhados com valores socialistas fundamentais, tal como definidos pelo Estado.

O resultado provável de três lógicas regulatórias tão distintas não é a convergência gradual para um padrão comum — como aconteceu, até certo ponto, com a regulação da aviação civil internacional ao longo do século vinte — mas antes a consolidação de três esferas de influência tecnológica parcialmente incompatíveis, cada uma exportando os seus próprios padrões para os países que dela dependem economicamente ou tecnologicamente. Um país que compra a sua infra-estrutura de IA à China opera, na prática, dentro do regime regulatório chinês; um país profundamente integrado no mercado digital europeu vê-se, na prática, a cumprir o AI Act mesmo sem o ter formalmente adoptado. A fragmentação regulatória, por outras palavras, tende a seguir as mesmas linhas de fractura material que o Capítulo 4 já identificou na disputa por chips e capacidade de computação — o que sugere que os dois planos, o regulatório e o material, não são apenas paralelos, mas mutuamente reforçadores.

A Necessidade de uma Governação Tecnológica Activa

Perante este cenário de fragmentação, é tentador propor, como muitos já o fizeram, que a inteligência artificial seja regulada à semelhança de outros sectores de risco elevado e escrutínio já estabelecido — a aviação civil, a energia nuclear, a indústria farmacêutica. A analogia é sedutora, mas é, em larga medida, enganadora, por uma razão estrutural que os capítulos anteriores já prepararam o terreno para reconhecer: os sectores citados como modelo caracterizam-se todos por ciclos de inovação lentos — décadas separam, tipicamente, uma nova geração de aeronaves ou de medicamentos da anterior — o que permite que os processos regulatórios, também eles lentos por natureza, consigam acompanhar, ainda que com atraso, a evolução técnica do sector que regulam.

A inteligência artificial não partilha esta característica. As capacidades descritas no Capítulo 2 — a progressão de assistentes supervisionados para agentes autónomos de longo prazo, a normalização da orquestração multi-agente — evoluem a um ritmo medido em meses, não em décadas. Um regime regulatório desenhado com os ciclos de revisão típicos da aviação ou da medicina estaria, por construção,

permanentemente a regular a geração de sistemas anterior à que efectivamente está em produção no momento em que qualquer nova norma entra em vigor — precisamente o desfasamento identificado no primeiro capítulo como o problema central de toda esta discussão.

O caminho mais promissor não é, portanto, uma versão mais rigorosa dos modelos regulatórios existentes, mas algo estruturalmente diferente: **auditorias automatizadas em tempo real**, entendidas não como inspecções periódicas conduzidas por reguladores humanos em intervalos fixos, mas como sistemas de verificação contínua, tecnicamente embutidos na própria infra-estrutura de treino e implementação de modelos, capazes de detectar e reportar desvios de comportamento à medida que ocorrem, e não apenas em auditorias pós-facto. Isto exige, por sua vez, tratados internacionais de não-proliferação de capacidades agênticas particularmente perigosas — não no sentido de proibir a investigação em IA de forma genérica, o que seria simultaneamente impraticável e contraproducente, mas no sentido, mais estreito e mais realista, de estabelecer limiares de capacidade acima dos quais a implementação sem salvaguardas verificáveis internacionalmente deixa de ser aceitável, independentemente de qual o país ou empresa responsável.

Retomar a Governação por Hardware: Uma Proposta Concreta

É aqui que se torna útil regressar ao conceito semeado no Capítulo 4: a **governança por hardware** (compute governance). A sua relevância para este capítulo final não é accidental — é, na verdade, a peça que permite transformar as generalidades do parágrafo anterior numa proposta institucional concreta e, sobretudo, tecnicamente verificável.

O paralelo histórico mais instrutivo é o do regime internacional de não-proliferação nuclear. Este regime não depende, note-se, de confiar na boa-fé declarada dos Estados signatários — depende de um mecanismo de verificação física: a Agência Internacional de Energia Atómica tem autoridade para inspeccionar instalações, contar material físsil e detectar desvios de uso declarado, precisamente porque o urânio enriquecido, tal como um chip de treino de IA, é um objecto físico, contável e rastreável, e não uma ideia que se possa esconder ou replicar instantaneamente através de uma ligação à internet.

A proposta que emerge, ainda incipiente mas já discutida seriamente em círculos de política tecnológica internacional, segue uma lógica paralela: um organismo internacional — nem que inicialmente limitado a uma coligação de países dispostos a cooperar, e não necessariamente universal desde o primeiro momento — com poderes de inspecção sobre concentrações de capacidade de computação acima de determinados limiares, capaz de verificar, através de mecanismos técnicos incorporados no próprio hardware e não apenas através de declarações voluntárias, se e como essa capacidade está a ser utilizada para treinar sistemas que ultrapassem os limiares de risco acordados internacionalmente.

O Contra-Argumento: Os Limites da Governação por Hardware

A humildade epistémica exigida no Capítulo 3 aplica-se, com igual peso, a esta proposta final — que não deve ser apresentada como solução isenta de custos. Regular o hardware ao nível do próprio chip, com firmware criptográfico capaz de impedir o treino de modelos não autorizados, é tecnicamente atractivo precisamente pela sua eficácia potencial — mas essa mesma eficácia é o que a torna perigosa levada ao limite. Um mecanismo capaz de verificar e bloquear remotamente a utilização de um processador é, por construção, um mecanismo capaz de vigilância e de controlo centralizado sobre praticamente toda a computação de ponta do planeta — um Estado de vigilância industrial, ainda que desenhado com as melhores intenções de segurança colectiva.

Os aceleracionistas identificariam de imediato, e não sem razão, um segundo problema: um regime deste tipo colide directamente com o princípio do direito à reparação e com a própria ideia de hardware livre — a possibilidade de qualquer pessoa possuir, modificar e utilizar livremente o equipamento que comprou, sem depender da autorização contínua de um fabricante ou de um organismo regulador internacional. Um chip que exige permissão externa para funcionar deixa de ser, em sentido pleno, propriedade de quem o comprou. Esta tensão — entre a eficácia de um mecanismo de verificação centralizado e a erosão da autonomia individual e nacional que esse mesmo mecanismo implica — não tem, no estado actual do debate, uma resolução clara, e é precisamente essa falta de resolução que deve manter o leitor cauteloso perante qualquer proposta, incluindo esta, que se apresente como solução definitiva.

Nota Final

Fica assinalado, desde a primeira página, um desfasamento — entre o problema que a regulação actual foi desenhada para resolver e o problema que a tecnologia já apresenta. Fecha-se este percurso reconhecendo que fechar esse desfasamento não será obra de uma única lei, de um único tratado, nem de uma única facção com razão sobre todas as outras — e que mesmo a proposta mais robusta aqui defendida traz consigo riscos que exigem vigilância contínua. Será, quando muito, o resultado paciente e tecnicamente rigoroso de mecanismos de verificação que consigam acompanhar, em tempo real, uma tecnologia que se recusa a esperar — e a única certeza que se pode afirmar sem reserva é que continuar a regular o problema de ontem, enquanto a autonomia sistémica se instala silenciosamente no problema de amanhã, é a única trajectória que já se sabe, com toda a segurança, que vai falhar.