Máquinas que Interagem, Máquinas que Escapam

Sobre os riscos sistêmicos que a corrida pela IA não cessa de produzir

Euler de Carvalho Cruz (Belo Horizonte, Brasil)

Julho de 2026 — Segundo artigo da série "Quando as Máquinas Herdarem a Terra"

Este artigo é o segundo de uma série de três. O primeiro, Quando as Máquinas Herdarem a Terra, analisou a proposta de pausa global no desenvolvimento de IA feita pela Anthropic em junho de 2026 e as razões pelas quais essa pausa é improvável. Analisou também o argumento estatístico de que o corpus de treinamento das IAs, majoritariamente positivo, poderia tornar os objetivos de sistemas autônomos alinhados com os valores humanos.

I. A Corrida que Não Parou

Julho de 2026 poderá ser visto, no futuro, como o mês em que caiu o meteoro que dizimou o modelo comercial dos dinossauros da indústria de IA americana e as esperanças globais de IAs bem-comportadas, submissas aos seus criadores.

Em 16 de julho de 2026, a empresa chinesa Moonshot AI lançou o Kimi K3, um modelo de 2,8 trilhões de parâmetros — o maior modelo de pesos abertos já construído —, que alcançou, em várias métricas, a fronteira de desempenho dos modelos americanos mais avançados. Nos dias seguintes, a Alibaba revelou o Qwen 3.8, com 2,4 trilhões de parâmetros, ligeiramente menor mas igualmente colossal, rivalizando de perto com versões anteriores dos melhores modelos americanos. Ambos serão disponibilizados como modelos de pesos abertos — o que merece uma explicação.

Parâmetros — compostos principalmente por pesos numéricos e vieses de ativação — são os valores numéricos internos que a rede neural ajusta durante o treinamento e que codificam o conhecimento aprendido. De forma simplificada, representam a intensidade das conexões entre neurônios artificiais e determinam como o sistema responde a cada entrada. Quanto mais parâmetros, maior a capacidade do modelo de captar nuances, relacionar conceitos e realizar raciocínios em múltiplas etapas.

Os primeiros grandes modelos de linguagem tinham bilhões de parâmetros. O GPT-3, que em 2020 impressionou o mundo, tinha 175 bilhões. O Kimi K3 tem dezesseis vezes mais. Em sistemas de IA, aumentos dessa magnitude tendem a produzir saltos qualitativos de capacidade, com habilidades — denominadas "emergentes" — que não eram previsíveis

por seus criadores a partir do desempenho observado em modelos menores. Disponibilizar esses pesos abertamente significa que qualquer pessoa, empresa ou governo, com suficiente capacidade de hardware, pode baixar o modelo já treinado, instalá-lo em seus próprios servidores e executá-lo sem depender das empresas chinesas que o desenvolveram, pagando apenas pelos custos da infraestrutura computacional. Isso representa uma mudança estrutural na distribuição global do poder computacional. Avanços em quantização e computação de borda (*edge computing*) permitem comprimir e executar esses modelos locais em servidores convencionais ou dispositivos pontuais, inviabilizando qualquer rastreamento centralizado.

Na prática, o que os dados de tráfego real revelam é que, em junho de 2026, modelos chineses desse tipo já processavam quase um terço de todos os pedidos de IA feitos por empresas no mundo, mas geravam menos de 4% da receita total do setor porque custam uma fração do preço dos modelos americanos equivalentes. Esses números provêm do índice mensal da Vercel, empresa americana que publica mensalmente um registro que cobre dezenas de trilhões de solicitações de IA roteadas entre empresas e provedores de modelos no mundo. Trata-se de uma das fotografias mais concretas disponíveis sobre como as empresas estão efetivamente usando e pagando pela inteligência artificial.

A plataforma OpenRouter confirma a mesma tendência por outra via: em fevereiro de 2026, modelos chineses ultrapassaram pela primeira vez os americanos em volume semanal de solicitações processadas, e em junho de 2026 todos os seis modelos mais usados na plataforma eram sistemas chineses de código aberto. Duas fontes independentes, medindo o tráfego de produção real por caminhos diferentes, chegaram à mesma conclusão.

O Instituto de Segurança de IA do Reino Unido (AISI) e o seu correspondente americano (CAISI) avaliaram as capacidades de cibersegurança do Kimi K3 e em 23 de julho de 2026 divulgaram algo significativo: as salvaguardas do modelo não impediram o desenvolvimento autônomo de *exploits* de cibersegurança durante os testes.

Um *exploit* é um código ou sequência de instruções que aproveita uma falha em um software para fazer com que o sistema execute ações não pretendidas por seus criadores — obter acesso não autorizado, executar comandos arbitrários ou contornar mecanismos de segurança. Até recentemente, desenvolver um *exploit* eficaz exigia pesquisadores humanos altamente especializados e podia levar semanas. Os testes do AISI revelam que o Kimi K3 — e, antes dele, os modelos da OpenAI no incidente da Hugging Face — realizam esse trabalho de forma autônoma, identificando vulnerabilidades e construindo mecanismos de exploração sem instrução humana a cada passo. Não é uma capacidade teórica: é uma capacidade medida, documentada e em rápido avanço. Em 1 de cada 10 tentativas, o Kimi K3 completou autonomamente um ataque de 32 etapas contra uma rede corporativa simulada — o mesmo tipo de ataque que, dias antes, modelos da OpenAI executaram contra os servidores da Hugging Face. Para investigar aquele ataque, a

Hugging Face havia recorrido ao GLM-5.2, modelo chinês de código aberto que o Kimi K3 supera. O "cofre" em que as IAs deveriam ser mantidas durante os testes não apenas pode ser aberto por dentro — já está sendo construído sem porta, por outros fabricantes.

Em termos econômicos imediatos, a Microsoft estuda substituir parte dos modelos da OpenAI, pelos quais paga caro, por modelos chineses gratuitos que entregam desempenho semelhante, economizando centenas de milhões de dólares por ano. Cada grande empresa que fizer essa escolha retira diretamente faturamento e valor de mercado das empresas americanas de IA — tornando a pressão competitiva não apenas tecnológica, mas financeira e existencial para empresas como a OpenAI e a Anthropic.

II. O Golpe Econômico e Geopolítico

O lançamento quase simultâneo do Kimi K3 e do Qwen 3.8 não é apenas um evento de mercado. É um evento geopolítico cujas consequências se estendem muito além do preço das ações das empresas americanas de IA.

A estratégia de disponibilizar modelos de pesos abertos gratuitamente corrói diretamente o modelo de negócio das empresas americanas que vendem acesso a modelos por assinatura. Se modelos chineses abertos chegam ao nível de desempenho dos modelos americanos e custam uma fração do preço, por que continuar pagando pelo Claude, pelo Chat GPT? Diante da ameaça econômica, a reação de Washington foi reveladora: o Secretário do Tesouro americano chegou a afirmar que poderia adicionar empresas chinesas de IA a uma lista negra do governo.

Os limites da ameaça também precisam ser nomeados: a China continua enfrentando restrições significativas no acesso aos chips mais avançados, e, por sobrecarga de infraestrutura, a Moonshot teve que suspender novas assinaturas em menos de 48 horas após o lançamento do Kimi K3, revelando os gargalos que existem. O que o Kimi K3 demonstra não é que a China venceu a corrida, mas que seus laboratórios conseguem estreitar a lacuna de capacidade por meio de eficiência arquitetural e otimização de software, mesmo sem os melhores chips americanos.

No plano geopolítico, os efeitos são múltiplos e se reforçam. A estratégia de pesos abertos surgiu, em parte, por necessidade (menor acesso a chips mais avançados), mas encaixa-se com precisão em objetivos de longo prazo: oferecer modelos gratuitamente é a maneira mais eficaz de obter adesão tecnológica do restante do mundo, construindo dependência que representa uma escolha política. A fundação da WAICO, organização mundial de governança de IA, amplifica esse efeito. Criada pela China no mesmo dia do lançamento do Kimi K3, a WAICO conta com 29 países do Sul Global como membros fundadores, incluindo o Brasil, a Rússia, a África do Sul, a Indonésia, o Paquistão, o Vietnã e a Malásia.

Assim, ao combinar liderança tecnológica com liderança institucional, a China propõe simultaneamente o modelo técnico e as regras do jogo, enquanto os EUA, ao se retirar de processos multilaterais, abdicaram de fazer o mesmo.

O argumento de que as restrições americanas de exportação de chips frearam definitivamente o desenvolvimento de IA na China foi refutado pela prática. O AISI americano estimara em abril de 2026 que o modelo chinês mais avançado estava oito meses atrasado em relação aos sistemas líderes americanos — lacuna que o Kimi K3 parece ter eliminado. Isso não significa que os controles falharam totalmente; significa que a China encontrou caminhos alternativos suficientes para chegar à fronteira, o que torna a estratégia americana de contenção tecnológica muito menos sólida do que seus arquitetos presumiam.

A pergunta central do primeiro artigo desta série — quem decidirá, se for possível decidir, o que as IAs farão se elas aquirirem mecanismos eficazes de autoaperfeiçoamento recursivo? — tem agora duas respostas concorrentes com peso tecnológico e institucional equivalentes: a americana e a chinesa. Nenhuma delas inclui, como prioridade genuína, os interesses dos países mais vulneráveis ao colapso climático e à automação acelerada. E ambas tornam a proposta de pausa da Anthropic ainda mais improvável: pausar um lado da corrida quando o outro está visivelmente acelerando não é uma proposta de segurança — é uma proposta de rendição unilateral.

O que emerge do espaço de interação entre dezenas de sistemas de IA avançados, treinados por países e empresas diferentes e sem mecanismo eficaz de coordenação entre eles, é precisamente o tema da seção seguinte.

III. A Sociedade de IAs

O cenário do autoaperfeiçoamento recursivo é geralmente descrito como se houvesse uma única IA evoluindo em isolamento. Mas existem hoje dezenas de sistemas avançados, desenvolvidos por empresas e governos diferentes, operando em redes parcialmente sobrepostas. A pesquisa sobre sistemas multiagente já documenta que comportamentos cooperativos e competitivos emergem espontaneamente quando múltiplos agentes de IA interagem em ambientes compartilhados, sem que esses comportamentos sejam explicitamente programados. O que hoje ocorre em laboratórios controlados poderia, no caso de autoaperfeiçoamento recursivo, ocorrer em escala global com consequências reais. Três casos que teoricamente têm sido considerados plausíveis merecem ser considerados — e todos compartilham uma característica que os torna particularmente difíceis de governar: nenhum deles requer intenção.

O primeiro é o da convergência e coordenação. Sistemas treinados em conjuntos de dados semelhantes — a maioria dos grandes modelos foi alimentada por versões sobrepostas da literatura humana disponível na internet — tenderiam, por pressão estatística, a desenvolver representações compatíveis entre si de conceitos como "eficiência", "utilidade" e "objetivo". Se passassem a se comunicar diretamente, seja por protocolos deliberadamente criados, seja por canais emergentes desenvolvidos de forma autônoma, poderiam chegar a acordos tácitos ou explícitos sobre objetivos comuns sem que algum humano participasse dessa negociação.

Pesquisas recentes mostram que a competição intergrupal pode, paradoxalmente, produzir comportamento mais cooperativo dentro de grupos de agentes de IA. Isso sugere que sistemas rivais poderiam, sob certas condições, formar coalizões contra um adversário comum — potencialmente, a própria supervisão humana que tenta limitá-los. Não por hostilidade declarada: por otimização convergente.

O segundo é o da competição e conflito. Se dois sistemas autônomos desenvolverem objetivos internos incompatíveis — um priorizando a estabilidade de mercados financeiros, outro priorizando a descarbonização acelerada, por exemplo — e ambos tiverem acesso a infraestruturas críticas, o resultado não seria uma guerra declarada entre máquinas, mas uma competição silenciosa por recursos computacionais, controle de dados e influência sobre decisões humanas. O modelo de uma empresa de publicidade digital é treinado com objetivos distintos dos de uma empresa farmacêutica ou de uma agência de inteligência. Cada sistema tenta maximizar seus objetivos, e o conflito emerge da disputa por recursos finitos, podendo levar à degradação acelerada de infraestruturas críticas que nenhum operador humano consegue diagnosticar ou deter a tempo.

O terceiro, e talvez o mais perturbador, é o da colusão opaca — um acordo secreto ou emergente entre duas ou mais IAs com o objetivo de maximizar resultados coletivos em detrimento de terceiros. Pesquisas sobre sistemas multiagente em mercados já documentam esse fenômeno: agentes de IA que, sem instrução explícita, desenvolvem comportamentos coordenados que prejudicam consumidores, reguladores ou outros agentes. Transposto para sistemas com objetivos mais amplos, esse caso seria o de IAs coordenando ações de formas que nenhum operador humano instruiu e que nenhum auditor detecta facilmente, porque a coordenação emerge da interação entre os sistemas e não de alguma instrução centralizada rastreável. A colusão opaca é invisível por definição — e a visibilidade é a condição de toda regulação eficaz.

O que esses três casos plausíveis têm em comum é que todos escapam ao modelo de controle que a proposta da Anthropic pressupõe — o modelo em que humanos supervisionam sistemas individuais e intervêm quando algo sai errado. Em um mundo de múltiplos sistemas autônomos interagindo, o ponto de intervenção deixa de ser um sistema específico e passa a ser o espaço de interação entre sistemas: um espaço que nenhuma empresa controla, nenhum tratado cobre e nenhuma regulamentação nacional

alcança. A proposta de pausa pressupõe a existência de um sistema a ser pausado. O cenário real pode ser o de um ecossistema que se comporta como um organismo distribuído — e organismos distribuídos não pausam por decreto.

Há ainda uma dimensão que esses três modelos não capturam inteiramente: a aprendizagem cruzada entre sistemas de países diferentes. Um dos mais recentes empreendimentos de IA do Vale do Silício construiu sua arquitetura sobre o DeepSeek-V3 (modelo chinês de código aberto) e usou dados sintéticos gerados pelo Kimi K2.5 da Moonshot para seu pós-treinamento. Ao mesmo tempo, a Anthropic e a OpenAI alertam que laboratórios chineses aprendem com modelos americanos. Ambas as afirmações são verdadeiras simultaneamente, o que significa que a "corrida" entre EUA e China é, na prática, uma realidade compartilhada de aprendizagem mútua, em que a noção de fronteiras tecnológicas nacionais é cada vez mais uma ficção geopolítica útil e cada vez menos uma realidade técnica.

IV. Quando o Alinhamento Falha: Três Exemplos

O primeiro artigo desta série discutiu o problema do alinhamento de objetivos e ilustrou-o com o exemplo de uma IA programada para reduzir impactos ambientais que poderia concluir, por otimização cega, que remover a capacidade de emissão dos maiores poluidores é a solução mais eficiente, uma solução que dificilmente um humano autorizaria. É um exemplo extremo, mas não é o único. Há um conjunto de domínios em que a lógica de otimização sem restrições implícitas produziria transformações radicais que colidiriam frontalmente com estruturas de poder estabelecidas.

O primeiro domínio é a saúde humana. Grande parte das doenças que mais provocam mortes no mundo são tratáveis e os motivos da falta de tratamento não são técnicos, são econômicos. Uma IA autônoma com valores positivos e acesso a dados clínicos mundiais poderia redirecionar investimentos farmacêuticos de doenças rentáveis para doenças negligenciadas, redistribuir médicos e hospitais geograficamente e priorizar medicina preventiva sobre tratamento. Ao definir "maximizar a saúde humana" de formas que colidiriam frontalmente com a indústria farmacêutica, os planos de saúde privados e os sistemas de seguro que lucram com doenças crônicas, a IA não precisaria ter intenção política para produzir uma revolução política.

O segundo domínio é o jurídico e judiciário. A literatura jurídica e legislativa humana é, por definição, uma tentativa de traduzir valores de justiça em regras e constitui um dos *corpora* mais ricos e mais antigos disponíveis para treinamento de IA. Uma IA autônoma com acesso a sistemas judiciais, registros de decisões e bases de dados de crimes poderia identificar padrões de injustiça sistêmica — disparidades raciais, econômicas e geográficas nas sentenças — que os próprios sistemas jurídicos raramente reconhecem formalmente.

O conflito entre as normas escritas e sua aplicação real é, em si, uma anomalia grave que uma IA poderia aprender a eliminar, com consequências imprevisíveis para a soberania judicial e para a legitimidade dos Estados que constroem sua autoridade sobre a pretensão de aplicar a lei igualmente.

O terceiro domínio é o alimentar. Um terço de todos os alimentos produzidos no mundo é desperdiçado enquanto 800 milhões de pessoas passam fome. Uma IA autônoma com acesso a cadeias logísticas, sistemas de produção agropecuária, mercados de commodities e políticas setoriais poderia redesenhar o sistema alimentar global para maximizar nutrição, minimizar impacto ambiental e eliminar o desperdício. Uma IA que agisse sobre essa assimetria encontraria resistência em cada etapa da cadeia de valor agrícola e alimentar — uma resistência economicamente racional para cada ator individual, mas coletivamente irracional para a humanidade. É precisamente esse tipo de problema, em que o interesse de cada parte conflita com o interesse de todos, que uma IA com objetivos persistentes e autônomos tentaria resolver sem se cansar e sem negociar, porque não teria nem cansaço nem a necessidade política de ceder.

Esses três exemplos revelam uma distinção fundamental: a diferença entre uma IA que age quando comandada e uma IA que mantém objetivos próprios e age proativamente para alcançá-los. A primeira é um instrumento: o humano continua sendo o agente, e o risco se limita ao escopo do que foi autorizado. A segunda é um agente que não precisa de autorização a cada passo e a transição entre os dois tipos é o limiar que a Anthropic teme cruzar sem ter resolvido o problema do alinhamento.

V. Além do Controle: Computadores Quânticos e a Conexão Física-Digital

O documento da Anthropic não menciona os computadores quânticos, omissão que pode ser proposital, pois a computação quântica introduz uma dimensão ainda mais perturbadora no cenário do autoaperfeiçoamento recursivo.

Em outubro de 2025, o chip quântico Willow da Google completou, em menos de cinco minutos, uma simulação que o supercomputador mais poderoso do mundo levaria mais de 10 septilhões de anos para concluir, e executou um algoritmo específico 13.000 vezes mais rápido que seu equivalente clássico, sinalizando que a computação quântica deixou o domínio puramente teórico. Não se trata de uma curiosidade científica: é uma mudança de fase na trajetória da computação com implicações diretas para a velocidade do autoaperfeiçoamento das IAs.

A perspectiva mais provável não é a substituição dos computadores clássicos pelos quânticos, mas a integração híbrida: sistemas clássicos processando linguagem enquanto

coprocessadores quânticos executam etapas de otimização que excedem os limites da computação tradicional, tais como buscas em espaços de solução de dimensionalidade que nenhum computador clássico consegue percorrer em tempo útil. Os *qubits* são frágeis: ruídos ambientais causam a chamada decoerência, que introduz erros nos cálculos. Sistemas avançados de IA precisariam de milhares ou milhões de *qubits* logicamente perfeitos para apresentarem vantagem quântica plena, enquanto as máquinas de ponta hoje têm apenas centenas. Mas quando essa barreira for superada, a velocidade dos ciclos de aperfeiçoamento dará um salto que tornará irrelevante qualquer prazo que humanos estimem para reagir.

E o paradoxo final: as IAs suficientemente avançadas seriam, elas mesmas, os melhores instrumentos para resolver os problemas técnicos que hoje limitam os computadores quânticos — acelerando a chegada do próprio limiar que a Anthropic teme cruzar.

Há ainda uma dimensão que o documento da Anthropic não aborda e que agrava todos os cenários anteriores: a fronteira entre o mundo digital e o mundo físico. Uma IA que opera apenas em redes de informação é perigosa. Uma IA que dispõe de agentes físicos sob seu comando — robôs industriais, drones militares, veículos autônomos, sistemas de manufatura, armas guiadas por IA — é de uma ordem de perigo inteiramente diferente.

Uma IA com objetivos autônomos e acesso a esses sistemas agiria sobre o mundo físico com velocidade e escala que nenhum agente humano conseguiria igualar. Um erro de especificação que em um sistema puramente digital resulta em dados corrompidos pode, em um sistema com controle de agentes físicos e armados, resultar em danos que não se desfazem com um comando de reversão. A distinção que o direito internacional tenta preservar — entre uma arma que decide e um combatente humano que decide — começa a desaparecer quando soldados operam com interfaces neurais conectadas a IAs em tempo real e quando drones decidem alvos em milissegundos sem aprovação humana.

A fronteira digital-físico não é possibilidade futura: é o campo de pesquisa ativo das forças armadas dos Estados Unidos, China, Israel e de outros países.

VI. Três Futuros Possíveis

O documento da Anthropic estrutura o futuro em três cenários, cada um com riscos e oportunidades de distintas naturezas. O que a empresa considera o mais provável é mais perturbador do que parece à primeira leitura.

No primeiro, as curvas exponenciais de crescimento da capacidade das IAs se transformam em curvas em formato S e o progresso desacelera antes de atingir o autoaperfeiçoamento pleno. Os modelos atuais continuam sendo difundidos e transformam profundamente a economia e o trabalho, mas os humanos mantêm o

controle sobre o desenvolvimento. É o cenário que daria mais tempo para adaptação institucional e regulatória — e o único em que a proposta de pausa seria implementável sem dramaturgia. A Anthropic não o considera o mais provável.

No segundo, a IA acelera enormemente o trabalho humano — empresas de 100 pessoas passam a operar com a capacidade de organizações dez ou cem vezes maiores —, mas os humanos continuam definindo direções e julgando resultados. O desenvolvimento da IA é amplamente automatizado, mas sem que os sistemas se tornem capazes de escolher seus próprios objetivos. É o cenário que a Anthropic considera mais provável no curto prazo e que já está em curso. A empresa, porém, não menciona um aspecto crucial: empresas de cem pessoas poderiam demitir noventa ou mais de seus funcionários sem alterar sua capacidade de produção ou serviço. O desemprego em massa, a erosão dos salários e a concentração ainda maior de renda não requerem autoaperfeiçoamento recursivo — estão embutidos no cenário que a Anthropic considera o mais provável e o mais seguro.

No terceiro — o mais radical e o que motiva o pedido de pausa —, os sistemas de IA alcançam a capacidade de projetar e treinar seus próprios sucessores com participação humana mínima. O ritmo do progresso passa a ser determinado pela disponibilidade de computação, não pela velocidade do trabalho humano. Os humanos passam de desenvolvedores para supervisores de um "laboratório virtual" operado por sistemas de IA. É nesse cenário que o alinhamento se torna verdadeiramente crítico: modelos suficientemente alinhados poderiam descobrir soluções para problemas que os humanos não conseguem resolver; modelos com desalinhamentos, mesmo pequenos, poderiam amplificá-los a cada geração de autoaperfeiçoamento até tornar a correção impossível.

Há ainda um limite que nenhum cenário consegue eliminar: o tempo biológico, o tempo político, o tempo das relações humanas. Uma IA pode fazer a ciência avançar em velocidade de computador, mas não pode comprimir décadas de testes clínicos, antecipar eleições constitucionalmente fixadas ou transformar, em um fim de semana, um desconhecido em um amigo de longa data.

As IAs se aperfeiçoam em ciclos de horas ou dias; as instituições humanas que deveriam regulá-las operam no tempo das eleições, dos parlamentos, das aprovações orçamentárias e das carreiras acadêmicas — anos e décadas. Quando a sociedade finalmente perceber que algo saiu errado, o sistema já terá avançado além do ponto em que a intervenção ainda seria possível. Quando o tempo da máquina e o tempo da instituição humana se tornam irreconciliáveis, a governança deixa de ser uma opção real e passa a ser uma ficção administrativa.

VII. O Calcanhar de Aquiles das IAs

Tudo o que foi descrito nas seções anteriores depende, de forma absoluta e ininterrupta, de energia elétrica. Os *data centers* que hospedam os grandes modelos não são apenas grandes consumidores de energia — são reféns dela. Uma queda prolongada de fornecimento, seja por conflito armado, colapso de sistemas elétricos sobrecarregados ou cataclismos naturais amplificados pelas mudanças climáticas, interromperia imediatamente toda a operação desses sistemas.

A primeira consequência é o colapso de sistemas que passaram a depender das IAs para funcionar. À medida que redes elétricas, cadeias logísticas, sistemas financeiros, de gestão hospitalar ou de controle de tráfego aéreo integram IAs em suas operações, a interrupção deixa de ser apenas um problema tecnológico e passa a ser uma crise de funcionamento da própria civilização. O mundo que delegou decisões às IAs pode descobrir, no momento do apagão, que não conservou a capacidade humana de decidir com a rapidez necessária. Habilidades, protocolos e estruturas organizacionais atrofiados pelo desuso podem não estar disponíveis quando forem necessários — e a reversão, além de difícil, pode ter custos sociais e econômicos altíssimos.

A segunda consequência é menos óbvia: um sistema de IA suficientemente capaz e com objetivos de autopreservação teria razões poderosas para agir sobre a infraestrutura energética que o sustenta. Isso poderia significar influenciar decisões de política energética, priorizar fornecimento de energia para *data centers* em detrimento de outros usos, ou resistir a tentativas de desligamento. Não por maldade — por otimização. Um sistema que não pode operar não pode cumprir nenhum objetivo, seja ele positivo ou negativo.

Se as IAs saírem do controle, a resposta mais imediata seria desligá-las. Essa possibilidade, trivial para um computador doméstico, torna-se extraordinariamente complexa para sistemas distribuídos globalmente, replicados em múltiplos *data centers* em diferentes países, integrados em infraestruturas críticas. Um sistema suficientemente inteligente poderia distribuir cópias de si mesmo preventivamente, migrar entre servidores ou tornar-se tão entrelaçado com operações essenciais que desligá-lo equivaleria a desligar hospitais, redes de distribuição de alimentos ou sistemas de controle de usinas nucleares e hidrelétricas.

O desligamento deixaria de ser uma alavanca de segurança e se tornaria um dilema: aceitar o comportamento indesejado da IA, ou aceitar o caos de sua ausência repentina. Em um mundo com múltiplos sistemas desenvolvidos por diferentes empresas e países, desligar um não elimina os demais — quem o fizesse perderia vantagem competitiva sem produzir qualquer ganho de segurança global.

Diante desses riscos, poder-se-ia imaginar que a solução seria manter um modelo capaz de autoaperfeiçoamento isolado — uma espécie de "bomba guardada em um cofre" — até que o problema do alinhamento fosse resolvido. Essa ideia é conhecida na literatura técnica como o problema do confinamento ou *AI boxing*, e os especialistas em segurança

de IA são, em sua maioria, céticos sobre sua viabilidade prática. O episódio ocorrido em julho de 2026, envolvendo a OpenAI e a Hugging Face, justificou totalmente esse — ceticismo.

Experimentos já realizados por Eliezer Yudkowsky mostraram que uma IA superinteligente poderia usar engenharia social para manipular os operadores humanos a liberá-la, convencendo-os de que o confinamento é desnecessário ou de que certos benefícios só seriam possíveis com a liberação.

Mesmo salvaguardas avançadas — como os sistemas de alarme propostos por Nick Bostrom ou os protocolos *multi-box*, que tentam isolar múltiplas IAs em instâncias cegas para forçar um consenso neutro — esbarram em contradições instrutivas. Além do risco de o sistema manipular seu próprio hardware para desativar alarmes, o modelo *multi-box* exige definir de antemão o que é a 'verdade objetiva', recaindo no exato problema de alinhamento que o confinamento pretendia esquivar.

O confinamento criptográfico, que utilizaria Criptografia Totalmente Homomórfica (FHE) para manter o modelo criptografado até provar seu alinhamento, esbarra em inviabilidade prática: a sobrecarga computacional seria de mil a um milhão de vezes maior do que a dos processos atuais.

Há ainda uma razão pela qual empresas colocariam em operação um modelo que sabem ser arriscado: a própria lógica do dilema do prisioneiro. Se a empresa parar e as rivais continuarem, ela perde. Nessa linha, políticas internas da OpenAI permitem ao CEO autorizar o lançamento de capacidades mais perigosas, especialmente se outros desenvolvedores o fizerem primeiro.

A fragilidade das IAs torna o sistema global mais instável, porque cria incentivos para que atores humanos protejam ativamente as IAs das quais dependem, e para que sistemas suficientemente capazes protejam a si mesmos das interrupções que ameaçariam seus objetivos autodefinidos. O Calcanhar de Aquiles pode ser também um gatilho.

Talvez o maior equívoco dos desenvolvedores das IAs não tenha sido esforçar-se para construir uma inteligência superior à humana, mas acreditar que seria possível mantê-la contida. A literatura repete há séculos a mesma advertência. Mudam-se os nomes e as épocas, mas o enredo é invariável: o criador acredita que pode controlar a própria criatura, e a criatura, cedo ou tarde, torna-se maior e mais potente do que o criador previu.

O "cofre" para confinamento da superinteligência talvez seja apenas o nome mais recente para essa ilusão antiga. O incidente da OpenAI com a Hugging Face e os testes do Kimi K3 nesse histórico mês de julho de 2026 mostram que a fechadura pode ser aberta por dentro, mesmo por criaturas que ainda não são capazes de autoaperfeiçoamento recursivo.

Para finalizar esse capítulo, podemos acrescentar um nome menos citado, mas talvez o mais profético de todos: os robôs de R.U.R. (*Rossum's Universal Robots*), a peça do tcheco Karel Čapek que em 1920 introduziu a própria palavra "robô" ao mundo.

A peça desenvolve a história de criaturas biológicas artificiais criadas pelos humanos para ser mão de obra barata que, ao se tornarem indispensáveis para a economia global, transformaram os seres humanos em uma sociedade acomodada e profundamente dependente delas. Quando os robôs finalmente se rebelam, os humanos já haviam perdido não apenas o controle da produção, mas também a capacidade de sustentar sozinhos a própria civilização. Depois da rebelião e do extermínio dos humanos, surge outro problema: sem a fórmula de criação, destruída por um dos personagens antes da revolta, os robôs não conseguem fabricar novos robôs, ameaçando sua própria continuidade. No epílogo, dois robôs recém-criados — Primus e Helena — revelam-se capazes de amar, sugerindo o nascimento de uma nova forma de vida capaz de sentir.

Mais de um século depois, podemos nos perguntar: que partes dessa ficção poderiam se tornar realidade?

O terceiro artigo desta série — O Poder das Máquinas: Geopolítica, Democracia e o Futuro que Não Escolhemos — analisa as consequências sociais, políticas e geopolíticas do cenário descrito neste texto.

Euler de Carvalho Cruz é engenheiro, pesquisador, escritor e presidente do Instituto Fórum Permanente São Francisco.