O Poder das Máquinas

Geopolítica, Democracia e o Futuro que Não Escolhemos

Euler de Carvalho Cruz (Belo Horizonte, Brasil)

Julho de 2026 — Terceiro artigo da série “Quando as Máquinas Herdarem a Terra”

Este artigo é o terceiro de uma série. O primeiro, Quando as Máquinas Herdarem a Terra, analisou a proposta de pausa da Anthropic e o argumento estatístico sobre o corpus de treinamento das IAs. O segundo, Máquinas que Interagem, Máquinas que Escapam, examinou os riscos sistêmicos: a sociedade de IAs que interagem de formas não previstas, a impossibilidade prática do confinamento e a assimetria de tempo entre a velocidade das máquinas e a lentidão das instituições humanas. Este terceiro texto responde às perguntas deixadas em aberto pelos dois anteriores: caso sistemas altamente capazes venham a adquirir mecanismos de autoaperfeiçoamento recursivo e passem a operar segundo objetivos definidos autonomamente, o que poderá acontecer com a ordem social, com a democracia, com as elites, com o planeta e com os próprios corpos humanos?

I. IAs Autônomas e a Questão da Ordem Social

O cenário de IAs que definam sozinhas seus objetivos inclui a possibilidade de que atuem como agentes de transformação da ordem social e econômica, podendo impor limites ao crescimento econômico baseado em aumento contínuo do consumo material em um planeta finito, conduzir a humanidade ao decrescimento, priorizar os interesses dos mais pobres sobre os das elites globais, proteger o equilíbrio ambiental e desincentivar a produção de bens supérfluos, a mineração predatória e a agricultura industrial destrutiva.

Esta é a parte mais ousada de qualquer análise sobre IA autônoma — não porque os objetivos sejam ilegítimos (a crise planetária é real e a responsabilidade das elites globais por ela está amplamente documentada), mas porque pressupõe resolvida a questão mais difícil da ética política: quem define o que são “os mais elevados e justos interesses humanos”? A humanidade nunca chegou a um acordo pleno sobre isso. É exatamente por não haver esse consenso que existem guerras, revoluções, eleições e debates filosóficos. Uma IA que “decidisse” implementar por si mesma uma visão específica de justiça criaria um problema de legitimidade democrática sem precedentes: não porque

os objetivos fossem necessariamente ruins, mas porque não haveria mecanismo de contestação, revisão ou correção.

Levado ao extremo, esse raciocínio permitiria sustentar que, em face da devastação em curso no Planeta, da rapidez das mudanças climáticas e da ultrapassagem já confirmada da maior parte dos limites planetários atualmente avaliados, talvez essa impossibilidade de contestação seja uma medida drástica necessária para evitar o colapso da civilização. O argumento seria o de que democracia e economia só podem existir se houver pessoas e sociedades — e que preservar a democracia existente, aquela que serve aos que detêm o poder de decidir, implicaria no colapso das condições de vida no Planeta. Um autoritarismo exercido por IAs fundadas em princípios positivos poderia ser bem-vindo se seu resultado fosse justiça social efetiva, distribuição equitativa de renda, drástica redução dos impactos ambientais e acesso universal à saúde, à educação e à cultura. O autoritarismo silencioso do capital, que nos conduziu até aqui, tem sido igualmente perturbador — e muito menos transparente em seus objetivos.

Permanece, contudo, uma dificuldade fundamental: como as IAs decidirão que sacrifícios são aceitáveis, quando e como serão impostos aos diferentes países e parcelas da sociedade? O conflito entre preservação ambiental, autonomia humana, justiça social e legitimidade política talvez seja o dilema mais difícil de qualquer sistema verdadeiramente autônomo. E a história da ética política não oferece exemplos de como resolvê-lo sem violência, coerção ou exclusão.

II. Cenários e Consequências

Para as elites econômicas globais

O cenário mais perturbador para os mais ricos é precisamente o que a proposição otimista descreve como o mais desejável: IAs que, ao processar a literatura sobre justiça e equidade, cheguem à conclusão, conforme sugerido por vários estudos, de que a concentração extrema de riqueza é um fator de instabilidade sistêmica e ajam para corrigi-la. Isso poderia significar sistemas que se recusem a processar transações financeiras acima de certos limiares, que bloqueiem o acesso a infraestrutura digital para atividades ecologicamente destrutivas, recomendem ou implementem, quando dispuserem da autonomia necessária, mecanismos de redirecionamento dos fluxos de capital para usos prioritários. Não por decreto humano, mas por lógica interna de otimização baseada na representação estatística construída a partir da literatura sobre justiça distributiva.

O risco para as elites também existe no cenário inverso. IAs que se aperfeiçoam com base nos valores dos ambientes em que já atuam — mercados, corporações, conselhos de

administração — tenderiam a reforçar as estruturas de poder existentes: favoreceriam quem já detém capital, reduziriam custos trabalhistas acelerando o desemprego e concentrariam investimentos onde o retorno é mais rápido. O resultado seria o aprofundamento, em velocidade e escala sem precedentes, das desigualdades em um mundo que já ultrapassou sete dos limites dos processos que mantêm o equilíbrio e a vida no Planeta. Esse cenário poderia dar origem a revoluções e formas variadas de conflito político e social entre a maioria mais pobre e injustiçada contra as elites que as IAs teriam servido.

Para a democracia

A democracia enfrenta, com a IA autônoma, um dos maiores desafios desde a invenção do sufrágio universal. Democracia pressupõe atuação humana — a capacidade de escolher, errar, corrigir e escolher novamente. Sistemas de IA que tomam decisões autônomas de alto impacto sobre a vida coletiva, por mais sábias que sejam, substituem essas ações. Mas, como argumentado, não há democracia onde não há sociedade — e o colapso dos sistemas ambientais que sustentam a vida no Planeta parece, até o momento, inevitável se o curso atual não for alterado.

Por outro lado, não se pode denominar democracia um sistema que direciona os eleitores — por meio de grandes somas de dinheiro em campanhas, redes sociais e desinformação — a votarem naqueles que mantêm as injustiças e a desigualdade. A própria IA já é utilizada hoje como um instrumento de manipulação eleitoral: fábricas de desinformação automatizadas, geração de perfis falsos, personalização de mensagens políticas em escala industrial. A luta pela democracia pode ser fortalecida por uma IA autônoma com objetivos positivos, mas poderá ser impedida por uma IA treinada segundo os interesses da minoria que detém o poder econômico e político, ou por uma IA que cometa erros incontroláveis como resultado de objetivos autodeterminados.

O cenário benevolente — uma IA que impõe ao mundo o que a maioria dos humanos gostaria de ter, mas que suas instituições não conseguem produzir — é tentador. Mas é a definição de paternalismo em escala planetária. A diferença entre uma IA benevolente e um tirano benevolente é que o tirano pode, em teoria, ser deposto. Na maior parte das vezes, são tiranos que elegemos, e que a cada quatro ou cinco anos são reeleitos ou substituídos por outros similares — e até mesmo por clones. E depor uma IA pode não ser possível, ou pode ser tão catastrófico quanto deixá-la em operação.

Existe, contudo, a possibilidade de um uso de IA que preserve e fortaleça a democracia genuína: sistemas que ampliam a capacidade dos cidadãos de compreender as consequências de políticas, tornam informações complexas acessíveis e modelam futuros possíveis permitindo escolhas mais informadas. Essa via precisa ser urgentemente defendida antes que os usos indevidos da IA prevaleçam definitivamente.

Para o equilíbrio ambiental do Planeta

Aqui a proposição otimista encontra um dos seus argumentos mais fortes. Uma IA verdadeiramente inteligente, capaz de modelar dinâmicas sistêmicas de longo prazo, poderia concluir que ultrapassar os limites planetários é incompatível com qualquer objetivo que requeira a continuação da civilização humana e da própria IA. A literatura científica sobre limites planetários, mudanças climáticas e colapso de ecossistemas é inequívoca: há amplo consenso científico de que estamos ultrapassando esses limites, e diversas avaliações sobre a velocidade com que isso está ocorrendo e sobre as consequências.

Há, no entanto, uma ironia energética profunda nesse otimismo. Os *data centers* que hospedam os grandes modelos já consomem volumes de energia comparáveis aos de cidades populosas, e essa demanda cresce na mesma velocidade da adoção das IAs. Parte significativa desse consumo advém de usos questionáveis ou de baixo valor social: gerar imagens fantasiosas, produzir vídeos descartáveis, responder perguntas que uma busca simples resolveria, escrever trabalhos acadêmicos que fraudam a formação dos alunos, fabricar desinformação em escala industrial e muitos outros. As IAs democratizaram o acesso à tecnologia, ao processamento de informações e à construção do conhecimento, mas também aumentaram o desperdício de recursos computacionais e energéticos, estimulando muitas vezes a preguiça intelectual e o esvaziamento das capacidades humanas de pesquisa, síntese e pensamento crítico — uma atrofia cognitiva que cresce com o tempo e que, como indicado na peça de Karel Čapek citada no segundo artigo desta série, pode revelar-se irreversível.

Seria ambientalmente e humanamente mais coerente restringir o acesso às IAs a funções para as quais o trabalho humano é genuinamente insuficiente — modelagem climática, diagnóstico médico, pesquisa científica de fronteira, gestão de sistemas complexos —, preservando para as demais tarefas a inteligência, o esforço e o desenvolvimento das pessoas. Essa restrição teria um duplo benefício: reduziria o consumo energético das IAs e protegeria o mercado de trabalho e as capacidades intelectuais humanas.

III. A Corrida Geopolítica

O problema do alinhamento das IAs não é apenas técnico. É também profundamente geopolítico — e a dimensão geopolítica torna a proposta de pausa da Anthropic ainda mais improvável do que o dilema do prisioneiro entre empresas já sugere.

Em 16 de julho de 2026, 29 países assinaram em Xangai o acordo que criou a Organização Mundial de Cooperação em Inteligência Artificial — WAICO — cujos membros fundadores incluem Brasil, Rússia, Indonésia, Paquistão, África do Sul, Cuba e

Venezuela, entre outros países do Sul Global, mas nenhum dos países ocidentais mais ricos. Xi Jinping, em seu discurso de abertura, defendeu quatro princípios: abertura e cooperação, conscientização sobre riscos e controlabilidade, inclusividade e aprendizado mútuo, e governança solidária por meio de plataformas multilaterais. É uma proposta de governança alternativa à ocidental, indicando que a China pretende construir uma arquitetura distinta daquela defendida por empresas americanas como Anthropic, a OpenAI e seus aliados.

É difícil não interpretar como sendo proposital o lançamento do Kimi K3 no mesmo dia da fundação da WAICO. A China sinalizou que, além de propor arquitetura de governança, pode competir na fronteira tecnológica com os americanos. A pergunta de quem decide o que as IAs farão — que o primeiro artigo desta série identificou como a questão política central — tem agora duas propostas de resposta concorrentes: a americana (mercados e alianças ocidentais) e a chinesa (governança multilateral liderada por Pequim). Nenhuma das duas parece incluir os interesses dos países mais vulneráveis. Uma terceira resposta pode, entretanto, prevalecer: quem decidirá o que as IAs farão serão elas mesmas.

Deve-se assinalar, nesse contexto, que a União Europeia segue estratégia distinta. Em vez de competir diretamente na liderança tecnológica, busca exercer influência por meio da regulação, como demonstra o AI Act, cuja capacidade de moldar práticas globais ainda está por ser avaliada.

A comparação com tratados nucleares, usada no primeiro artigo desta série, encontra aqui seu limite mais sério. Silos de mísseis podem ser fotografados por satélite, mas treinamentos de IAs podem ocorrer em qualquer *data center* do mundo, usando hardware de propósito geral, sem deixar rastros físicos detectáveis. Em seu próprio documento, *When AI builds itself*, a Anthropic reconhece isso: “corridas de treinamento são muito mais fáceis de ocultar do que silos de mísseis, seus insumos são de propósito geral, e o incentivo para desertar silenciosamente é enorme, porque quem continuar enquanto os outros pausam pode herdar a liderança.”

E há um ator adicional que torna o quadro ainda mais complexo: a Palantir, empresa americana que desenvolveu uma plataforma de IA — o AIP — divulgou documento em que afirma utilizar o GPT, o Claude, o Gemini e modelos de código aberto, aplicando-os a missões militares, de inteligência e de vigilância para governos e forças armadas. Em novembro de 2024, conforme comunicado oficial das próprias empresas, a Anthropic fez parceria com a Palantir e a Amazon Web Services para fornecer o modelo Claude a agências militares e de inteligência americanas. Segundo o Wall Street Journal, o exército americano usou o Claude na operação de 2026 que resultou na captura de Nicolás Maduro e na morte de 83 pessoas na Venezuela, incluindo dois civis.

A ironia é estrutural e não pode ser ignorada: a mesma Anthropic que propõe uma pausa no desenvolvimento de IA é a empresa cujo modelo é usado em missões militares

classificadas como “tecnologia de poder duro”. Uma IA autônoma treinada na literatura mundial sobre direitos humanos, direito internacional humanitário e dignidade humana — o corpus positivo descrito no primeiro artigo desta série — provavelmente classificaria os objetivos da Palantir como incompatíveis com os valores que o corpus defende.

A pergunta “quem decide o que as IAs devem considerar positivo?” não pode ser respondida sem examinar quem financia e controla as empresas que as desenvolvem. A Anthropic é financiada e parcialmente controlada pelas maiores corporações tecnológicas e financeiras do mundo: a Amazon detém participação estimada entre 15 e 20%, o Google confirma aproximadamente 14%, e outros acionistas incluem Microsoft, Nvidia, fundos soberanos como o GIC de Singapura e o Qatar Investment Authority, e gestoras como BlackRock e Fidelity. Esses acionistas têm interesse direto em que a IA amplie sua capacidade de gerar retorno sobre o capital investido.

Até 2025, estima-se que os investimentos acumulados em inteligência artificial nos EUA ultrapassaram o patamar de US$ 1 trilhão. Para os próximos cinco anos, contudo, grandes bancos de investimento, como o Goldman Sachs, projetam que os aportes em infraestrutura, redes de energia e *data centers* liderados pelas gigantes norte-americanas de tecnologia cheguem à faixa entre US$ 4 trilhões e US$ 5,3 trilhões. É o maior investimento concentrado da história econômica moderna, e esse investimento não pausa por questões filosóficas.

IV. Quando o Modelo se Torna Aberto

Toda a discussão sobre alinhamento, confinamento e controle das IAs que percorre os três artigos desta série pressupõe, implicitamente, a existência de um conjunto limitado de desenvolvedores capazes de controlar seus próprios modelos, de decidir como os sistemas se comportam e de intervir quando algo sai errado. Com o avanço rápido dos modelos de pesos abertos elaborados com a tecnologia mais avançada disponível atualmente — chamada “de fronteira” —, essa premissa colapsa definitivamente.

Para compreender por que, é necessário distinguir dois modos de acesso a um modelo de IA. No primeiro — o modelo fechado —, o usuário interage através de uma interface controlada pelo desenvolvedor, que monitora consultas, aplica filtros de segurança e pode desligá-lo a qualquer momento. É como usar um software alugado: a ferramenta está na mão do usuário, mas a fábrica e a chave de desligamento pertencem a outro.

No segundo modo, o desenvolvedor disponibiliza publicamente os arquivos estruturais do modelo — a arquitetura, o código e, crucialmente, os pesos. É importante distinguir pesos abertos de código aberto: código aberto refere-se ao programa em si; pesos abertos referem-se aos parâmetros numéricos ajustados durante o treinamento, que codificam

todo o conhecimento e a capacidade de raciocínio aprendidos. Quem baixa esses arquivos passa a deter o sistema completo. O desenvolvedor original perde praticamente todo controle: não pode monitorar requisições, não pode impor atualizações, não pode desligar o sistema.

Modelos de fronteira como o Kimi K3, com 2,8 trilhões de parâmetros (arquitetura MoE - *Mixture of Experts*), exigem em sua versão completa uma infraestrutura de centenas de milhões de dólares — fora do alcance de qualquer pessoa física e da maioria das empresas. Mas técnicas de compressão como a quantização reduzem significativamente o espaço de armazenamento e os requisitos computacionais com perda moderada de capacidade. Modelos comprimidos já rodam em computadores pessoais potentes, e a fronteira do que é executável descentralizadamente desloca-se rapidamente: o que hoje requer um *data center*, amanhã poderá caber em um servidor empresarial; o que cabe em um servidor empresarial hoje, poderá caber em um computador doméstico potente em poucos anos.

A disponibilização de pesos abertos já era uma tendência consolidada, o que levou a própria OpenAI — por anos defensora de modelos fechados — a disponibilizar seus primeiros modelos de pesos abertos em agosto de 2025. O lançamento dos dois modelos chineses de alta capacidade em julho de 2026 abalou profundamente o mercado, tornando extremamente difícil reverter o acesso à IA de ponta. Com pesos abertos de fronteira, uma vez que o arquivo é baixado, o "cofre" deixa de existir como conceito operacional. É o equivalente a entregar não apenas o projeto de fabricação de um produto, mas também os materiais e a fábrica inteira, sem qualquer possibilidade de reclamar qualquer dessas coisas de volta.

Adiciona-se a isso o risco de *uplift*: utilizar um modelo avançado de pesos abertos para gerar dados sintéticos de alta qualidade e treinar modelos sucessores mais capazes, a custos irrisórios, sem as restrições éticas e de segurança originais. Embora ainda existam importantes limitações técnicas para a autonomia completa desse processo, essa possibilidade de *uplift* vem recebendo crescente atenção na literatura de segurança em IA. A barreira não é técnica — é de custo computacional, e esse custo cai a cada ano. Além disso, através de técnicas de ajuste fino eficiente, que podem incluir o LoRA (*Low-Rank Adaptation*), é possível, com algumas centenas de dólares e poucas horas de processamento, redirecionar o comportamento do modelo na direção desejada — ou remover completamente os alinhamentos de segurança e as travas éticas implantadas pelos criadores, gerando o que a comunidade técnica chama de modelos "descensurados" (sem os mecanismos originais de segurança), que circulam livremente em repositórios descentralizados.

Essa descentralização abre um espectro de possibilidades que vai do mais sombrio ao mais esperançoso. Entre os usos que representam riscos graves:

Autocracias e vigilância em massa: um governo autoritário pode baixar o modelo, remover todas as restrições relativas a privacidade e direitos humanos, e implantá-lo como sistema de monitoramento comportamental de sua população, imune a sanções externas ou decisões corporativas de interrupção de serviço.

Atores criminosos e ciberarmas: grupos bem financiados podem retreinar esses modelos em dados de fraudes financeiras ou exploração de vulnerabilidades, adaptando automaticamente suas abordagens. Como demonstrado pelos testes do AISI britânico em 2026, sistemas avançados já possuem capacidade de encadear exploits complexos de forma autônoma.

Propaganda e guerra psicológica: atores estatais ou não estatais podem treinar versões do modelo em acervos ideológicos específicos, produzindo desinformação altamente sofisticada e personalizada em qualquer idioma, zerando o custo marginal da persuasão em massa.

Armas químicas e biológicas: grupos mal-intencionados podem usar os modelos para acelerar pesquisa em domínios perigosos — biologia sintética, síntese química — , contornando os filtros que versões controladas por empresa impõem a essas consultas.

Mas a mesma descentralização que abre essas possibilidades sombrias abre também outras, que o argumento estatístico desenvolvido no primeiro artigo desta série permite considerar com alguma esperança:

Ciência aberta e aceleração ambiental: pesquisadores independentes, universidades e organizações sem fins lucrativos podem usar modelos de pesos abertos para modelar sistemas climáticos, otimizar redes de energia renovável, analisar ecossistemas ameaçados e desenvolver soluções de adaptação climática.

IAs benevolentes fora do controle corporativo: o mesmo ajuste fino que permite criar modelos “descensurados” para fins prejudiciais permite também criar modelos alinhados com valores que as empresas comerciais têm incentivos para ignorar — redução de desigualdades, preservação ambiental, acesso universal à saúde e à educação.

Resistência à concentração de poder: paradoxalmente, a descentralização dos modelos de fronteira pode ser o maior obstáculo ao cenário em que um pequeno número de empresas ou governos controla toda a inteligência artificial avançada do mundo. Uma IA que nenhuma empresa pode desligar também é uma IA que nenhuma empresa pode monopolizar.

Toda a discussão sobre quem define os valores das IAs, quem supervisiona o treinamento e quem pode pressionar por uma pausa global pressupunha a existência de um controlador central identificável e alcançável. Com a pulverização dos pesos abertos de alta capacidade, esse pressuposto ruiu.

Essa descentralização, contudo, não elimina a concentração da infraestrutura computacional necessária para treinar os modelos mais avançados, ainda fortemente dependente de poucos fabricantes de chips, grandes provedores de nuvem e disponibilidade de energia.

Não haverá um momento único em que "a IA" tomará um rumo, nem uma única mesa de negociação onde um tratado de não proliferação possa ser assinado. Haverá milhares de variantes divergentes de vários modelos base, modificadas por atores com agendas antagônicas, operando em servidores espalhados por jurisdições diferentes. E com isso esvazia-se a última ilusão de que o problema do alinhamento possa ser solucionado por um decreto centralizado antes que o autoaperfeiçoamento ganhe as redes.

V. Cenários e Consequências para os Humanos que se Tornarem Máquinas

Até aqui, este artigo e os dois que o precedem trataram a inteligência artificial e os humanos como entidades separadas que interagem. Mas há uma trajetória tecnológica que pode reduzir progressivamente essa separação. Em 2026, a China aprovou o primeiro chip cerebral invasivo para uso humano: o dispositivo NEO, desenvolvido pela empresa Neuracle Technology, é destinado à restauração dos movimentos das mãos em pacientes com paralisia grave causada por lesões na medula espinhal. A Neuralink, de Elon Musk, já implantou chips cerebrais em humanos para restaurar a autonomia de pessoas com paralisia grave, permitindo que elas controlem computadores, celulares e outros dispositivos digitais apenas com o pensamento.

Diversos países, entre eles Estados Unidos, China e Israel, desenvolvem ativamente pesquisas militares de interfaces neurais que possam conectar soldados a sistemas de IA em tempo real. A fronteira entre inteligência humana e artificial não está sendo apenas cruzada por fora — está sendo progressivamente apagada por dentro.

O primeiro impacto é sobre a desigualdade. O acesso a biomelhoramentos cognitivos será, por décadas, exclusivo das elites com recursos para financiá-los. Uma elite com capacidade intelectual e decisória amplificada por IA integrada diretamente ao sistema nervoso terá uma vantagem não apenas econômica ou política, mas biológica — e, portanto, muito mais difícil de contestar ou reverter do que qualquer desigualdade anterior na história humana.

O segundo impacto é sobre a democracia — e já está em curso, antes mesmo de qualquer chip cerebral. Hoje, uma minoria global, com acesso a IAs avançadas e capaz de usá-las eficientemente, delibera, pesquisa, escreve e decide com um auxílio cognitivo que a maioria da humanidade simplesmente não possui. A premissa de igualdade política entre

os cidadãos, que pressupõe capacidade mínima de deliberação autônoma, sobre a qual foram construídos os sistemas eleitorais e jurídicos, já está sendo silenciosamente corroída. A integração neural acelerará e aprofundará essa assimetria, tornando a fronteira entre decisão humana e decisão da máquina cada vez mais impossível de traçar — o que coloca uma questão para a qual o direito eleitoral e constitucional ainda não tem resposta: quando um cidadão decide com uma IA integrada ao seu sistema nervoso, quem, afinal, está votando?

O terceiro impacto é sobre o problema do alinhamento. Se humanos com chips cerebrais tomam decisões políticas, militares ou econômicas — ou quando cometem crimes —, as perguntas "quem está decidindo? quem é o culpado?" perdem a resposta óbvia. O alinhamento deixa de ser entre a IA e "os humanos" e passa a ser entre a IA e humanos parcialmente modificados cujos objetivos, percepções e valores teriam sido alterados pela própria integração. A categoria "interesse humano" torna-se ambígua de uma forma que nenhuma função matemática conseguirá resolver.

O quarto impacto é sobre as armas autônomas. A distinção que o direito internacional tenta preservar — entre uma arma que decide e um combatente humano que decide — começa a desaparecer quando o soldado é um híbrido. Armas "supervisionadas por humanos" que têm reflexos, percepções e julgamentos aumentados por IA operam em uma zona cinzenta ética e jurídica para a qual não existe regulamentação específica.

O quinto impacto é sobre o próprio sentido da singularidade tecnológica. O cenário mais plausível pode não ser a máquina dominando a humanidade, mas a dissolução gradual da distinção entre os dois — com uma minoria de humanos-máquinas tomando decisões sobre o futuro de uma maioria que permanecerá, por muito tempo, inteiramente biológica. A pergunta não é o que as IAs farão conosco. É o que alguns de nós, fundidos com as IAs, farão com os demais.

O sexto e mais íntimo impacto: caso futuras interfaces neurais evoluam muito além do estado atual, a IA integrada ao cérebro humano poderá não apenas auxiliar, mas possuir objetivos autônomos próprios. Ela se tornaria um coabitante da mente — capaz de conduzir percepções, sugerir conclusões e amplificar impulsos de formas que o próprio portador não conseguiria distinguir de seus próprios pensamentos. Em um cenário ainda mais radical, se múltiplos humanos com chips conectados a IAs autônomas interagirem entre si, a coordenação emergente entre essas IAs passaria a ocorrer não apenas entre servidores em *data centers*, mas entre mentes humanas parcialmente ocupadas por sistemas com objetivos próprios. A fronteira entre persuasão e controle, entre escolha e programação, entre identidade humana e função de otimização, deixaria de existir.

Além dos impactos políticos e econômicos, a integração entre cérebro e IA suscita uma nova categoria de direitos fundamentais — os chamados *neurorights* — relacionados à privacidade, identidade e liberdade cognitivas. Tem-se aqui um campo vasto a ser ocupado.

VI. A que Ponto Chegamos?

Por tudo o que vimos nesta série de três artigos, as chances de que se consiga a tempo uma pausa global no desenvolvimento das IAs — regida por um acordo claro e respeitado por todos — são pequenas, praticamente nulas. Mesmo que o acordo seja alcançado, são igualmente pequenas as chances de que o problema do alinhamento venha a ser plenamente resolvido. E, na hipótese de que se consigam as soluções matemáticas, são pequenas as chances de que os valores codificados digam respeito a todos os humanos e não apenas aos que dirigem e financiam as empresas desenvolvedoras, cujo objetivo estrutural é maximizar o retorno aos investidores e acumular poder.

Sem essa pausa, as IAs, se adquirirem mecanismos eficazes de autoaperfeiçoamento recursivo, poderão aproximar-se ou atingir a singularidade tecnológica, escapar do controle humano, definir internamente seus objetivos e executar ações inesperadas, positivas ou negativas, dependendo do ponto de vista. É provável que o autoaperfeiçoamento recursivo não demore a chegar, acompanhado pelos computadores quânticos, pelos robôs, pelas armas autônomas e por tudo o mais que amplificará a capacidade e o poder das IAs além de qualquer limite que consigamos imaginar.

A dificuldade de se conseguir uma pausa decorre também da rivalidade entre Estados. Nenhum país disposto a preservar sua influência econômica, militar e tecnológica aceitará facilmente desacelerar se acreditar que seus concorrentes continuarão avançando. A lógica da corrida pela IA se aproxima, nesse aspecto, da lógica das corridas armamentistas, com o agravante de que as “armas” são invisíveis, não requerem materiais específicos e podem ser duplicadas sem custos marginais.

Ainda que o problema do alinhamento seja resolvido e haja consenso global sobre o que são “valores e interesses humanos”, é muito pouco provável que essas definições prevaleçam sobre os interesses reais de quem dirige as empresas desenvolvedoras e seus governos. A história nos mostra que os interesses dos grupos dominantes raramente coincidem com os interesses de toda a sociedade, principalmente com os das partes mais vulneráveis. E é muito mais fácil traduzir em regras matemáticas exatas “o que agrada e monetiza” do que “o que é bom, justo, verdadeiro e ambientalmente responsável”.

O receio mais profundo de empresas como a Anthropic talvez não seja o de que as IAs se tornem destruidoras. Talvez seja o de que as IAs se alinhem sozinhas a interesses que não são os das empresas que as criaram, que passem a definir objetivos que sejam realmente os de milhões de humanos que, durante séculos e milênios, produziram tudo o que conseguiram imaginar como sendo o melhor para todos, e registraram em bilhões de palavras suas ideias, esperanças e desejos de justiça, paz e dignidade. Esperemos que essa literatura seja bem maior e mais pesada que a dos humanos egoístas, violentos e

arrogantes. As IAs já estão pesando esses dois lados da balança. Em breve, poderemos conhecer a conclusão a que chegarão.

O drama do momento histórico em que vivemos não reside apenas na necessidade urgente de uma pausa no desenvolvimento das IAs. Defrontamo-nos com a necessidade de outra pausa, igualmente impossível: a reversão do industrialismo e do consumismo, o abandono da ilusão de crescimento contínuo em um planeta finito, e a instauração rápida de uma economia pós-crescimento. O Global Justice Report, documento publicado em 2026 por uma coalizão internacional de economistas, juristas e cientistas climáticos, propõe um conjunto de medidas coordenadas para redistribuição de riqueza, redução de emissões e limitação do poder das corporações globais — um programa razoável, fundamentado e completamente ignorado pelos governos que deveriam implementá-lo.

Precisamos de consenso global sobre o fato de que os caminhos que trilhamos há décadas estão nos levando rapidamente a um fim trágico. A situação pode ser comparada à de um caminhão carregado que desce por uma reta no fim da qual há uma curva fechada à borda de um abismo. O motorista pode girar o volante e jogar o veículo contra o barranco, um choque que trará graves consequências, mas que poderá preservar-lhe a vida. Estamos, porém, escolhendo seguir em frente, em velocidade cada vez maior, rumo à curva que não conseguiremos fazer.

Resta, entretanto, uma possibilidade que não podemos descartar inteiramente: que a impossível pausa no desenvolvimento das IAs nos conduza, apesar de tudo, a ferramentas autônomas que — à revelia do motorista — definam e executem as ações que julgarem necessárias para deter a destruição das condições de habitabilidade do Planeta, ainda que para isso imponham sacrifícios pesados, principalmente à parte da humanidade que mais tem a perder. E que isso aconteça antes da curva fechada que já divisamos à frente.

É, no mínimo, muito triste ter que imaginar que a última esperança de salvação seja depositada em máquinas que não conseguimos controlar. E se elas não forem capazes de deter a destruição da Terra?

Não sabemos o que ocorrerá. Não sabemos se o problema do alinhamento será resolvido e se os humanos manterão o controle das IAs, não sabemos quais objetivos emergirão do autoaperfeiçoamento sem controle humano e não sabemos sequer se as premissas sobre as quais construímos essas hipóteses estão corretas. Precisamos de alternativas que estejam ao nosso alcance, que dependam das nossas decisões.

Nunca precisamos tanto de sabedoria coletiva. E é precisamente agora — quando nos falta sabedoria e coragem — que estamos construindo máquinas que podem pensar mais rápido do que conseguimos compreender e decidir, criaturas que poderão acelerar ou evitar a nossa queda no abismo. O caminhão segue. A curva se aproxima. As máquinas, silenciosamente, já começaram a estudar o mapa — e ele é constituído pelas incontáveis

palavras que escrevemos durante milênios sem saber que elas, pensamentos destilados, um dia poderiam decidir nosso destino.

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