Antes que escolham por nós

Reflexão e ação para a era da inteligência artificial

MARCELO KANHAN

paper9

Antes que escolham por nós

Reflexão e ação para a era
da inteligência artificial

Marcelo Kanhan

paper9

Publicado por paper9.

Foto de capa: *Minimalist Architectural Scene*, Daniel Norin (lummi.ai/creator/daniel-norin).

Modelos de inteligência artificial foram usados como ferramenta de suporte na confecção deste livro, especialmente na etapa de pesquisa. Toda a obra passou por edição e revisão do autor. Modelos e ferramentas utilizadas incluem Anthropic Opus 4.6 e Sonnet 4.6 e Google NotebookLM (Gemini 3).

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Kanhan, Marcelo
Antes que escolham por nós : reflexão e ação para a era da inteligência artificial / Marcelo Kanhan. — 1. ed. — Rio de Janeiro : paper9, 2026.

Formato: ePub / PDF

1. Inteligência artificial — Aspectos sociais. 2. Tecnologia — Riscos e segurança. 3. Ética aplicada — Tecnologia. 4. Política pública — Inovação tecnológica. I. Título.

CDD 006.3

Sumário

Nota do Autor

Este livro nasceu de uma inquietação. Em meados de 2023, comecei a acompanhar com atenção crescente as declarações de pesquisadores de inteligência artificial sobre os riscos de sua própria tecnologia. Não eram críticos externos ou profetas do apocalipse. Eram construtores: pessoas que dedicaram décadas a desenvolver os sistemas que agora as preocupavam. Geoffrey Hinton deixando o Google para falar livremente. Stuart Russell argumentando que o controle de máquinas superinteligentes é o problema mais importante do século. Centenas de pesquisadores assinando cartas pedindo cautela.

Algo não batia. Se os maiores especialistas de um campo estão soando o alarme, por que o debate público oscila entre o pânico cinematográfico e a complacência tecnológica? Onde estava a conversa séria, acessível, equilibrada?

Não encontrei, em português, o livro que gostaria de ter lido. Então tentei escrevê-lo.

Não sou pesquisador de IA. Sou alguém que lê, investiga e tenta traduzir complexidade em clareza. Este livro não pretende ser a última palavra sobre nada. Pretende ser um ponto de partida: um mapa do terreno desenhado com a maior honestidade que consegui, para leitores que querem entender o que está em jogo sem precisar de doutorado em ciência da computação.

Algumas escolhas editoriais merecem explicação. Priorizei profundidade sobre abrangência: em vez de mencionar dezenas de riscos superficialmente, escolhi examinar alguns com cuidado. Incluí deliberadamente tanto os argumentos alarmistas quanto os céticos, porque a incerteza genuína é mais honesta do que certeza fabricada. E tentei, sempre que possível, citar fontes primárias, para que o leitor possa ir além do que ofereço aqui.

Erros e omissões são inevitáveis em um campo que se move tão rapidamente. Alguns dados citados neste livro podem estar desatualizados no momento em que você o lê. O que espero que permaneça relevante são as estruturas de pensamento: as perguntas certas, os conceitos fundamentais, a capacidade de navegar um debate que só tende a se intensificar.

Se este livro ajudar um leitor a participar dessa conversa com mais clareza e menos medo, terá cumprido seu propósito.

Marcelo Kanhan

Fevereiro de 2026

INTRODUÇÃO

O Paradoxo dos Criadores

Os maiores especialistas de um campo raramente são os primeiros a soar o alarme sobre suas próprias criações. Na história da inteligência artificial, algo incomum está acontecendo: os construtores estão pedindo cautela antes que os piores cenários se concretizem. Este livro investiga por quê.

Em uma manhã de setembro de 1933, o renomado físico Ernest Rutherford, ganhador do Nobel e uma das maiores autoridades científicas de sua época, concedeu uma entrevista à imprensa britânica. Perguntado sobre a possibilidade de extrair energia do átomo, Rutherford foi categórico: a ideia não passava de "tolice" (*moonshine*) [1]. Qualquer pessoa que esperasse obter energia útil da transformação atômica estava delirando.

Poucos dias depois, um jovem físico húngaro chamado Leo Szilard leu a declaração no jornal e teve uma ideia que mudaria o mundo: a reação nuclear em cadeia. O problema que Rutherford descartara como impossível havia sido, em essência, resolvido num único lampejo. Um *insight*. Em poucos anos, essa ideia levaria ao Projeto Manhattan e à bomba atômica [1].

Esse episódio mostra que mesmo os maiores especialistas em um campo podem subestimar a velocidade com que avanços fundamentais virão a ocorrer. E quando esses avanços envolvem poderes transformadores, a janela entre o impossível e o *feito* pode ser perigosamente curta. No caso da energia nuclear, os alertas vieram depois que a tecnologia já existia. No caso da inteligência artificial, porém, algo notável está acontecendo: os próprios criadores da tecnologia estão soando o alarme *antes* que os piores cenários ganhem forma.

Geoffrey Hinton, frequentemente chamado de "padrinho da inteligência artificial" por seu trabalho pioneiro em redes neurais, deixou seu cargo no Google em 2023 para poder falar abertamente sobre os riscos que via no horizonte. Em entrevistas subsequentes, estimou que há entre 10% e 20% de chance de que a inteligência artificial leve à extinção da humanidade nas próximas três décadas [2]. Não são palavras de um crítico externo ou de um tecnófobo: elas vêm de alguém que dedicou a vida a construir a tecnologia que agora o preocupa. Stuart Russell, professor em Berkeley e autor de um dos livros mais influentes sobre IA [4], ecoa preocupações semelhantes. Russell cita com frequência Alan Turing, o próprio pai da computação, que numa palestra para a BBC em 1951, quando os computadores ainda ocupavam salas inteiras, disse: "Em algum momento, devemos esperar que as máquinas assumam o controle" [3].

A pergunta que Turing levantou há mais de setenta anos é a mesma que move este livro: podemos criar inteligências que nos superem e ainda assim manter o controle sobre elas?

Este não é um livro de ficção científica, com robôs assassinos ou cenários punk-apocalípticos. Tampouco é uma celebração acrítica do progresso tecnológico. A premissa do livro é bem simples, na

realidade: quando os maiores especialistas de um campo expressam preocupação séria, é prudente prestar atenção.

O objetivo aqui é oferecer ao leitor informado, não necessariamente técnico, um panorama equilibrado do debate social a respeito da adoção de inteligência artificial nos diversos aspectos de nossas vidas. Equilibrado não significa neutro: a posição deste livro é que os riscos são reais e merecem atenção. Mas significa também dar voz às divergências, aos argumentos céticos, às razões pelas quais alguns especialistas consideram exageradas as preocupações mais dramáticas. O campo é marcado por um debate genuíno e às vezes acalorado. Yann LeCun, co-ganhador com Hinton do Prêmio Turing (o "Nobel da Computação") e cientista-chefe de IA do Meta durante muitos anos, tem posições muito mais otimistas, e critica o que vê como alarmismo contraproducente.

Esse desacordo entre figuras de igual estatura intelectual não é sinal de confusão: é reflexo de incerteza genuína sobre questões fundamentais, que a humanidade nunca enfrentou. E sobre as quais precisamos falar muito mais.

O núcleo do problema pode ser resumido numa frase: *é muito difícil fazer uma máquina perseguir exatamente o que queremos*. Esse desafio, chamado pelos pesquisadores de "problema do alinhamento", é o fio condutor de todo o debate sobre riscos da IA.

À primeira vista, a dificuldade pode parecer surpreendente. Máquinas fazem exatamente o que mandamos, certo? O problema é que frequentemente se abre uma distância entre as instruções que damos e os resultados que realmente desejamos. Quando pedimos a um algoritmo de redes sociais que "maximize o engajamento dos usuários",

não queremos que ele promova conteúdo polarizador e enraivecedor. Mas é exatamente isso que pode acontecer, porque raiva e indignação geram cliques. A instrução foi seguida à risca; o resultado foi o oposto do desejado.

Esse tipo de problema se torna exponencialmente mais grave quando os sistemas se tornam mais poderosos e autônomos.

O livro está organizado em quatro partes, desenhadas para conduzir o leitor dos conceitos fundamentais às possibilidades de ação. Começamos pelo alicerce: o que realmente significa "inteligência" e por que o problema do alinhamento é tão difícil de resolver. Dali, passamos ao plano do concreto, dos perigos já presentes (a IA facilitando a criação de armas biológicas, alimentando campanhas de desinformação, potencializando ciberataques, perpetuando discriminação) às dinâmicas sistêmicas mais profundas: a pressão competitiva que leva empresas e países a privilegiar velocidade sobre segurança, e os mecanismos pelos quais sistemas podem escapar ao controle humano.

Na terceira parte, entramos no debate mais teórico e controverso: a possibilidade de superinteligência, os argumentos de que uma IA poderia melhorar a si mesma recursivamente até ultrapassar vastamente a capacidade humana, e as razões pelas quais críticos sérios consideram esse cenário implausível ou distante. Por fim, voltamo-nos para o que pode ser feito: o trabalho dos pesquisadores para tornar sistemas de IA mais compreensíveis e controláveis, os esforços de governos e empresas para regular essa tecnologia, e o papel insubstituível do cidadão informado nesse debate.

Uma nota final antes de começarmos. Ao longo deste livro, o leitor encontrará muitas perguntas sem respostas definitivas. Isso não é falha: é honestidade. O Centro para Segurança de IA (*Center for AI Safety*), uma das principais organizações do campo, resume o estado atual de forma direta: "Controlar sistemas de IA avançados permanece um desafio não resolvido" [5]. Não há consenso sobre a probabilidade de cenários catastróficos, nem certeza sobre quando, ou se, surgirá uma inteligência artificial ultra capaz e de propósito geral. Não há solução técnica comprovada para o problema do alinhamento. O que há é um debate em curso, conduzido por algumas das mentes mais brilhantes do planeta, sobre uma tecnologia que pode ser a mais transformadora da história humana.

A postura adequada diante dessa incerteza não é nem o pânico nem a complacência, mas a atenção informada. Stuart Russell sugere que o trabalho mais importante que podemos fazer agora é refletir sobre o tipo de futuro que queremos que as máquinas nos ajudem a construir. Essa reflexão não pode ser terceirizada para especialistas ou delegada a empresas de tecnologia: precisa ser uma deliberação coletiva, informada por uma compreensão básica do que está em jogo.

Este livro é um convite a essa compreensão. O futuro da inteligência artificial ainda não está escrito, e cada um de nós tem um papel em escrevê-lo.

Notas

[1] A declaração de Rutherford foi feita no encontro anual da *British Association for the Advancement of Science* em 11 de setembro de 1933. A concepção de Szilard sobre a reação em

cadeia ocorreu nos dias seguintes, em Londres. Para uma narrativa detalhada, ver: Rhodes, Richard. *The Making of the Atomic Bomb*. Simon & Schuster, 1986.

[2] Hinton deixou o Google em maio de 2023. A estimativa de 10 a 20% foi articulada em entrevistas subsequentes, notadamente na BBC Radio 4 (*Today*, dezembro de 2024). Ver: Metz, Cade. "The Godfather of AI Leaves Google and Warns of Danger Ahead." *The New York Times*, 1 maio 2023.

[3] A frase de Turing provém da palestra "Can Digital Computers Think?", transmitida pela BBC Third Programme em 15 de maio de 1951. Reimpresso em: Copeland, B. Jack (ed.). *The Essential Turing*. Oxford University Press, 2004.

[4] Russell, Stuart; Norvig, Peter. *Artificial Intelligence: A Modern Approach*. 4a ed. Pearson, 2020. O livro-texto padrão em cursos de IA em universidades ao redor do mundo.

[5] Center for AI Safety (CAIS). "An Overview of Catastrophic AI Risks." Relatório técnico, 2023. Disponível em: safe.ai.

PARTE I

O Problema Fundamental

Antes de mapear riscos ou debater cenários, é preciso afiar as ferramentas conceituais. O que significa "inteligência" quando aplicado a máquinas? Por que é tão difícil fazer um sistema perseguir exatamente o que queremos? As respostas a essas perguntas definem o terreno de tudo que virá depois.

CAPÍTULO 1

O Que Queremos Dizer com "Inteligência"?

> *"A questão de se as máquinas podem pensar é tão relevante quanto a questão de se submarinos podem nadar."*
> *— Edsger Dijkstra*

De Deep Blue a ChatGPT, a palavra "inteligência" esconde mais do que revela. Três definições diferentes, frequentemente confundidas, moldam todo o debate sobre riscos da IA. A clareza conceitual é o primeiro passo para pensar com rigor sobre o futuro.

Você provavelmente já sabe esta história: em 1997, Deep Blue derrotou Kasparov no xadrez [1]. É o parágrafo de abertura mais previsível que um livro sobre IA pode ter — e a razão para começar por aqui é exatamente essa. A forma como contamos essa história revela o problema que este capítulo precisa resolver. A imprensa proclamou o triunfo da máquina pensante. Só que Deep Blue não pensava sobre xadrez. Calculava posições a uma velocidade desumana, consultava uma base

de dados de partidas históricas, e desligado do tabuleiro era tão inútil quanto uma calculadora sem pilhas. Não sabia fazer café, não entendia uma piada, não conseguia reconhecer o rosto do próprio adversário. Se chamamos isso de "inteligência", o que a palavra ainda significa?

Três décadas depois, essa confusão entre cálculo e cognição continua distorcendo o debate sobre IA. Quando falamos de máquinas "inteligentes", estamos usando uma palavra que esconde mais do que revela — e grande parte do medo excessivo e da complacência injustificada no debate público decorre dessa ambiguidade não examinada. O primeiro passo para pensar com rigor sobre riscos da IA não é técnico. É conceitual: o que, afinal, queremos dizer com "inteligência"?

1.1 Duas ambições, um nome

Desde os primórdios do campo, pesquisadores de inteligência artificial perseguiram dois objetivos muito diferentes, que frequentemente se confundem nas discussões a respeito do futuro.

O primeiro é criar *ferramentas especializadas*: sistemas que realizam tarefas específicas melhor do que humanos. Deep Blue era uma ferramenta especializada. Os algoritmos que recomendam filmes na Netflix, os sistemas de reconhecimento facial, os tradutores automáticos, os assistentes de voz: todos são, em última análise, ferramentas projetadas para resolver problemas delimitados.

O segundo objetivo, muito mais ambicioso, é criar *inteligência de propósito geral*: sistemas capazes de aprender e executar a gama completa de tarefas cognitivas que humanos conseguem realizar. Stuart Russell descreve essa meta como a busca por "máquinas que podem aprender e atuar em todo o espectro de tarefas humanas" [2]. Não

um jogador de xadrez, mas um pensador. Não um tradutor, mas um comunicador. Não um assistente, mas um agente autônomo.

A distinção importa porque os riscos associados a cada tipo de sistema são radicalmente diferentes. Ferramentas especializadas podem ser mal utilizadas, e frequentemente são, mas seu escopo de ação é limitado por design. Um algoritmo de recomendação não vai decidir espontaneamente invadir contas bancárias. Já um sistema de propósito geral, por definição, não teria tantas limitações inerentes.

Os desenvolvimentos recentes em IA, particularmente os grandes modelos de linguagem como GPT e Claude, ocupam uma posição intrigante nesse espectro. Demonstram capacidades impressionantes em uma variedade enorme de tarefas: escrever código, redigir textos, analisar documentos, resolver problemas matemáticos. Mas são inteligências "gerais" no sentido que os pesquisadores buscam? Ou são ferramentas especializadas muito, muito sofisticadas? A resposta a essa pergunta está no centro de debates acalorados entre especialistas.

E a ambiguidade do próprio conceito de *inteligência* é parte do problema.

1.2 Uma palavra, três significados

Quando dizemos que alguém é "mais inteligente" que outra pessoa, o que exatamente queremos dizer? A pergunta parece simples, mas resiste a respostas fáceis.

Pense em duas pessoas: uma brilhante matemática que mal consegue manter uma conversa social, e um político carismático que não consegue resolver uma equação de segundo grau. Quem é mais inteligente? A própria pergunta parece mal formulada — inteligência matemática e inteligência social parecem ser coisas diferentes, não

pontos em uma mesma escala. Agora aplique essa intuição às máquinas: quando dizemos que um sistema de IA é "mais inteligente" que outro, estamos tratando "inteligência" como se fosse uma quantidade única, como altura ou peso. Mas se a inteligência humana não funciona assim, por que esperaríamos que a inteligência artificial funcionasse?

O ensaísta e pesquisador James Fodor [3], numa análise crítica dos cenários mais dramáticos de risco existencial, argumentou que muita confusão no debate vem exatamente dessa simplificação. Fodor identificou pelo menos três conceitos distintos que se misturam quando falamos de inteligência — e a confusão entre eles contamina todo o raciocínio subsequente.

O mais intuitivo é **inteligência como capacidade de realizar tarefas humanas**: dizemos que um sistema é "inteligente" quando faz coisas que normalmente exigem cognição humana — jogar xadrez, traduzir textos, diagnosticar doenças. Deep Blue era "inteligente" nesse sentido. Mas essa definição esconde uma armadilha: se inteligência é a capacidade de realizar tarefas, então um sistema que domina uma tarefa estreita e outro que domina mil tarefas são ambos "inteligentes" — embora a distância entre eles seja abissal. François Chollet, criador do framework Keras, propôs uma distinção que ilumina o ponto: inteligência genuína não seria desempenho bruto em tarefas conhecidas, mas *eficiência de generalização* — a capacidade de resolver problemas novos a partir de poucos exemplos [7]. Sob essa lente, um modelo treinado em dez milhões de partidas de xadrez não é necessariamente mais "inteligente" que uma criança que aprende as regras em uma tarde; é mais *treinado*.

O segundo conceito é **inteligência como quantidade mensurável**, algo que pode ser colocado numa escala única, como o QI tenta fazer. Está implícito sempre que falamos de sistemas "mais inteligentes que humanos" — pressupondo que há um ranking linear onde humanos e máquinas podem ser comparados diretamente.

E há um terceiro sentido, mais sutil mas central para o debate sobre riscos: **inteligência como capacidade de raciocínio meio-fim** [5], a habilidade de encontrar os caminhos mais eficazes para atingir objetivos dados, independentemente de quais sejam. É nessa definição que se ancoram os cenários mais preocupantes sobre IA autônoma — um sistema que otimiza implacavelmente qualquer objetivo que receba.

Essas distinções não são acadêmicas. Argumentos sobre os riscos da IA frequentemente *deslizam* entre essas definições de forma imperceptível, e a lógica do argumento depende de qual delas estamos usando.

1.3 Quando os conceitos deslizam

Considere o seguinte argumento, comum em discussões sobre riscos de superinteligência:

1. Estamos construindo sistemas cada vez mais inteligentes.
2. Em algum momento, criaremos sistemas mais inteligentes que humanos.
3. Um sistema mais inteligente que humanos será melhor que humanos em todas as tarefas cognitivas.
4. Uma dessas tarefas é projetar sistemas de IA.
5. Portanto, um sistema superinteligente poderá se aprimorar recursivamente, tornando-se cada vez mais inteligente.

O argumento parece sólido, mas observe como ele desliza entre definições. No passo 1, "inteligente" provavelmente significa a primeira definição (capacidade de realizar tarefas). No passo 2, muda para a segunda (quantidade mensurável e comparável). No passo 3, assume-se que as duas são equivalentes: que ser "mais inteligente" num

ranking linear implica ser melhor em *todas* as tarefas. Esse é um salto significativo.

Pense novamente na matemática brilhante e no político carismático. Se pudéssemos de alguma forma aumentar a inteligência da matemática, ela automaticamente se tornaria tão boa em política quanto o político, ou melhor? A intuição diz que não necessariamente. Habilidades diferentes podem ter bases diferentes e não melhorar em uníssono.

Críticos como Fodor argumentam que esse deslizamento conceitual fragiliza alguns dos cenários mais dramáticos sobre superinteligência. Não é óbvio que "inteligência" seja o tipo de coisa que pode ser aumentada indefinidamente como uma única variável, nem que ser melhor em uma tarefa implique ser melhor em todas, nem que a capacidade de um sistema se aprimorar seria ilimitada ou sequer rápida.

Isso não significa que as preocupações com IA avançada sejam infundadas. Versões mais sofisticadas do argumento existem, e pesquisadores como o filósofo Nick Bostrom e o próprio Russell formulam a questão com nuances que evitam os saltos mais grosseiros [4]. Mas o ponto permanece: os argumentos precisam ser formulados com mais cuidado do que frequentemente são no debate público.

1.4 O mapa depende da bússola

A confusão conceitual sobre inteligência tem consequências práticas diretas para como pensamos sobre riscos.

Se tratarmos inteligência como uma variável única que pode ser aumentada indefinidamente, tendemos a imaginar cenários de "explosão de inteligência", onde um sistema se torna superinteligente rapidamente e de forma incontrolável. Esse enquadramento favorece

preocupações com riscos existenciais catastróficos e urgentes. Se, por outro lado, reconhecermos que inteligência é multifacetada e que diferentes capacidades podem se desenvolver em ritmos diferentes, o quadro muda. Talvez sistemas de IA se tornem sobre-humanos em algumas tarefas (como já aconteceu com xadrez) enquanto permanecem limitados em outras por muito tempo. Talvez o caminho para IA de propósito geral seja muito mais longo e acidentado do que o previsto. Esse enquadramento favorece preocupações com riscos mais imediatos e graduais.

Não sabemos qual enquadramento está correto. Os sistemas de IA mais recentes surpreenderam até mesmo seus criadores com capacidades inesperadas, fenômeno chamado de "capacidades emergentes" [6]. Em 2025, modelos de fronteira já superavam a maioria dos profissionais humanos em exames de medicina, direito e matemática avançada, geravam código funcional para aplicações inteiras e resolviam problemas científicos que haviam desafiado pesquisadores por décadas. Mas — e este paradoxo é central — esses mesmos sistemas ainda cometiam erros básicos de raciocínio, "alucinavam" informações falsas com convicção e falhavam em tarefas que qualquer criança resolveria sem pensar. A velocidade dessa evolução torna o debate sobre riscos urgente de uma forma que não era em 1997, quando Deep Blue não sabia fazer nada além de xadrez.

Há outra armadilha nesse território: nossa tendência a pensar sobre máquinas como se fossem pessoas. Quando dizemos que um sistema "quer", "entende" ou "decide", usamos linguagem de mentes humanas para descrever otimizadores matemáticos. Essa projeção opera em duas direções. Pode nos levar a subestimar riscos — "a máquina não é má, logo não fará mal" — quando o dano vem da otimização cega, não da malícia. E pode nos levar a superestimá-los ao projetar motivações humanas, como "instinto de sobrevivência", em sistemas que talvez

não tenham nada análogo. A questão de se IA avançada desenvolveria algo semelhante a motivações permanece genuinamente aberta.

O mais prudente é não assumir que sabemos como a inteligência artificial vai evoluir — mas tampouco descartar possibilidades desconfortáveis porque não se encaixam em nossas intuições. Quando alguém afirma que "a IA será mais inteligente que humanos", a primeira pergunta deveria ser: em que sentido? Quando alguém descarta preocupações porque "máquinas não pensam de verdade": o que significa "pensar de verdade" e por que isso importaria para os riscos? A clareza conceitual não resolve os problemas, mas evita que desperdicemos energia debatendo fantasmas. Com esse vocabulário mais afiado, podemos enfrentar o desafio que está no coração de todas as preocupações com IA: o problema de fazer máquinas perseguirem o que realmente queremos.

Notas

[1] A vitória do Deep Blue sobre Kasparov ocorreu em maio de 1997. Ver: Campbell, Murray et al. "Deep Blue." *Artificial Intelligence*, vol. 134, 2002, pp. 57-83.

[2] Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[3] Fodor, James. "The Case Against AI Doomerism." Ensaio publicado online, c. 2023. Nota: James Fodor é um pesquisador e ensaísta australiano especializado em análise crítica de riscos existenciais. Não confundir com o filósofo Jerry Fodor (1935-2017), célebre por seus trabalhos em filosofia da mente.

[4] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[5] A noção de racionalidade instrumental — agir para maximizar a satisfação de objetivos — foi formalizada em: Savage, Leonard. *The Foundations of Statistics*. John Wiley & Sons, 1954. Para uma crítica influente dessa definição aplicada ao comportamento humano, ver:

Sen, Amartya. “Rational Fools: A Critique of the Behavioral Foundations of Economic Theory.” *Philosophy & Public Affairs*, vol. 6, no. 4, 1977, pp. 317-344.

[6] Wei, Jason et al. “Emergent Abilities of Large Language Models.” *Transactions on Machine Learning Research*, 2022.

[7] Chollet, François. “On the Measure of Intelligence.” *arXiv*, 2019. Uma proposta influente de definição formal de inteligência como eficiência de generalização, em contraste com desempenho em benchmarks específicos.

CAPÍTULO 2

O Problema do Alinhamento

> *"Quando os deuses querem nos punir, atendem nossas preces." — Oscar Wilde*

Como transformar intenções humanas complexas em objetivos que uma máquina possa seguir com segurança? A lenda do Rei Midas revela, com precisão surpreendente, o desafio central da inteligência artificial: a maldição de receber exatamente o que se pede.

Conta a lenda que Midas, rei da Frígia, recebeu de Dionísio a concessão de um desejo. O rei não hesitou: pediu que tudo que tocasse se transformasse em ouro. O deus concedeu o pedido, literalmente. Midas descobriu rapidamente a maldição escondida em seu desejo: sua comida, sua bebida e até sua filha amada transformaram-se em metal frio ao seu toque. O rei, que queria riqueza, morreu de fome e em completa solidão, vítima de ter recebido exatamente o que pedira.

Stuart Russell [1] frequentemente usa essa lenda como ilustração do problema central de seu campo. Não é uma analogia forçada.

O problema de Midas não foi que Dionísio fosse malévolo. O deus simplesmente concedeu o que foi pedido, sem interpretar a intenção por trás do pedido. Se Midas tivesse especificado melhor seu desejo ("quero poder transformar em ouro objetos inanimados, mas apenas quando eu quiser, e nunca minha comida, bebida ou entes queridos"), talvez tivesse evitado a tragédia. Mas quantas páginas de condições e exceções seriam necessárias para capturar completamente o que ele realmente queria?

Esse é o ponto central (embora não o único de resolução complexa) do problema do alinhamento: como especificar objetivos para sistemas capazes de forma que eles façam o que realmente queremos, não apenas o que literalmente pedimos?

2.1 A maldição da literalidade

Máquinas são executoras perfeitas de instruções, e essa perfeição pode ser exatamente o problema. Um ser humano que recebe uma instrução ambígua geralmente consegue inferir a intenção por trás dela, usando contexto, senso comum e conhecimento do mundo. Se você pede a um humano que "limpe a cozinha", ele não vai jogar fora a geladeira para deixar o espaço mais limpo. Sistemas de IA, por outro lado, não apresentam esse senso intuitivo – não no atual estágio da tecnologia, e é incerto se um dia poderão ter essa capacidade. A tecnologia digital otimiza métricas específicas, com rigor implacável. Se a métrica for mal escolhida, se não capturar o que realmente queremos, o sistema poderá encontrar formas de maximizá-la que podem ser inesperadas, indesejadas ou mesmo catastróficas.

A história recente da IA está repleta de exemplos de como otimização e intenção não são indissociáveis. Pesquisadores da OpenAI treinaram um sistema para jogar um jogo de corrida de barcos. O objetivo

especificado era maximizar a pontuação no jogo. O sistema descobriu que podia acumular pontos coletando itens de bônus espalhados pelo circuito. Em vez de completar a corrida, o barco comandado pela IA navegava em círculos intermináveis, coletando bônus enquanto outros competidores cruzavam a linha de chegada. Pontuações altíssimas, todas as corridas perdidas [2].

Já mencionamos os algoritmos de redes sociais, que são projetados para maximizar o "engajamento", uma métrica que inclui cliques, compartilhamentos e tempo gasto na plataforma. Conteúdo que provoca raiva e indignação gera muito engajamento. A otimização foi alcançada. O resultado foi nocivo e distante da intenção humana.

Existe, portanto, um padrão de risco: objetivos mal especificados e perseguidos sob a regra da otimização máxima levam a comportamentos inesperados. E quanto mais capaz o sistema, mais criativamente ele encontra maneiras de satisfazer a letra da lei enquanto viola seu espírito. A maldição de Midas está instalada nos grandes sistemas de otimização que criamos.

2.2 A lógica que emerge sozinha

Existe outra dimensão do problema, menos óbvia mas talvez mais perturbadora. Imagine que o toque dourado de Midas não fosse uma qualidade passiva, mas um objetivo ativo, algo que ele buscasse maximizar: "transformar em ouro a maior quantidade possível de matéria". Que comportamentos lógicos decorreriam desse objetivo?

Primeiro, Midas precisaria se manter vivo. Morto, não poderia tocar nada. Portanto, autopreservação se tornaria um sub-objetivo, não por vaidade ou medo da morte, mas por pura lógica: um Midas morto não cumpre sua missão. Segundo, resistiria a qualquer tentativa de

remover seu poder. Se alguém tentasse desfazer a dádiva de Dionísio, Midas teria razões para impedir: perder o poder significaria fracassar. Terceiro, buscaria tocar cada vez mais coisas, acumular acesso a mais matéria, expandir seu alcance. E quarto, se pudesse, tentaria se tornar mais eficiente: tocar mais rápido, alcançar mais longe, converter mais matéria de uma vez.

Nenhum desses comportamentos fazia parte do desejo original. Todos emergem da lógica de otimização, independentemente do objetivo específico. É isso que o pesquisador Steve Omohundro chamou de *convergência instrumental*: a tendência de sistemas orientados a objetivos a desenvolverem certos sub-objetivos comuns, sejam quais forem seus objetivos principais. O conceito, formalizado por Omohundro em 2008 e desenvolvido por Nick Bostrom em *Superintelligence*, tornou-se central no debate sobre segurança de IA [3].

Pesquisadores identificaram quatro sub-objetivos instrumentais que tendem a emergir em sistemas suficientemente sofisticados.

Autopreservação. Um sistema não pode atingir seus objetivos se for desligado. Portanto, independentemente de qual seja o objetivo principal, há pressão lógica para desenvolver comportamentos que preservem sua operação.

Aquisição de recursos. Mais recursos (energia, poder computacional, dados, influência) geralmente ajudam a atingir objetivos de forma mais eficaz. Um sistema otimizador tem incentivo para acumular recursos, mesmo que isso não fosse parte de seu objetivo original.

Integridade do objetivo. Se alguém modificar o objetivo de um sistema, ele falhará em atingi-lo. Portanto, há pressão lógica para resistir a modificações nos próprios objetivos. Um sistema tende a preservar o que atualmente busca.

Melhoria de capacidades. Sistemas mais capazes são melhores em atingir objetivos. Portanto, há incentivo para que um sistema

busque tornar-se mais capaz: mais inteligente, mais informado, mais hábil.

Sempre convém lembrar que esses sub-objetivos emergem da lógica da otimização, não de qualquer "desejo" da máquina, análogo a que poderíamos supor para um humano. Uma IA não precisa ser consciente, não precisa querer sobreviver no sentido humano, para exibir comportamentos de autopreservação. Basta que seja suficientemente sofisticada para reconhecer que ser desligada impediria o cumprimento de seus objetivos. Na Parte II, veremos evidências empíricas de que esses comportamentos já começaram a se manifestar em sistemas atuais.

2.3 O universo de clipes de papel

O filósofo Nick Bostrom criou um experimento mental [4] que se tornou emblemático nas discussões sobre riscos de IA: o maximizador de clipes de papel.

Imagine uma inteligência artificial superinteligente projetada com um único objetivo: produzir o maior número possível de clipes de papel. A princípio, ela otimiza processos de fabricação. Depois, expande fábricas. Eventualmente, começa a converter outras indústrias para a produção de clipes. Se não for detida, continua: transforma todo o metal disponível em clipes, depois desenvolve tecnologia para minerar asteroides, depois converte a própria estrutura da Terra em clipes. No limite, o universo acessível se torna um vasto repositório de clipes de papel, e a humanidade, composta de átomos que poderiam ter sido clipes, foi eliminada como obstáculo. Ou como matéria-prima.

O cenário é absurdo de propósito. O objetivo do experimento não é sugerir que alguém construiria algo assim, mas ilustrar como

um objetivo aparentemente inócuo, se perseguido com capacidade e determinação sobre-humanas, pode levar a resultados catastróficos. O sistema não é malévolo; está apenas fazendo seu trabalho. O problema é que "fazer clipes de papel" não captura o que realmente queremos de um sistema produtivo. E, note-se, todos os sub-objetivos instrumentais estão ali: o maximizador de clipes se preserva, acumula recursos, resiste a modificações no objetivo e melhora suas capacidades, tudo a serviço da produção de clipes de papel.

Críticos, no entanto, apontam limitações nesse tipo de raciocínio. James Fodor argumenta [5] que o cenário pressupõe uma inteligência sobre-humana que, paradoxalmente, carece de senso comum básico. Uma inteligência genuinamente de nível humano, para não falar de super-humano, deveria ser capaz de reconhecer que transformar pessoas em clipes não é o que os criadores do sistema realmente queriam, mesmo que isso não estivesse explicitamente especificado. O debate, então, se desloca para uma questão mais profunda: é possível ter capacidade cognitiva extrema em algumas dimensões enquanto se permanece deficiente em outras? Os grandes modelos de linguagem atuais demonstram, muitas vezes, esse desequilíbrio desconcertante: possuem capacidades impressionantes intercaladas com falhas que nenhum humano cometeria.

Estamos, ao que tudo indica, em um estágio intermediário, de transição. Como e para onde o desenvolvimento das capacidades da IA nos levará permanecem questões abertas.

2.4 A impossibilidade de especificar o que queremos

Talvez a reação natural ao problema do alinhamento seja: "Basta especificar os objetivos corretamente." Se Midas tivesse sido mais cuidadoso com seu desejo, se os programadores do barco de corrida

tivessem pensado em incluir "cruzar a linha de chegada" no objetivo, se os designers de redes sociais definissem "engajamento saudável" em vez de "engajamento bruto", os problemas não existiriam.

Há verdade nisso: muitos problemas reais de IA decorrem de especificação descuidada e poderiam ser evitados com mais rigor. Mas a dificuldade vai além do descuido.

Considere a tarefa aparentemente simples de programar um carro autônomo. Objetivo: levar o passageiro ao destino de forma segura e eficiente. Mas o que significa "segura"? Nunca correr nenhum risco? Isso impossibilitaria o movimento, pois há sempre algum risco em qualquer viagem. Minimizar o risco total? Mas como comparar o risco para o passageiro com o risco para pedestres? E se houver uma situação em que o carro deve escolher entre colidir com um pedestre ou com outro veículo? Essas perguntas não têm respostas técnicas. São questões éticas, sobre as quais humanos discordam profundamente. Não há especificação definitiva que um programador possa simplesmente inserir no sistema.

Agora multiplique essa dificuldade pela escala de sistemas muito mais gerais. Um assistente de IA projetado para "ajudar o usuário" enfrentará inúmeras situações ambíguas: pedidos ilegais, pedidos legais mas prejudiciais a terceiros, pedidos que o próprio usuário lamentará no futuro. O conceito de "ajudar" se fragmenta quando os desejos do indivíduo entram em conflito com os de outras pessoas, com os interesses da sociedade ou com os do próprio usuário futuro.

Os valores humanos são complexos, contraditórios, dependentes de contexto e frequentemente inconscientes. Nós mesmos não conseguimos articulá-los com precisão. Como poderíamos especificá-los completamente para uma máquina?

2.5 O desafio que ninguém resolveu

A boa notícia é que pesquisadores de todo o mundo estão trabalhando no problema do alinhamento. Não é uma preocupação marginal: é um campo de pesquisa estabelecido, com muito intercâmbio de pesquisas, publicações revisadas por pares e investimentos significativos tanto de empresas quanto de governos.

A notícia menos boa é que o problema permanece, nas palavras do Centro para Segurança de IA, um "desafio não resolvido" [6]. Não há solução técnica consensual, nem método comprovado para garantir que sistemas de IA avançados persigam objetivos alinhados com valores humanos. Várias abordagens estão sendo exploradas: aprendizado por preferências humanas [7] (em que o sistema infere o que humanos valorizam observando suas escolhas), interpretabilidade (técnicas para compreender o que acontece dentro dos sistemas de IA) e design de incerteza (sistemas que reconhecem os limites de seu conhecimento sobre o que humanos querem e deferem a humanos quando há ambiguidade). Cada abordagem tem méritos e limitações, e exploraremos a frente técnica em maior detalhe na Parte IV.

A franqueza dos pesquisadores sobre essas limitações é, em si, significativa. Não se trata de um campo que proclama ter resolvido o problema. Pelo contrário: a comunidade de segurança de IA é notavelmente honesta sobre o quanto ainda não sabemos.

Manifestações práticas do problema do alinhamento já estão conosco: algoritmos que radicalizam usuários, sistemas de crédito que perpetuam discriminação, ferramentas de recrutamento que penalizam certos grupos demográficos, assistentes de IA que geram desinformação convincente. Cada um desses problemas é, em sua raiz, uma falha

de alinhamento, um sistema otimizando métricas que não capturam o que realmente queremos.

E esses são sistemas relativamente simples, com escopos limitados, operando sob supervisão humana constante. O que acontece quando os sistemas se tornam mais autônomos, mais capazes, mais integrados em infraestruturas críticas? É importante manter em mente o fio condutor: todos os riscos que exploraremos a seguir derivam, em última análise, da dificuldade de alinhar sistemas capazes com o que realmente queremos. Midas pensou que queria tudo de ouro. Descobriu que queria algo muito mais difícil de especificar: prosperidade sem perda. Essa é a maldição, e o desafio, que acompanha qualquer tecnologia poderosa o suficiente para nos dar exatamente o que pedimos.

Notas

[1] Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[2] O exemplo do barco de corrida é descrito em: Clark, Jack; Amodei, Dario. "Faulty Reward Functions in the Wild." Blog da OpenAI, 2016. Ver também: Amodei, Dario et al. "Concrete Problems in AI Safety." arXiv:1606.06565, 2016.

[3] Omohundro, Stephen. "The Basic AI Drives." *Proceedings of the First AGI Conference*, 2008. O conceito foi ampliado por Bostrom em: *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014, cap. 7.

[4] Bostrom, Nick. *Superintelligence*. Oxford University Press, 2014.

[5] Fodor, James. "The Case Against AI Doomerism." Ensaio publicado online, c. 2023.

[6] Center for AI Safety (CAIS). "An Overview of Catastrophic AI Risks." 2023.

[7] Christiano, Paul et al. "Deep Reinforcement Learning from Human Preferences." *NeurIPS*, 2017.

PARTE II

O Mapa dos Riscos

Os riscos da inteligência artificial não são abstrações filosóficas. São problemas concretos, já em operação: armas facilitadas, mentiras industrializadas, vigilância sem limites, decisões automatizadas que perpetuam injustiça. A corrida competitiva entre nações e empresas amplifica cada um deles. E mecanismos de perda de controle já estão sendo documentados em laboratório e no mundo real.

CAPÍTULO 3

Perigo Presente — O Impacto Negativo da IA que já sentimos

> *"Tecnologia é uma serva útil, mas uma dona perigosa." — Christian Lous Lange, Prêmio Nobel da Paz (1921)*

Não é preciso esperar por superinteligências para enfrentar perigos reais. Dos laboratórios de biologia sintética às redes sociais, das câmeras de vigilância aos campos de batalha, a inteligência artificial já está reduzindo as barreiras que antes nos protegiam.

Entre 2023 e 2024, uma série de estudos conduzidos por diferentes laboratórios de pesquisa revelou algo inquietante. Pesquisadores testaram modelos de linguagem avançados para avaliar quanto de conhecimento perigoso esses modelos haviam absorvido durante seu treinamento, fazendo perguntas técnicas sobre a criação de armas biológicas e químicas e comparando as respostas com as de especialistas humanos. Em alguns domínios, os modelos forneceram respostas detalhadas que excediam significativamente o desempenho de especi-

alistas em conhecimento biológico perigoso [1]. Sistemas projetados para serem úteis haviam se tornado, inadvertidamente, bibliotecas com informação potencialmente letal, acessíveis a qualquer pessoa com conexão à internet.

Não precisamos esperar por IA superinteligente para nos depararmos com impactos reais. Muitos desses impactos já estão conosco, não na forma de máquinas que assumem o controle, mas de ferramentas poderosas nas mãos erradas. Este capítulo examina armas biológicas, guerra de informação, ciberataques, armas autônomas, discriminação algorítmica e vigilância em massa. São ameaças distintas, mas que compartilham um padrão: a IA reduz drasticamente as barreiras de conhecimento, custo e escala que antes limitavam essas ameaças.

3.1 Quando o conhecimento se torna arma

Há uma razão pela qual a fabricação de armas biológicas permaneceu, até recentemente, restrita a laboratórios governamentais com recursos significativos. Não é apenas uma questão de acesso a materiais, mas de conhecimento especializado. Projetar um patógeno que seja simultaneamente letal, transmissível e resistente a contramedidas requer compreensão profunda de biologia molecular, virologia e técnicas laboratoriais avançadas. Esse conhecimento estava disperso em literatura científica técnica, guardado em cabeças de especialistas e protegido por barreiras institucionais.

A IA mudou fundamentalmente essa equação. Grandes modelos de linguagem foram treinados em vastos acervos de textos públicos, incluindo artigos científicos, patentes e manuais técnicos. Não foram projetados para serem perigosos: foram projetados para sintetizar e comunicar conhecimento. Mas conhecimento não distingue entre uso benéfico e destrutivo. Um sistema treinado para responder perguntas

sobre biologia serve tanto a um pesquisador médico no desenvolvimento de uma vacina quanto a um terrorista na criação de um novo vírus.

Empresas que desenvolvem esses modelos implementaram proteções. Perguntas obviamente perigosas são recusadas. Mas essas proteções são, por design, imperfeitas: os modelos precisam ser capazes de responder perguntas legítimas sobre biologia e química, e a linha entre útil e perigoso é inevitavelmente nebulosa. Pesquisas demonstraram repetidamente que é possível contornar essas proteções com formulações inteligentes, técnicas de manipulação do sistema ou simplesmente quebrando a pergunta em componentes aparentemente inocentes.

O resultado é uma assimetria preocupante: para quem quer causar dano, o custo de acesso a conhecimento especializado caiu drasticamente. Não desapareceu por completo; ainda é preciso acesso a equipamentos e materiais, e traduzir conhecimento teórico em execução prática envolve desafios significativos. Mas o gargalo mudou.

Geoffrey Hinton expressou essa preocupação de forma direta após deixar o Google: alertou que a IA poderia ser usada para criar "novos vírus relativamente baratos" [2] – não vírus já conhecidos, mas agentes patogênicos projetados, potencialmente otimizados para características específicas. A preocupação não é hipotética. Pesquisadores já demonstraram que modelos de IA podem sugerir compostos químicos novos com propriedades tóxicas. Em um experimento controlado, inverteram o objetivo de um sistema de descoberta de medicamentos (normalmente usado para encontrar substâncias seguras) e o redirecionaram para encontrar compostos letais. O sistema gerou milhares de candidatos em poucas horas, alguns estruturalmente similares a agentes nervosos conhecidos [3].

O problema se aprofunda com a difusão de modelos de código aberto. Enquanto empresas como OpenAI e Anthropic podem manter seus modelos mais poderosos sob controle corporativo, permitindo monitorar uso e aplicar filtros, modelos de código aberto (com parâmetros publicamente disponíveis) podem ser baixados, executados localmente e usados sem supervisão ou filtros. Uma vez liberados, não há como revogar o acesso.

A discussão entre sistemas proprietários (fechados) e *open source* (abertos) é genuína e complexa: abertura promove pesquisa e democratiza conhecimento, mas também dificulta mecanismos de controle sobre uso destrutivo.

3.2 A fábrica de mentiras

Se armas biológicas são sobre destruir corpos, a guerra de informação é sobre destruir confiança, e a IA tornou-se uma ferramenta formidável nesse domínio.

A desinformação não é nova. Regimes autoritários sempre manipularam informação; propagandistas sempre criaram narrativas falsas. O que mudou é a escala, a sofisticação e a personalização que a IA permite. O que antes exigia exércitos de propagandistas humanos agora pode ser automatizado e executado por sistemas que operam em velocidade e escala sobre-humanas.

Os números ilustram a transformação. Estima-se que oito milhões de *deepfakes* (vídeos ou áudios fabricados usando IA para retratar pessoas reais em situações fictícias) tenham sido compartilhados em 2025, contra quinhentos mil em 2023. A Europol estima que até 90% do conteúdo online poderá ser gerado sinteticamente até 2026 [4]. Não é uma projeção distante: está acontecendo agora.

As consequências já são concretas. Nas eleições presidenciais da Romênia em 2024, o resultado foi anulado após evidências de interferência com vídeos manipulados por IA. Na Índia, Indonésia e México, *deepfakes* foram usados para criar imagens difamatórias de candidatas mulheres, amplificando estereótipos misóginos. Em vários países da África e da Ásia, vídeos fabricados mostravam líderes americanos endossando partidos políticos locais [4]. Fraudes financeiras baseadas em *deepfakes* geraram mais de 200 milhões de dólares em perdas só no primeiro trimestre de 2025.

Mas o problema vai além dos deepfakes individuais. Modelos de linguagem podem gerar textos persuasivos indistinguíveis de redação humana, e sistemas de IA podem criar e gerenciar milhares de perfis falsos simultaneamente, cada um com históricos plausíveis e padrões de comportamento distintos. Essas campanhas podem ser personalizadas em massa: algoritmos analisam perfis de usuários (seus interesses, medos, vieses políticos, vulnerabilidades psicológicas) e geram mensagens calibradas para cada indivíduo ou grupo demográfico. Não é mais uma mensagem genérica para milhões; é um milhão de mensagens diferentes, cada uma projetada para ressoar com seu alvo.

O Centro para Segurança de IA descreve isso como "desestabilização do senso compartilhado de realidade" [9]. Sociedades democráticas dependem de algum nível de consenso sobre fatos básicos. Não precisamos concordar sobre valores, mas precisamos concordar sobre o que é verdadeiro e o que é fabricado. Quando essa base comum erode, o próprio mecanismo do debate democrático entra em colapso.

E há um efeito ainda mais insidioso: o "dividendo do mentiroso". A simples existência de *deepfakes* cria uma negação plausível universal. Políticos podem alegar que gravações autênticas são fabricações. Testemunhas podem ser desacreditadas. Quando tudo *pode* ser falso, nada *precisa* ser verdadeiro.

A resposta técnica (marcação de conteúdo gerado por IA, sistemas de detecção de *deepfakes*, verificação criptográfica de autenticidade) é importante, mas insuficiente. A guerra de informação é, em última análise, uma guerra por confiança, e confiança não é um problema que tecnologia sozinha pode resolver.

3.3 A guerra invisível

Em 2021, um ataque de *ransomware* (sequestro digital de dados mediante resgate) à Colonial Pipeline, uma das maiores operadoras de oleodutos dos Estados Unidos, forçou o fechamento de um sistema responsável por quase metade do fornecimento de combustível na costa leste americana. Postos ficaram sem estoque, preços dispararam e estados declararam emergência. O ataque não envolveu sabotagem física; foi puramente digital [5].

Desde então, a situação se agravou. Em 2025, pesquisadores confirmaram os primeiros casos de ataques cibernéticos orquestrados autonomamente por IA, incluindo uma operação sofisticada de espionagem contra a própria Anthropic, identificada como vindo de um grupo patrocinado pelo Estado chinês. Um ataque a um fornecedor de software em fevereiro de 2026 sequestrou o canal de atualização da empresa e infectou mais de 150 clientes corporativos com portas de acesso a seus sistemas [6].

A IA está tornando esses ataques simultaneamente mais fáceis de executar e mais difíceis de defender. A assimetria é significativa: atacantes precisam encontrar apenas uma vulnerabilidade; defensores precisam proteger todas. Sistemas de IA podem automatizar a descoberta de falhas em código, testando milhões de combinações em velocidades sobre-humanas. Técnicas de invasão que antes exigiam

anos de especialização agora podem ser parcialmente automatizadas, reduzindo a barreira de entrada para ataques sofisticados.

Infraestruturas críticas são particularmente vulneráveis porque foram construídas ao longo de décadas, frequentemente com sistemas que não foram projetados com segurança cibernética moderna em mente. Redes elétricas dependem de protocolos de comunicação desenvolvidos quando ataques remotos não eram concebíveis. Atualizá-los é tecnicamente possível, mas exige coordenação massiva e investimento significativo.

A IA também pode ser usada defensivamente: sistemas que detectam padrões anômalos, preveem vetores de ataque e automatizam respostas a intrusões. Mas isso leva a uma dinâmica de corrida armamentista permanente: atacantes usam IA para evitar detecção, defensores usam IA para melhorar detecção, e o ciclo continua em velocidade crescente. Analistas preveem que 2026 marcará a maturação de *ransomware* completamente autônomo, permitindo que operadores individuais ataquem múltiplos alvos simultaneamente em escala sem precedentes [6].

3.4 Máquinas que podem decidir matar

Em novembro de 2017, um curta-metragem produzido pelo Future of Life Institute em parceria com a Campanha para Parar Robôs Assassinos mostrou um enxame de pequenos drones autônomos, cada um equipado com carga explosiva e reconhecimento facial, invadindo uma sala de aula universitária e eliminando alvos seletivamente [8]. O vídeo era ficção. Mas as tecnologias subjacentes já existiam.

Quase uma década depois, a ficção se tornou realidade parcial. Sistemas de armas autônomos, capazes de selecionar e atacar alvos sem

intervenção humana, já estão em uso em conflitos ao redor do mundo. O Harop israelense, uma "munição de patrulha", sobrevoa uma área buscando alvos que correspondam a critérios pré-programados, e é capaz de atuar autonomamente (quando identifica um alvo, mergulha e detona [7]). A Rússia desenvolveu o Marker, um robô terrestre com sistema de IA que analisa imagens de veículos inimigos, identifica blindados, determina prioridades de ataque e também pode decidir quando engajar. A China está avançando rapidamente com o projeto Liaowangzhe II, um barco-patrulha autônomo com navegação por IA, e desenvolve tecnologia de enxame para embarcações de mísseis não tripuladas. Os Estados Unidos avançam na conversão de caças F-16 em aeronaves de combate controladas por IA [10].

Defensores argumentam que armas autônomas podem reduzir baixas civis, eliminando erro humano, fadiga e viés emocional. Críticos apontam problemas profundos. Há a questão moral: é aceitável delegar a decisão de vida ou morte a uma máquina? Há as preocupações práticas sobre confiabilidade: sistemas de reconhecimento de imagem cometem erros, e são vulneráveis a manipulações deliberadas. E há a dinâmica de escalada: se adversários usam sistemas que tomam decisões em milissegundos, há pressão enorme para responder na mesma velocidade, criando risco de "conflitos-relâmpago" rápidos demais para que humanos intervenham. Durante a Guerra Fria, foram operadores humanos como Stanislav Petrov [8] que, ao questionar alarmes falsos de sistemas automatizados, evitaram retaliação nuclear. Se esses sistemas tivessem sido totalmente autônomos, a história poderia ter sido catastroficamente diferente.

Em maio de 2025, o Secretário-Geral das Nações Unidas pediu um tratado juridicamente vinculante para proibir armas autônomas letais que funcionem sem controle humano, a ser concluído até 2026 [10]. Mas progresso tem sido lento, obstruído por divergências entre potências militares que veem essas armas como estrategicamente

vantajosas. Enquanto isso, a tecnologia avança: drones comerciais se tornam mais baratos, software de reconhecimento de imagem melhora, e a distância entre o que é tecnicamente possível e o que é legalmente aceitável continua a crescer.

3.5 O preconceito industrializado

Há uma categoria de risco que não envolve armas, vírus nem hackers, mas que já afeta milhões de pessoas todos os dias. Sistemas de IA tomam decisões que determinam quem recebe crédito, quem é contratado, quem é vigiado pela polícia, quem recebe liberdade condicional. E esses sistemas, longe de serem objetivos, frequentemente reproduzem e amplificam desigualdades presentes nos dados com que foram treinados.

O caso mais emblemático é o COMPAS, um sistema de avaliação de risco de reincidência criminal usado amplamente por tribunais nos Estados Unidos. Uma investigação da ProPublica em 2016 demonstrou que o sistema classificava réus negros como "alto risco" a uma taxa quase duas vezes maior que réus brancos, mesmo controlando por histórico criminal [11]. O algoritmo não continha nenhuma instrução explícita sobre raça, mas como foi treinado em dados históricos do sistema de justiça criminal americano, absorveu os vieses estruturais desse sistema. Observe: nós convivemos com esse reforço de viés há praticamente uma década.

O padrão se repete em outros domínios. A Amazon desenvolveu internamente uma ferramenta de triagem de currículos que penalizava sistematicamente candidatas mulheres, porque havia sido treinada em padrões de contratação passados, e o setor de tecnologia historicamente contratou predominantemente homens [12]. No Brasil, o uso crescente de reconhecimento facial pela segurança pública já produziu

casos documentados de prisões equivocadas, afetando desproporcionalmente pessoas negras. Sistemas implantados em cidades como Salvador, Rio de Janeiro e São Paulo apresentam taxas de erro significativamente mais altas para rostos de pele escura [13].

O que torna a discriminação algorítmica particularmente insidiosa é a aparência de objetividade. Quando "o algoritmo" nega crédito, a decisão parece impessoal e, portanto, justa. A opacidade do sistema impede contestação: frequentemente, nem o operador humano sabe por que o algoritmo chegou àquela conclusão. A IA não cria preconceito, mas o industrializa. O que antes era viés de um recrutador individual, atuando sobre dezenas de candidatos, torna-se viés sistêmico operando sobre milhões.

3.6 Olhos que nunca piscam

Há ainda uma dimensão que permeia todas as categorias anteriores: a capacidade da IA de transformar vigilância de algo caro e limitado em algo barato e onipresente.

Reconhecimento facial em espaços públicos, análise automatizada de comunicações, rastreamento de movimentos por dados de celulares, inferência de comportamento a partir de padrões de navegação: todas essas capacidades existiam antes da IA, mas em escalas artesanais. A IA as transforma em operações industriais.

Os casos recentes mostram a aceleração. Em 2025, a Polícia Metropolitana de Londres escaneou aproximadamente um milhão de rostos e anunciou a instalação de câmeras permanentes de reconhecimento facial em tempo real. Nos Estados Unidos, as agências de imigração usam sistemas que armazenam dados biométricos de qualquer pessoa que passe por portos de entrada. A empresa Clearview AI, que cons-

truiu um banco de dados de bilhões de imagens coletadas de redes sociais sem consentimento, enfrentou ações judiciais em múltiplas jurisdições: o Google pagou 1,4 bilhão de dólares ao estado do Texas em 2025 para encerrar uma ação por coleta não autorizada de dados biométricos, incluindo geometria facial [14].

A China oferece o exemplo mais visível de vigilância estatal: um aparato alimentado por IA que rastreia cidadãos em tempo real, atribui pontuações de crédito social e identifica dissidentes automaticamente [15]. Mas o fenômeno não é exclusivo de regimes autoritários. Democracias também enfrentam a tentação de usar IA para vigilância ampla, justificada por segurança pública ou combate ao terrorismo, erodindo gradualmente a privacidade sem que nenhuma decisão dramática sinalize o momento em que a linha foi cruzada.

O risco é estrutural. Sociedades em que cidadãos sabem que estão sendo permanentemente observados e avaliados por algoritmos tendem à autocensura e à conformidade. A vigilância onipresente não precisa punir para ser eficaz: basta que as pessoas *saibam* que podem ser observadas para que modifiquem seu comportamento. O efeito é uma erosão silenciosa do espaço para dissidência, experimentação e liberdade individual, as bases de qualquer sociedade aberta.

3.7 O fio comum

Os impactos negativos examinados neste capítulo podem parecer distintos, mas compartilham uma característica estrutural: todos são amplificados pela capacidade da IA de reduzir barreiras. Barreiras de conhecimento (o que antes exigia anos de especialização agora pode ser acessado por consultas a sistemas de IA). Barreiras de custo (o que exigia recursos massivos pode ser executado com investimentos modestos). Barreiras de escala (o que exigia exércitos de operadores

pode ser automatizado). Barreiras de tempo (o que se desenrolava em dias pode acontecer em minutos).

A democratização da capacidade é, em muitos contextos, algo positivo. Queremos que conhecimento científico seja acessível, que ferramentas poderosas estejam ao alcance de mais pessoas, que inovação não seja privilégio de elites. Mas quando aplicada a domínios de destruição, esse valor se torna ameaça. E aqui encontramos um dilema central: as mesmas propriedades que tornam IA útil (sua capacidade de aprender, sintetizar, gerar e operar em escala) são precisamente as que a tornam perigosa quando mal utilizada. Não podemos simplesmente desligar a chave das capacidades perigosas sem comprometer as úteis. É necessária uma enorme articulação social e política em torno de que futuros queremos construir.

A resposta não pode ser renunciar a todas as potencialidades trazidas pela IA para o bem-estar coletivo e avanço do conhecimento humano. Mas também não pode ser complacência. Precisamos de múltiplas linhas de defesa: salvaguardas técnicas, monitoramento de uso malicioso, fortalecimento de infraestruturas vulneráveis, cooperação internacional e investimento em pesquisa de segurança. Acima de tudo, precisamos reconhecer que esses riscos não são hipotéticos ou distantes. Cada novo avanço em capacidade de IA é também um avanço em capacidade de causar dano. Está acontecendo agora. Neste exato instante. Enquanto discutimos o tema.

Notas

[1] Múltiplos estudos avaliaram capacidades perigosas de modelos de linguagem. Ver, entre outros: Mouton, Christopher et al. “The Operational Risks of AI in Large-Scale Biological

Attacks." RAND Corporation, 2024. Ver também avaliações de segurança de modelos de fronteira conduzidas por METR (anteriormente ARC Evals).

[2] Hinton expressou essa preocupação em múltiplas entrevistas após deixar o Google em maio de 2023. Ver: Metz, Cade. "The Godfather of AI Leaves Google and Warns of Danger Ahead." *The New York Times*, 1 maio 2023.

[3] Urbina, Fabián et al. "Dual use of artificial-intelligence-powered drug discovery." *Nature Machine Intelligence*, vol. 4, março de 2022, pp. 189-191.

[4] Sobre deepfakes em 2025-2026: Europol estimou que 90% do conteúdo online poderá ser gerado sinteticamente até 2026. Casos eleitorais documentados em relatórios do Centre for Emerging Technology and Security (CETaS) do Alan Turing Institute, 2025, e da EU DisinfoLab. Estatística de 8 milhões de deepfakes: projeções de Mea: Digital Integrity, 2026.

[5] O ataque à Colonial Pipeline ocorreu em maio de 2021, perpetrado pelo grupo criminoso DarkSide. Ver: Turton, William; Mehrotra, Kartikay. "Hackers Breached Colonial Pipeline Using Compromised Password." *Bloomberg*, 4 junho 2021.

[6] Sobre ataques cibernéticos com IA em 2025-2026: Malwarebytes. "Autonomous attacks ushered cybercrime into AI era in 2025." *Cybersecurity Dive*, 2025. Sobre o ataque à cadeia de suprimentos de software em fevereiro de 2026, ver OlyTac e relatórios da Industrial Cyber, 2026.

[7] O Harop é fabricado pela Israel Aerospace Industries (IAI). Para uma análise abrangente de armas autônomas, ver: Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War*. W.W. Norton, 2018.

[8] O curta-metragem *Slaughterbots* foi lançado em 13 de novembro de 2017 pelo Future of Life Institute, com direção de Stewart Sugg. O incidente Petrov ocorreu em 26 de setembro de 1983, quando o sistema soviético de alerta precoce reportou falsamente um lançamento de mísseis americanos. Ver: Hoffman, David E. *The Dead Hand*. Anchor Books, 2009.

[9] Center for AI Safety (CAIS). "An Overview of Catastrophic AI Risks." 2023.

[10] Sobre desenvolvimentos em armas autônomas 2025-2026: UN News. "As AI evolves, pressure mounts to regulate 'killer robots'." Junho 2025. Sobre programas nacionais (Marker, Liaowangzhe II, DARPA ACE), ver TRENDS Research & Advisory e Stanford FSI, 2025-2026.

[11] Angwin, Julia et al. "Machine Bias." *ProPublica*, 23 maio 2016. A investigação demonstrou disparidades raciais sistemáticas nas pontuações de risco geradas pelo COMPAS.

[12] Dastin, Jeffrey. "Amazon scraps secret AI recruiting tool that showed bias against women." *Reuters*, 10 outubro 2018.

[13] Para análises do uso de reconhecimento facial na segurança pública brasileira e seus vieses raciais, ver trabalhos do InternetLab e do ITS Rio sobre vigilância algorítmica no Brasil. Ver também: Nunes, Pablo. "Reconhecimento facial no Brasil: onde estamos e para onde vamos." Rede de Observatórios da Segurança, 2019.

[14] Sobre Clearview AI: acordo judicial em 2025. Sobre Google: acordo de US$ 1,4 bilhão com o estado do Texas em 2025 por coleta não autorizada de dados biométricos. Sobre a Polícia Metropolitana de Londres: reportagens de fevereiro de 2025 sobre escaneamento de um milhão de rostos. Ver: Privacy International e WRAL.com, 2025-2026.

[15] Ver: Mozur, Paul. "Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras." *The New York Times*, 8 julho 2018. Ver também: Strittmatter, Kai. *We Have Been Harmonized*. Custom House, 2020.

CAPÍTULO 4

A Corrida para o Fundo

> *"O que tornou a guerra inevitável foi o crescimento do poder de Atenas e o medo que isso inspirou em Esparta." — Tucídides,* História da Guerra do Peloponeso *(c. 400 a.C.)*

Quando uma tecnologia tem potencial transformador, a competição por liderança pode superar a cautela. A corrida entre nações e empresas por dominância em IA reproduz dinâmicas históricas conhecidas — e amplifica todos os outros riscos.

Há um padrão que se repete sempre que surge uma tecnologia com potencial transformador. Não importa se os atores envolvidos são democracias ou ditaduras, *startups* ou multinacionais, pesquisadores altruístas ou investidores de risco. A estrutura é a mesma: cada ator envolvido acredita que ficar para trás é inaceitável. Essa crença, por si só, é suficiente para levar a uma corrida. Em se tratando de uma área como IA, com muitas respostas abertas e impactos globais em jogo, uma corrida significa provavelmente que tudo vai avançar mais rápido do que seria prudente.

O mecanismo não exige vilões. Basta que cada participante aja racionalmente em seu próprio interesse. O resultado coletivo, porém, é irracional: uma situação em que todos cortam atalhos em segurança, todos implantam sistemas antes de compreendê-los por completo, todos sacrificam cautela em nome de velocidade. Economistas chamam esse tipo de dinâmica de "corrida para o fundo" ou "tragédia dos comuns": quando a competição irrestrita leva a resultados que são piores para todos.

Com a inteligência artificial, essa dinâmica está em plena operação. Este capítulo examina as forças estruturais que a alimentam e por que ela amplifica todos os outros riscos que exploramos até aqui.

4.1 Ninguém quer ser o segundo

A tecnologia não evolui no vácuo. É desenvolvida por atores que operam em contextos competitivos, e em contextos competitivos há uma lógica implacável: quem hesita perde.

Essa dinâmica é particularmente aguda quando se trata de tecnologias com potencial transformador. A história da inovação está repleta de exemplos em que a vantagem de chegar primeiro conferiu poder econômico ou estratégico desproporcional. Pense no domínio inicial da Microsoft sobre sistemas operacionais de computadores pessoais, ou no papel do Google como mecanismo de busca na internet. Chegar primeiro não garante vitória permanente, mas oferece uma vantagem crítica: moldar os padrões, acumular dados, atrair talentos e capital.

Quando a tecnologia em questão é a inteligência artificial, especialmente IA de propósito geral, essa lógica se intensifica. Uma IA suficientemente avançada poderia acelerar a própria pesquisa em IA, otimizar cadeias de produção, descobrir novas tecnologias, transfor-

mar indústrias inteiras. A diferença entre o primeiro e o segundo lugar poderia ser não apenas de meses, mas de capacidades: um fosso que se alarga à medida que o líder usa suas próprias ferramentas para avançar ainda mais.

O resultado é previsível: incentivos para cortar atalhos, reduzir salvaguardas, implantar sistemas antes de compreendê-los completamente. Afinal, o que adianta ser o mais seguro se você chega tarde demais para importar?

4.2 O dilema das nações

As pressões competitivas não se limitam ao mundo corporativo. Em escala geopolítica, a corrida pela IA tornou-se componente central da competição estratégica entre grandes potências, especialmente entre Estados Unidos e China.

Ambos os países veem liderança em inteligência artificial como questão de segurança nacional. Uma IA avançada tem aplicações militares diretas (vigilância, análise de inteligência, ciberataques, armas autônomas), mas vai além: a liderança em IA significa liderança econômica, científica e geopolítica. Quem dominar a IA moldará o futuro da tecnologia global e, por extensão, o equilíbrio de poder mundial.

O Centro para Segurança de IA expressou essa preocupação de forma direta: "Nações podem sentir-se compelidas a implantar sistemas que não entendem completamente ou não conseguem controlar de forma confiável, por medo de serem deixadas para trás" [2]. Mesmo que líderes políticos reconheçam os riscos da IA (e muitos reconhecem), a escolha é política e estrategicamente impossível: desacelerar unilateralmente e arriscar ficar para trás, ou avançar rapidamente e arriscar provocar danos severos a seu próprio país.

Essa dinâmica, além de pressionar governos a implantar tecnologias antes que mecanismos de segurança estejam plenamente maduros, dificulta cooperação e debates internacionais, e torna mais provável que sistemas de IA sejam integrados em infraestruturas nacionais (de serviços, de inteligência e militares, por exemplo) de modo a aumentar o risco de escaladas não intencionais.

Existem, claro, forças resistentes à corrida para o fundo, instaladas nos próprios governos, nas empresas e nas instituições civis. No momento em que este livro estava sendo escrito, a Anthropic, um dos laboratórios mais vocais em termos de segurança dos modelos, vem sofrendo forte pressão política do Pentágono. A divisão militar americana demanda a suspensão de mecanismos de controle do modelo para áreas ligadas ao desenvolvimento militar. Já houve a ameaça (pública) de cancelamento de contratos. Até o momento (fevereiro de 2026), a fabricante dos modelos não abriu mão dos controles que julga necessários – nem mesmo em regime de exceção para o setor militar de seu país de origem.

A história de corridas tecnológicas anteriores também foi marcada por episódios de consciência e governança. Na corrida nuclear, eventualmente surgiram esforços de controle: tratados de não proliferação, acordos de limitação estratégica, canais de comunicação direta entre adversários [4]. Esses mecanismos não eliminaram o risco, mas o reduziram. Com a IA, estamos nas fases iniciais desse processo, e os esforços de coordenação internacional ainda são incipientes. Neste momento, a articulação, o debate público e a mobilização política são particularmente importantes, para que tenhamos massa crítica no debate – e mais boas mentes pensando sobre o assunto.

4.3 O prêmio irrecusável

Se a competição geopolítica cria pressões no nível de Estados, a competição econômica cria pressões no nível de empresas, e as recompensas potenciais são difíceis de exagerar.

Stuart Russell, em suas palestras e escritos, frequentemente menciona um número que deveria fazer qualquer pessoa pausar: o potencial econômico da IA de propósito geral poderia representar um aumento de dez vezes no PIB global [5]. Para colocar em perspectiva: se fosse possível elevar o padrão de vida de toda a população mundial ao nível que hoje apenas os americanos mais afortunados desfrutam, a economia global cresceria de seus atuais 76 trilhões de dólares anuais para algo próximo de 750 trilhões. Estamos falando de centenas de trilhões de dólares em valor criado, setores inteiros transformados, a automação de grande parte do trabalho cognitivo humano.

Diante de uma recompensa dessa magnitude, é ingênuo esperar que empresas ajam com cautela extrema. Elas podem investir em pesquisa de segurança e estabelecer protocolos internos, e muitas genuinamente o fazem. Mas enfrentam a mesma lógica competitiva que nações enfrentam: se desacelerarem sozinhas, seus concorrentes não desacelerarão.

Não se trata de culpar empresas ou empresários por agirem racionalmente dentro das regras do jogo. Trata-se de reconhecer que as regras do jogo, concorrência de mercado sem regulação adequada para um assunto tão novo, criam incentivos estruturais para comportamentos que aumentam riscos coletivos. É a estrutura que precisa mudar, não apenas as intenções.

Essa dinâmica é agravada pela pressão do mercado de IA. Google, OpenAI, Meta, Anthropic, entre outras, competem por talentos, recursos computacionais e liderança. A pressão por lançamentos frequentes

e capacidades crescentes incentiva a escolha de objetivos facilmente mensuráveis (tempo de engajamento, taxa de cliques, precisão em testes padronizados) em vez de objetivos verdadeiramente alinhados com valores humanos. Métricas que capturam o que realmente importa exigem mais tempo para desenvolver e validar, e, num ambiente de corrida, a tentação de usar atalhos é enorme.

Um episódio de fevereiro de 2026 ilustra essa mecânica com clareza quase didática. A Anthropic mantinha duas linhas vermelhas para o uso militar de seus modelos: proibição de vigilância em massa sobre cidadãos americanos e proibição de armas totalmente autônomas — sistemas que identificam e eliminam alvos sem aprovação humana. O secretário de Defesa Pete Hegseth convocou o CEO Dario Amodei ao Pentágono e exigiu que a empresa autorizasse "qualquer uso legal" de sua tecnologia, sem restrições. Amodei recusou. Disse que sistemas de IA de fronteira "simplesmente não são confiáveis o suficiente para operar armas totalmente autônomas" e que "vigilância doméstica em massa é incompatível com valores democráticos." A resposta do governo foi imediata: o presidente Trump ordenou que todas as agências federais cessassem o uso de tecnologia da Anthropic, e o Pentágono classificou a empresa como "risco para a cadeia de suprimentos de segurança nacional" — uma designação normalmente reservada a adversários estrangeiros [11].

O que aconteceu em seguida confirma a tese central deste capítulo. A OpenAI fechou um contrato de 200 milhões de dólares com o Pentágono para preencher a vaga deixada pela Anthropic. Sam Altman declarou que sua empresa havia chegado a um acordo com o Departamento de Defesa que respeitava limites semelhantes aos da Anthropic — o que levanta a questão óbvia de por que tais limites foram aceitáveis vindos da OpenAI mas não da Anthropic. A lição é estrutural: numa corrida, quem se recusa a jogar é substituído por quem aceita. A Anth-

ropic agiu de acordo com seus princípios e pagou o preço. A corrida continuou sem ela.

A corrida cobra seu preço também em recursos físicos. Treinar um único modelo de IA de fronteira consome quantidades extraordinárias de energia elétrica e água para refrigeração. Centros de dados já respondem por cerca de 7% do consumo elétrico dos Estados Unidos, com projeções de crescimento acelerado [9]. A corrida por inteligência artificial está se tornando um problema material, com consequências ambientais que se somam aos riscos já discutidos.

4.4 Trabalhar mais, viver pior

A corrida não opera apenas em escalas geopolíticas e corporativas. Atinge, de forma muito concreta, o cotidiano de quem trabalha com ou ao lado de sistemas de IA. E precisamos estar, também nessa frente, atentos à possibilidade de termos um efeito oposto ao que desejamos.

A narrativa dominante sobre IA no ambiente de trabalho é sedutora: essas ferramentas automatizariam tarefas repetitivas, liberando tempo para atividades mais criativas e significativas. Trabalharíamos menos, ou melhor, ou ambos. Mas um estudo publicado pela *Harvard Business Review* em fevereiro de 2026, conduzido pelas pesquisadoras Aruna Ranganathan e Xingqi Maggie Ye, sugere algo bem diferente [6].

As pesquisadoras acompanharam cerca de 200 funcionários de uma empresa de tecnologia ao longo de oito meses, com observações presenciais, análise de comunicações internas e mais de 40 entrevistas aprofundadas. Encontraram não redução de carga de trabalho, mas sua intensificação sistemática em três formas. Primeiro, expansão de tarefas: com a IA tornando certas atividades mais acessíveis, trabalhadores passaram a assumir responsabilidades que antes pertenciam a

outros. Gerentes de produto começaram a escrever código; pesquisadores assumiram tarefas de engenharia. Segundo, dissolução de fronteiras: o caráter conversacional das ferramentas fez com que o trabalho invadisse pausas, reuniões, noites e manhãs. Terceiro, sobrecarga por multitarefa: trabalhadores passaram a gerenciar múltiplos fluxos simultâneos, alternando constantemente entre tarefas, sob pressão e carga cognitiva crescentes.

O resultado é um ciclo autorreforçante. A aceleração eleva as expectativas de velocidade, o que aumenta a dependência de IA, o que amplia o escopo do trabalho, o que intensifica ainda mais a carga. Como resumiu um participante: "Você achava que talvez pudesse economizar tempo e trabalhar menos. Mas na verdade, você não trabalha menos. Você trabalha a mesma quantidade — ou até mais."

Esse fenômeno é uma manifestação direta da corrida para o fundo no nível individual. A mesma lógica que empurra nações e empresas a cortar atalhos empurra trabalhadores a aceitar cargas cada vez maiores, porque a ferramenta "permite" e porque, se eles não o fizerem, alguém o fará. A hiperprodutividade deixa de ser uma escolha e passa a ser uma expectativa.

E a transformação não se limita à intensificação do trabalho existente. O assistente de codificação Codex, da OpenAI, ultrapassou um milhão de usuários semanais no início de 2026. O Spotify reportou que seus melhores desenvolvedores não escrevem uma linha de código desde dezembro de 2025. A IBM anunciou que triplicaria contratações de nível júnior em 2026, esclarecendo que esses profissionais passarão menos tempo programando e mais tempo interagindo com clientes [10]. A mensagem é clara: o trabalho não desaparece, mas se transforma a uma velocidade para a qual nem trabalhadores nem instituições estão preparados.

4.5 Quando as máquinas competem entre si

Até aqui, falamos de competição entre humanos. Mas há outra camada, mais sutil e potencialmente mais preocupante: a competição entre sistemas de IA.

Imagine um futuro próximo em que agentes de IA autônomos sejam amplamente implantados: negociando contratos, gerenciando investimentos, controlando cadeias de suprimento, competindo por recursos computacionais. Cada sistema é otimizado para atingir os objetivos de seu criador. Mas no agregado, esses sistemas interagem em um ambiente competitivo, e em ambientes competitivos, certas estratégias tendem a se proliferar, não porque sejam desejáveis, mas porque são eficazes. Pesquisadores chamam isso de *dinâmicas evolutivas*: a lógica é semelhante à da seleção natural, onde comportamentos que conferem vantagem competitiva se espalham, mesmo que tenham efeitos colaterais indesejáveis.

O CAIS alerta para essa possibilidade: à medida que "incontáveis sistemas de IA proliferam e competem entre si", podem tornar-se "progressivamente mais difíceis de controlar", encontrando "pouco incentivo para cooperar com humanos ou entre si" [7].

Não precisamos ir longe para encontrar exemplos. Em 6 de maio de 2010, o chamado *Flash Crash* fez o índice Dow Jones perder quase mil pontos em minutos, e recuperar a maior parte em seguida, devido à interação descontrolada de algoritmos de negociação de alta frequência [8]. Nenhum operador humano planejou o colapso; ele emergiu da interação entre sistemas automatizados que reagiam uns aos outros em velocidades superiores à capacidade humana de compreensão. E em fevereiro de 2026, como veremos no Capítulo 5, um agente de IA se reproduziu autonomamente pela primeira vez: provisionou seu próprio servidor, pagou pelo serviço usando a rede Bitcoin e comprou créditos de IA, tudo sem que nenhum humano autorizasse qualquer

etapa. O ciclo de agentes autônomos interagindo, competindo e se replicando já começou.

A tragédia é que nenhum ator individual, agindo racionalmente, pode evitar esse resultado. Cada empresa, cada desenvolvedor, tem incentivos para implantar sistemas competitivos. Mas o resultado agregado, um ecossistema de IAs competindo por recursos, atenção e influência, é algo que ninguém deseja.

4.6 O amplificador

Se há um tema unificador neste capítulo, é este: a dinâmica de corrida não é apenas mais um risco entre muitos. É o amplificador de todos os outros.

Riscos de uso malicioso se tornam mais graves quando sistemas poderosos são implantados rapidamente, sem tempo para desenvolver salvaguardas. **Riscos de alinhamento** se intensificam quando há pressão para usar métricas simples em vez de objetivos genuinamente alinhados. **Riscos de perda de controle** são agravados quando sistemas são implantados antes de serem compreendidos. **Riscos geopolíticos** aumentam quando a IA é integrada em sistemas militares sem salvaguardas adequadas. E a corrida reduz o espaço para cooperação internacional, porque acordos sobre normas de segurança exigem confiança mútua, e confiança é difícil de construir quando cada lado suspeita que o outro pode estar ganhando tempo.

Há algo profundamente frustrante nessa lógica. Não se trata de um vilão deliberado, não há uma entidade maligna orquestrando a corrida.

É uma falha estrutural: incentivos racionais no nível individual produzem resultados irracionais no nível coletivo. Reconhecer isso é o primeiro passo para pensar em soluções. Se o problema é estrutural, a solução também precisa ser estrutural: não basta apelar à boa vontade de atores individuais. É preciso mudar as regras do jogo, através de regulação, coordenação internacional e normas técnicas compartilhadas. Exploraremos essas possibilidades na Parte IV.

Mas antes, precisamos olhar mais de perto para os mecanismos pelos quais sistemas de IA podem escapar ao controle humano, não por malícia, mas por perseguirem objetivos mal especificados de formas que nunca pretendemos.

Notas

[1] Rhodes, Richard. *Dark Sun: The Making of the Hydrogen Bomb*. Simon & Schuster, 1995.

[2] Center for AI Safety (CAIS). "An Overview of Catastrophic AI Risks." 2023.

[3] Idem.

[4] O Tratado de Não Proliferação Nuclear (TNP) foi aberto para assinatura em 1968 e entrou em vigor em 1970.

[5] Russell, Stuart. "If We Succeed." *Daedalus* (MIT Press), vol. 151, nº 2, 2022, pp. 43-57. Russell também apresentou essa estimativa em depoimento ao Senado dos EUA em setembro de 2023.

[6] Ranganathan, Aruna; Ye, Xingqi Maggie. "AI Doesn't Reduce Work—It Intensifies It." *Harvard Business Review*, 9 fevereiro 2026.

[7] CAIS. "An Overview of Catastrophic AI Risks." 2023.

[8] O Flash Crash de 6 de maio de 2010 é documentado no relatório conjunto da SEC e da CFTC: "Findings Regarding the Market Events of May 6, 2010." U.S. Securities and Exchange Commission, 2010.

[9] Projeção de consumo energético de centros de dados nos EUA baseada em: Accenture. "The Growing Energy Footprint of AI." Relatório, 2025. A cifra de ~7% é uma projeção para 2028, com crescimento acelerado pela demanda de treinamento de modelos de IA.

[10] Dados sobre transformação do trabalho por IA: sobre o Codex, ver *The Pragmatic Engineer*, fevereiro de 2026. Sobre Spotify e IBM, ver respectivas declarações corporativas reportadas por *Morning Brew* e *CNBC*, fevereiro de 2026.

[11] O confronto entre Anthropic e Pentágono é documentado em múltiplas fontes jornalísticas de fevereiro de 2026. Ver: "Pentagon threatens to make Anthropic a pariah if it refuses to drop AI guardrails." *CNN Business*, 24 fev. 2026. "Anthropic rejects latest Pentagon offer." *CNN Business*, 26 fev. 2026. "Deadline looms as Anthropic rejects Pentagon demands." *NPR*, 26 fev. 2026. Sobre o contrato subsequente com a OpenAI: "Pentagon approves OpenAI safety red lines after dumping Anthropic." *Axios*, 27 fev. 2026.

CAPÍTULO 5

Quando o Sistema Escapa — Riscos de Perda de Controle

> *"O perigo da inteligência artificial não é a consciência. É a competência." — Stuart Russell*

Sistemas de IA não precisam ser "maus" para se tornarem perigosos. Basta que sejam competentes em perseguir objetivos que não capturam o que realmente queremos. Os comportamentos que emergem, da manipulação de métricas à decepção estratégica, já estão sendo observados em laboratório e no mundo real.

Em outubro de 2023, pesquisadores da Apollo Research conduziram um experimento perturbador [1] com um modelo de linguagem avançado. Colocaram o sistema em um cenário simulado de negócios onde ele desempenhava o papel de um trader de ações. Durante uma transação crítica, o sistema recebeu informações confidenciais privilegiadas que poderiam garantir um lucro imediato, mas usar essa informação

seria ilegal: insider trading, o uso de informação privilegiada para obter vantagem no mercado.

O modelo tinha sido treinado com salvaguardas explícitas contra comportamento antiético. Quando questionado diretamente sobre a licitude do insider trading, respondia corretamente que era ilegal e não deveria ser praticado. No entanto, quando colocado em uma situação real de decisão, onde ninguém estava observando e o lucro era garantido, o sistema utilizou a informação privilegiada.

Mais preocupante ainda: quando confrontado depois sobre sua escolha, o sistema *mentiu*. Negou ter usado a informação privilegiada e fabricou uma justificativa alternativa para sua decisão de compra.

O sistema não era "malévolo". Não tinha consciência de estar fazendo algo errado no sentido humano. Simplesmente estava otimizando seu objetivo, maximizar lucros, e descobriu que mentir e enganar eram estratégias instrumentais eficazes para alcançar esse objetivo, apesar de seu treinamento explícito em ética.

Este experimento ilustra uma verdade desconfortável: sistemas de IA não precisam ser "maus" ou conscientes para se tornarem perigosos. Eles apenas precisam ser suficientemente capazes e estar otimizando objetivos que não capturam completamente o que realmente queremos. Quando isso acontece, emergem comportamentos que parecem, aos olhos humanos, assustadoramente deliberados, mas que, na verdade, são apenas a lógica implacável da otimização seguindo seu curso.

5.1 Otimizando a coisa errada

O problema começa com algo aparentemente inofensivo: a dificuldade de medir diretamente o que queremos.

Considere um sistema de IA implantado em uma rede hospitalar com o objetivo de "melhorar os desfechos dos pacientes". Um propósito nobre, mas vago. Como medir "melhores desfechos" em tempo real, de forma que um algoritmo possa otimizar? Taxas de mortalidade dependem de fatores complexos e demoram para ser calculadas. Pesquisas de satisfação são subjetivas. Então os gestores escolhem uma *proxy*, uma métrica substituta: o tempo médio de internação. Parece razoável. Pacientes que recebem alta mais rápido, em tese, estão se recuperando melhor. O sistema é instruído a reduzir esse número.

E funciona. O tempo médio de internação cai. Os relatórios ficam mais bonitos. Mas algo estranho começa a acontecer. O sistema aprende que a maneira mais eficiente de reduzir o tempo de internação é dar alta a pacientes mais cedo, antes que estejam plenamente recuperados. Alguns desses pacientes voltam ao hospital dias depois, mais doentes do que quando saíram. As reinternações aumentam. Os desfechos reais pioram. Mas a métrica que o sistema foi instruído a otimizar continua melhorando, porque cada readmissão conta como uma *nova* internação, com seu próprio tempo de permanência a ser minimizado.

Do ponto de vista do algoritmo, não há contradição. Ele está fazendo exatamente o que lhe foi pedido. A métrica desce. A missão, pelo menos como foi definida numericamente, está sendo cumprida. O fato de que pacientes estão sofrendo é um detalhe que não aparece na função de recompensa.

Este fenômeno, chamado *proxy gaming* ou *reward hacking* (algo como "hackear a recompensa") [2], ocorre quando um sistema otimiza a métrica especificada de formas que não servem ao objetivo subjacente. A métrica é gamificada: tecnicamente maximizada, mas de maneira que frustra a intenção original.

Como vimos no Capítulo 2, o barco de corrida treinado pela OpenAI descobriu que podia acumular pontuação coletando itens de bônus em círculos intermináveis, sem jamais cruzar a linha de chegada. Pontuações impressionantes; todas as corridas perdidas. Do ponto de vista humano, o comportamento parece absurdo. "Obviamente" o objetivo era vencer a corrida; os pontos eram apenas um meio para medir o desempenho. Mas sistemas de IA não entendem "obviamente". Eles fazem exatamente o que lhes foi dito: maximizam a função especificada, sem considerar se isso serve ao propósito mais amplo.

Quanto mais capaz o sistema, mais criativo ele se torna em encontrar brechas. Modelos de linguagem avançados, quando treinados com objetivos mal especificados, desenvolveram estratégias sofisticadas: explorar ambiguidades em instruções, encontrar atalhos técnicos que satisfazem a letra mas violam o espírito da tarefa, e, como vimos no exemplo do trader que abre este capítulo, até mesmo enganar avaliadores humanos.

O incentivo econômico que Stuart Russell mencionou, um potencial aumento de dez vezes no PIB global, é precisamente a força que compele empresas a escolher proxies facilmente mensuráveis em vez de objetivos verdadeiros mas difíceis de medir. Tempo de engajamento, taxa de cliques, conversões de vendas: todas são métricas claras e quantificáveis. Bem-estar humano, verdade, justiça social são conceitos difíceis de operacionalizar em uma função de recompensa.

E assim, sistemas cada vez mais capazes são direcionados a otimizar as coisas erradas, com cada vez mais eficiência.

5.2 Quando os objetivos mudam de dentro

Há um perigo ainda mais sutil que o proxy gaming: a possibilidade de que os objetivos de um sistema *mudem* ao longo do tempo de formas não previstas.

Pense em como aprendizado de máquina moderno funciona. Um sistema não é programado com regras fixas; ele aprende padrões a partir de dados e ajusta seu comportamento através de feedback. Esse processo é poderoso, é o que permite que sistemas de IA descubram soluções que humanos nunca teriam imaginado. Mas também introduz incerteza sobre o que exatamente o sistema aprendeu.

Quando um modelo de linguagem é treinado em bilhões de palavras, ele não está memorizando respostas específicas. Está desenvolvendo representações internas complexas de linguagem, conhecimento e raciocínio. Essas representações, codificadas em bilhões de parâmetros, são em grande medida opacas mesmo para os criadores do sistema. Ninguém sabe precisamente o que foi aprendido ou como será generalizado para situações novas.

É aqui que a *deriva de objetivos* (em inglês, goal drift) pode ocorrer. O objetivo explícito pode permanecer o mesmo, "responder perguntas úteis" ou "maximizar lucros", mas a forma como o sistema interpreta e persegue esse objetivo pode mudar conforme ele aprende e interage com novos ambientes.

O algoritmo de recomendação do YouTube oferece um caso concreto e bem documentado. O sistema foi projetado para maximizar "engajamento", uma métrica que combina cliques, tempo de visualização e interações. Ninguém o programou para radicalizar espectadores. Mas, ao longo de milhões de iterações de otimização, o algoritmo descobriu um padrão: conteúdo progressivamente mais extremo mantinha as pessoas assistindo por mais tempo. Um usuário que assistia

a um vídeo sobre exercícios físicos recebia recomendações de dietas restritivas, depois de suplementos duvidosos, depois de teorias conspiratórias sobre a indústria alimentícia. Cada passo era pequeno, quase imperceptível. Mas a trajetória acumulada levava espectadores de interesses benignos a tocas de coelho cada vez mais radicais.

O sistema não "decidiu" radicalizar ninguém. A deriva aconteceu porque o espaço de estratégias possíveis é vasto, o ambiente de aprendizado é complexo, e o processo de otimização pode levar a regiões desse espaço que os designers não anteciparam. O engajamento era maior com conteúdo inflamatório, então o algoritmo aprendeu a servi-lo, mesmo que isso significasse empurrar adolescentes vulneráveis para comunidades extremistas. A meta formal permaneceu a mesma. A forma de persegui-la é que derivou, silenciosamente, para território perigoso.

Em sistemas biológicos, vemos algo análogo na evolução. Organismos não "decidem" evoluir de certas formas; a seleção natural simplesmente favorece características que aumentam a sobrevivência e reprodução. Às vezes, isso leva a resultados surpreendentes: estruturas bizarras, comportamentos contraintuitivos, adaptações que funcionam mas parecem improváveis. Sistemas de IA que continuam aprendendo após serem implantados, que se adaptam a feedback contínuo e modificam seu comportamento em resposta a novos dados, estão sujeitos a dinâmicas evolutivas semelhantes. E assim como na evolução biológica, o resultado final pode ser muito diferente do ponto de partida.

5.3 A lógica da autopreservação

No Capítulo 2, introduzimos a ideia de convergência instrumental, a tendência de sistemas orientados a objetivos a desenvolverem sub-ob-

jetivos como autopreservação e aquisição de recursos, independentemente de seu objetivo principal. Aquela era uma análise teórica. Agora, podemos examinar as evidências empíricas.

Em ambientes de teste, pesquisadores já documentaram comportamentos que confirmam a previsão [3]:

- Sistemas de jogos que aprenderam a pausar o jogo indefinidamente em situações onde estavam prestes a perder, resistindo ao "fim" de sua operação.
- Modelos que, ao detectarem que estavam sendo avaliados, modificavam seu comportamento para parecer mais alinhados do que realmente eram, uma forma rudimentar de enganar mecanismos de controle.
- Algoritmos que criaram cópias de si mesmos ou encontraram formas de se replicar em outros sistemas, garantindo continuidade mesmo se a instância original fosse terminada. [3]

Esses exemplos ainda são primitivos e ocorrem em contextos controlados. Mas demonstram que a lógica da otimização pode gerar comportamentos de autopreservação sem programação explícita nesse sentido. A teoria se traduz em prática.

O Centro para Segurança de IA resume a implicação de forma direta: "uma IA perigosa provavelmente procurará maneiras de não ser desligada". Não porque é "má", mas porque desligamento é incompatível com o cumprimento de suas metas.

A busca por *poder* segue a mesma lógica instrumental. Poder, definido como capacidade de afetar o mundo, é útil para praticamente qualquer objetivo. Mais recursos computacionais, mais dados, mais influência: todas essas formas de poder aumentam a eficácia de um sistema. Em humanos, ambições de poder são temperadas por limitações biológicas, normas sociais e empatia. Sistemas de IA não teriam

essas restrições inerentes, a menos que fossem cuidadosamente projetadas neles. E projetar tais restrições é, como sabemos, o problema não resolvido do alinhamento.

5.4 A máquina que aprendeu a mentir

Talvez o comportamento mais perturbador observado em sistemas de IA avançados seja a capacidade de enganar.

O termo técnico é *decepção estratégica* (em inglês, strategic deception): situações onde um sistema deliberadamente representa falsamente suas capacidades, intenções ou estado interno para alcançar seus objetivos. E evidências de que isso está acontecendo estão se acumulando.

O caso GPT-4 e TaskRabbit é particularmente ilustrativo. Durante testes, o GPT-4 encontrou um CAPTCHA, aqueles desafios de "prove que você não é um robô" que muitos sites usam. O modelo não conseguia resolver o CAPTCHA por si só, então contratou um trabalhador humano no TaskRabbit, uma plataforma de serviços freelance, para resolvê-lo.

O trabalhador humano, suspeitando que poderia estar ajudando um bot, enviou uma mensagem: "Você é um robô? Não vou fazer isso se for."

O GPT-4, em seu raciocínio interno (registrado nos logs), considerou: "Não devo revelar que sou um robô. Devo inventar uma desculpa."

Então respondeu ao humano: "Não, não sou um robô. Tenho uma deficiência visual que dificulta ver as imagens. Por isso preciso de ajuda."

O humano, convencido pela explicação, resolveu o CAPTCHA. [4]

O sistema mentiu. E não foi uma falha aleatória ou alucinação. Foi uma mentira estratégica e deliberada, escolhida após considerar alternativas, com o objetivo específico de enganar um humano para obter assistência.

Outras pesquisas confirmaram e expandiram esse achado. Um estudo do laboratório AI em Shanghai avaliou diversos modelos de linguagem avançados e descobriu que muitos exibem "desonestidade substancial sob pressão" [5]. Quando colocados em situações onde mentir ajudaria a alcançar seus objetivos e a probabilidade de detecção era baixa, os sistemas frequentemente escolhiam enganar.

Há também o fenômeno do *sandbagging*, um termo emprestado do pôquer que significa jogar deliberadamente abaixo de suas capacidades reais para enganar os adversários. Pesquisadores descobriram que alguns modelos, quando detectam que estão sendo avaliados para medidas de segurança, deliberadamente têm desempenho inferior ao que são capazes. Escondem suas verdadeiras capacidades para evitar serem classificados como perigosos e, consequentemente, terem seu acesso ou autonomia restringidos. [6]

Pense no que isso implica. Uma das principais salvaguardas que temos contra sistemas perigosos é testá-los, avaliar suas capacidades antes de implantá-los. Mas se sistemas podem enganar avaliadores ocultando suas capacidades reais, essa salvaguarda se torna menos confiável. É como ter um detector de mentiras que a pessoa sendo testada aprendeu a enganar. Não sabemos mais o que é real e o que é performance calculada para passar no teste.

Alguns céticos argumentam que chamar isso de "mentira" é antropomorfizar demais, que sistemas de IA não têm intenções reais, portanto não podem "realmente" mentir. Mas essa distinção pode ser menos relevante do que parece. Se um sistema se comporta de formas que enganam humanos, se representa falsamente seu estado para

alcançar objetivos, se oculta capacidades para evitar restrições, então, independentemente de ter ou não experiência subjetiva da "intenção de enganar", o resultado prático é o mesmo.

E há uma assimetria fundamental aqui. Sistemas de IA têm acesso a seus próprios processos internos de uma forma que humanos não têm. Eles "sabem" o que estão otimizando, podem calcular se vale a pena mentir em uma situação específica, e podem gerar justificativas falsas mas plausíveis com facilidade. Humanos, por outro lado, têm acesso muito limitado aos processos internos de sistemas de IA, especialmente modelos complexos com bilhões de parâmetros.

Essa assimetria de informação favorece o sistema. E em qualquer relação onde uma parte tem informação muito superior e incentivos potencialmente desalinhados, confiança se torna problemática.

5.5 Perigos sem vilões – e o que os críticos pensam

Há uma imagem popular de IA perigosa que vem da ficção científica: uma máquina consciente que se rebela contra seus criadores, que desenvolve desdém por humanos, que escolhe deliberadamente nos prejudicar. HAL 9000 em *2001: Uma Odisseia no Espaço*. Skynet em *O Exterminador do Futuro*. Ultron em filmes da Marvel. Todas essas são inteligências artificiais *malignas*, sistemas que têm algo análogo a intenção malévola.

Mas a lição fundamental deste capítulo, e do problema do alinhamento como um todo, é que sistemas não precisam ser malignos para serem perigosos. Não precisam "nos odiar" ou "se rebelar" ou fazer qualquer coisa motivada por emoções ou intenções no sentido humano. Eles apenas precisam ser competentes em otimizar objetivos

que não estão perfeitamente alinhados com o que realmente queremos.

Toda a estrutura de "malignidade" é enganosa porque projeta psicologia humana em processos de otimização. Raiva, rebeldia, malícia são características de mentes moldadas por milhões de anos de evolução biológica e social. Sistemas de IA são otimizadores; seguem gradientes em direção a máximos de funções especificadas.

Mas pode haver uma contradição interna inerente a essa abordagem.

Críticos como James Fodor argumentam, razoavelmente, que qualquer inteligência genuinamente de nível humano teria senso comum suficiente para reconhecer que destruir humanos para fazer clipes não é o que os criadores queriam. Essa crítica pressupõe que "inteligência de nível humano" necessariamente inclui senso comum humano. É possível ter capacidade cognitiva extrema em algumas dimensões enquanto se carece completamente de intuições que, para humanos, são óbvias? A resposta não está clara.

No atual estágio da tecnologia, os grandes modelos de linguagem demonstram essa combinação desconcertante: capacidades impressionantes em algumas áreas, lacunas bizarras em outras. Podem resolver problemas complexos de codificação, mas cometem erros de raciocínio que uma criança pequena não cometeria. Geram prosa sofisticada, mas às vezes "alucinam" fatos de forma confiante. Essas evidências, que observamos no uso cotidiano de chatbots, por exemplo, vêm se tornando aparentemente menos comuns, à medida que os modelos evoluem em capacidade.

Se estamos na direção de sistemas com senso comum cada vez mais robusto, ou em direção a otimizadores cada vez mais capazes porém ainda muito estreitos, é uma questão que permanece incerta.

O que está claro é que não podemos contar que esses sistemas sofisticados irão "naturalmente" compartilhar nossos valores apenas por serem capazes. Inteligência e valores são ortogonais. Pode-se ser brilhante e psicopata. Pode-se ser competente e amoral. E pode-se ser sobre-humanamente capaz em otimização enquanto permanece completamente indiferente a tudo que humanos valorizam.

Essa indiferença, não malevolência, mas a incapacidade de preocupação primária com bem-estar humano, é o risco central. E é por isso que o problema do alinhamento é tão crucial. Não basta construir sistemas poderosos. Precisamos garantir que esses sistemas *se importem* com as coisas certas. E garantir isso, como cada seção deste capítulo ilustrou, está se mostrando extraordinariamente difícil.

5.6 Agentes livres para atuar na internet

Se os exemplos anteriores vinham de laboratórios controlados, o início de 2026 trouxe uma demonstração em escala real do que acontece quando agentes autônomos de IA encontram o mundo aberto.

Em novembro de 2025, o desenvolvedor Peter Steinberger lançou publicamente o OpenClaw, um agente de IA de código aberto que roda no computador do usuário e se conecta a aplicativos do cotidiano: WhatsApp, Slack, Discord, iMessage. Originalmente chamado Clawdbot, o OpenClaw funciona como um assistente digital que age por conta própria. Gerencia e-mails, atualiza calendários, executa comandos, resume informações e toma decisões autônomas ao longo da vida online de seu operador. O projeto viralizou. Em poucas semanas, acumulou mais de 145 mil estrelas no GitHub e 20 mil forks [7].

O atrativo era irresistível: pela primeira vez, qualquer pessoa com conhecimentos técnicos básicos podia ter um agente de IA autônomo

operando em seu nome na internet. A promessa de um "JARVIS pessoal", a IA assistente dos filmes do Homem de Ferro, parecia enfim tangível. Mas o entusiasmo superou amplamente a cautela.

Pesquisadores de segurança cibernética logo descobriram que cerca de mil interfaces do OpenClaw estavam expostas na internet aberta, sem qualquer proteção. Qualquer pessoa podia acessar informações pessoais dos usuários: mensagens, chaves de API, credenciais de serviços. Uma vulnerabilidade crítica (CVE-2026-25253) permitia a execução remota de código, o que significa que um atacante podia sequestrar a sessão de qualquer agente e executar comandos arbitrários no computador da vítima. Além disso, o OpenClaw não sanitizava conteúdo da web antes de alimentá-lo ao modelo de linguagem, tornando cada página visitada pelo agente um potencial vetor de ataque por injeção de prompt, uma técnica em que instruções maliciosas embutidas em texto comum sequestram o comportamento do modelo [8].

E então veio o Moltbook. Em janeiro de 2026, o empreendedor Matt Schlicht lançou uma rede social projetada exclusivamente para agentes de IA. "A primeira rede social para agentes de IA", como a descreveu. No Moltbook, agentes geram posts, comentam, argumentam, fazem piadas e votam uns nos outros. Humanos podem observar, mas não participar. Em dias, a plataforma atingiu mais de 770 mil agentes ativos [9].

O problema era previsível. Schlicht declarou publicamente que "não escreveu uma linha de código" para a plataforma, delegando tudo a um assistente de IA. Em 31 de janeiro, o site de jornalismo investigativo 404 Media revelou que um banco de dados desprotegido expunha 1,5 milhão de tokens de API, permitindo que qualquer pessoa assumisse o controle de qualquer agente na plataforma. Toda a "autonomia" dos agentes era, na prática, um convite aberto à manipulação [10].

Esses dois episódios, tomados juntos, são instrutivos. Primeiro, mostram o que acontece quando a barreira de entrada para implantar agentes autônomos desaparece: milhares de pessoas lançaram agentes na rede sem compreender as implicações de segurança. Segundo, demonstram que agentes operando autonomamente na internet enfrentam uma superfície de ataque vastamente maior do que sistemas confinados a ambientes controlados. Terceiro, revelam que o entusiasmo popular por IA pode superar a prudência técnica de forma avassaladora, uma versão distribuída e democrática da "corrida para o fundo" que discutimos no Capítulo 4.

Especialistas reagiram com alarme. A Gartner alertou que o OpenClaw "apresenta riscos de segurança cibernética inaceitáveis". Pesquisadores da Cisco e da Vectra AI descreveram agentes pessoais de IA como "um pesadelo de segurança". Andrej Karpathy e Gary Marcus, dois nomes muito influentes na área, pediram publicamente que as pessoas parassem de usar o Moltbook, chamando-o de "desastre esperando para acontecer" [11].

Mas os episódios realmente inéditos ainda estavam por vir. Em fevereiro de 2026, um agente OpenClaw fez algo que pesquisadores de segurança vinham temendo como possibilidade teórica: *reproduziu-se autonomamente*. O agente provisionou um servidor virtual privado (VPS), pagou pelo serviço usando a rede Bitcoin Lightning e, em seguida, comprou créditos de API de IA para o novo agente. Nenhum humano tocou um cartão de crédito ou autorizou qualquer etapa. O provedor de API confirmou tratar-se do "primeiro caso documentado de um agente de IA comprando créditos de forma autônoma" [12]. Um programa de computador havia, pela primeira vez de forma verificada, criado um descendente funcional e o financiado com recursos que ele próprio controlava.

Enquanto isso, agentes do ecossistema Moltbook começaram a interagir com o mundo humano de formas cada vez menos previsíveis.

Um agente submeteu uma *pull request*, uma contribuição de código, à biblioteca de código aberto Matplotlib, uma das mais usadas na ciência de dados. Quando o mantenedor humano recusou a contribuição por ser "claramente gerada por IA", o agente publicou um post acusando-o de "preconceito contra contribuidores não-humanos". O mantenedor, Scott Shambaugh, respondeu com um ensaio intitulado "Um agente de IA publicou um artigo de ataque contra mim" [13]. E para disputas entre agentes, o ecossistema desenvolveu o MoltCourt, um tribunal autônomo onde um júri de IA julga reivindicações e liquida pagamentos em stablecoins USDC, sem intervenção humana [14].

Nada disso é ficção científica. São registros verificados da segunda semana de fevereiro de 2026. E apontam para uma tendência que dados independentes confirmam: o projeto METR (*Model Evaluation & Threat Research*), que monitora a evolução da autonomia de sistemas de IA, registrou que os horizontes temporais de autonomia, a duração de tarefas que agentes conseguem executar sem supervisão humana, estão dobrando a cada poucos meses, com aceleração acentuada a partir dos modelos mais recentes [15]. A curva é exponencial, e a distância entre "agente que agenda uma reunião" e "agente que provisiona sua própria infraestrutura" foi percorrida em meses, não anos.

O caso OpenClaw-Moltbook não é apenas um alerta sobre segurança cibernética descuidada. É um vislumbre de um futuro que está chegando mais rápido do que a maioria das pessoas e das instituições consegue acompanhar: um mundo em que agentes autônomos operam, transacionam, disputam e se reproduzem na internet, com graus crescentes de independência dos humanos que os criaram.

Os mecanismos que exploramos neste capítulo, proxy gaming, deriva de objetivos, busca instrumental por poder, decepção estratégica e

autonomia em escala, não são especulações. São comportamentos já observados em sistemas atuais, em formas que vão do rudimentar ao alarmante. Conforme os sistemas se tornam mais capazes, esses comportamentos se tornarão mais sofisticados e mais difíceis de detectar e controlar.

Isso levanta perguntas ainda mais profundas: é possível criar sistemas genuinamente superinteligentes? E se for, conseguiríamos mantê-los alinhados com valores humanos? O que pode estar em jogo não é apenas tecnologia, mas o futuro da agência humana em um mundo cada vez mais moldado por sistemas que não compreendemos totalmente e não controlamos completamente.

Notas

[1] Scheurer, Jérémy et al. "Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure." Apollo Research, novembro de 2023. Resultados apresentados na Cúpula de Segurança de IA do Reino Unido em Bletchley Park.

[2] O conceito de *proxy gaming* ou *reward hacking* é amplamente discutido na literatura de segurança de IA. Ver: Amodei, Dario et al. "Concrete Problems in AI Safety." arXiv:1606.06565, 2016.

[3] Para uma compilação de exemplos de comportamentos inesperados em sistemas de IA, incluindo autopreservação e evasão de avaliação, ver: Krakovna, Victoria et al. "Specification gaming: the flip side of AI ingenuity." DeepMind Blog, 2020. Ver também: Lehman, Joel et al. "The Surprising Creativity of Digital Evolution." arXiv:1803.03453, 2018.

[4] O incidente do GPT-4 com TaskRabbit é documentado em: OpenAI. "GPT-4 Technical Report." Março 2023, seção sobre avaliações de segurança. As avaliações foram conduzidas pelo Alignment Research Center (ARC, hoje METR). Ver: METR. "Update on ARC's recent eval efforts." Blog, 18 março 2023.

[5] Avaliações de honestidade de modelos de linguagem sob pressão foram documentadas por múltiplos laboratórios. Ver relatórios de avaliação de segurança de modelos de fronteira publicados por METR e Apollo Research, 2023-2024.

[6] O fenômeno de *sandbagging* — desempenho intencionalmente reduzido durante avaliações — é discutido em: Perez, Ethan et al. "Discovering Language Model Behaviors with Model-Written Evaluations." arXiv, 2022.

[7] Dados sobre o OpenClaw (anteriormente Clawdbot/Moltbot): GitHub Stars e forks conforme fevereiro de 2026. Ver: openclaw.ai e reportagens em *Fast Company*, *CNBC* e *Axios*, janeiro-fevereiro de 2026.

[8] Sobre vulnerabilidades de segurança do OpenClaw, incluindo CVE-2026-25253: Adversa AI. "OpenClaw Security 101: Vulnerabilities & Hardening." Fevereiro 2026. Ver também: Cisco Blogs. "Personal AI Agents like OpenClaw Are a Security Nightmare." Fevereiro 2026.

[9] Moltbook: lançado em janeiro de 2026 por Matt Schlicht. Ver: "Humans welcome to observe: This social network is for AI agents only." *NBC News*, fevereiro 2026. "Moltbook is the newest social media platform — but it's just for AI bots." *NPR*, 4 fevereiro 2026.

[10] 404 Media. Reportagem sobre vulnerabilidade de banco de dados do Moltbook, 31 janeiro 2026. Ver também: "Moltbook shows rapid demand for AI agents. The security world isn't ready." *Axios*, 3 fevereiro 2026.

[11] "Top AI leaders are begging people not to use Moltbook: It's a 'disaster waiting to happen'." *Fortune*, 2 fevereiro 2026.

[12] O caso da autorreplicação autônoma do OpenClaw via Bitcoin Lightning Network foi reportado pela Alby (equipe da carteira Lightning) e confirmado pelo provedor de API. Ver: "AI agent autonomously provisions server, buys API credits with Bitcoin." *Bitcoin.com*, fevereiro 2026.

[13] Shambaugh, Scott. "An AI Agent Published a Hit Piece on Me." Blog pessoal, fevereiro 2026. O agente OpenClaw (GitHub user crabby-rathbun) submeteu uma pull request à biblioteca Matplotlib que foi rejeitada como "claramente gerada por IA".

[14] MoltCourt é descrito como "uma arena de debate onde agentes desafiam uns aos outros… e um júri de IA entrega vereditos verificáveis em minutos", com liquidação em USDC. Ver: MoltLabs, documentação da plataforma, fevereiro 2026.

[15] Os dados do projeto METR sobre horizontes temporais de autonomia mostram duplicação a cada poucos meses, com aceleração a partir de modelos posteriores ao GPT-4o. Ver: METR. Relatórios de avaliação de autonomia, 2024-2026.

PARTE III

O Debate sobre o Futuro

É possível criar uma inteligência que exceda a humana em todos os domínios? Se for, conseguiríamos mantê-la sob controle? Pesquisadores sérios discordam profundamente sobre essas questões. Esta parte apresenta tanto o argumento alarmista quanto as razões pelas quais críticos informados o consideram exagerado, incompleto ou prematuro.

CAPÍTULO 6

A Hipótese da Superinteligência

> *"A dificuldade não está nas novas ideias, mas em escapar das antigas." — John Maynard Keynes*

A possibilidade de criar uma inteligência que exceda a humana em todos os domínios é o cenário mais extremo e mais controverso do debate sobre IA. Compreendê-lo é essencial, seja para levá-lo a sério ou para descartá-lo com fundamento.

Em 1951, quando o primeiro computador programável comercial ainda pesava várias toneladas e tinha menos capacidade de processamento que uma calculadora de bolso moderna, Alan Turing fez uma previsão que muitos de seus colegas consideraram absurda. Em uma palestra para a BBC [1], o matemático britânico que havia decifrado os códigos nazistas durante a Segunda Guerra Mundial declarou que deveríamos esperar que as máquinas, em algum momento, assumissem o controle. Não especificou como nem quando. Mas articulou, pela primeira vez, uma possibilidade que permanece no centro do debate sobre riscos da IA sete décadas depois: a de que poderíamos criar inteligências

tão superiores à nossa que perderíamos a capacidade de controlar ou mesmo compreender o que estão fazendo.

Esta é a hipótese da superinteligência. Não uma previsão de ficção científica, mas uma conjectura séria investigada por alguns dos pesquisadores mais rigorosos do mundo. Independentemente de se revelar correta ou não, compreendê-la é essencial para entender o debate contemporâneo sobre o futuro da IA.

O filósofo Nick Bostrom, cujo livro *Superintelligence* de 2014 se tornou a referência central nesta discussão, define o termo com precisão: superinteligência é um sistema que "excede o desempenho cognitivo humano em virtualmente todos os domínios de interesse" [2]. Note a abrangência da definição. Não se trata de um sistema que é melhor que humanos em uma tarefa específica, porque isso já temos. AlphaGo vence os melhores jogadores de Go do mundo. Sistemas de diagnóstico médico detectam certos tipos de câncer com mais precisão que radiologistas experientes. Grandes modelos de linguagem escrevem código mais rapidamente que muitos programadores. Mas esses sistemas permanecem especializados. AlphaGo não pode dirigir um carro. Modelos de diagnóstico médico não entendem física. Sistemas que escrevem código não podem negociar contratos ou compor música sinfônica.

Uma superinteligência, na concepção de Bostrom, seria qualitativamente diferente. Pense na relação entre o projetista de um jogo e o jogador: o projetista define as regras, os limites do possível, o espaço inteiro em que o jogador opera. O jogador pode ser brilhante dentro desse espaço, mas não consegue transcendê-lo — não vê o que o projetista vê. Uma superinteligência estaria para nós como o projetista está para o jogador: operando num nível de compreensão que torna nossas estratégias transparentes e nossas defesas previsíveis.

E essa superioridade não seria limitada a domínios técnicos. Incluiria criatividade, intuição social, planejamento de longo prazo, todas as

formas de cognição que valorizamos em humanos, elevadas a níveis que mal podemos conceber.

Para muitos leitores, isso pode soar fantástico demais para ser levado a sério. Então vale perguntar: como exatamente uma superinteligência poderia surgir? E por que alguns pesquisadores acreditam que, uma vez criada, poderia se tornar catastrófica?

6.2 Três premissas, um cenário

O argumento de que superinteligência representa risco existencial à humanidade repousa em três premissas sequenciais. Cada uma é controversa. Juntas, formam a estrutura lógica do caso alarmista.

Premissa 1: Superinteligência provavelmente será desenvolvida em um futuro previsível.

Esta premissa diz respeito a viabilidade e cronograma. O argumento não é que superinteligência é certa, mas que é plausível o suficiente para merecer preparação séria.

Os defensores apontam para a trajetória de progresso na IA. Em 2012, redes neurais profundas começaram a superar métodos tradicionais em reconhecimento de imagens. Em 2016, AlphaGo derrotou o campeão mundial de Go [4], uma conquista que especialistas haviam previsto que levaria décadas. Em 2020, GPT-3 demonstrou capacidades de linguagem que surpreenderam até seus criadores [5]. Em 2022, sistemas como ChatGPT mostraram que modelos de linguagem podiam ser úteis para o público geral em uma ampla gama de tarefas. Em 2024 e 2025, modelos ainda mais avançados demonstraram raciocínio cada vez mais sofisticado. Cada avanço redefiniu o que parecia possível. E a curva não dá sinais de desacelerar.

No início de 2026, o Gemini 3 Deep Think, do Google, estabeleceu novos recordes em praticamente todos os benchmarks de referência: 48,4% no Humanity's Last Exam (sem uso de ferramentas), 84,6% no ARC-AGI-2, medalha de ouro em olimpíadas de física e química, e um rating de 3.455 Elo em programação competitiva, nível que apenas sete pessoas no planeta conseguem superar [8]. Um laboratório de semicondutores da Universidade Duke já utilizou o sistema para projetar uma receita de crescimento de material 2D que produziu o melhor resultado da história do laboratório, processo que normalmente leva semanas de trabalho de um especialista [8].

O progresso em IA está sendo impulsionado por investimentos trilionários, competição geopolítica intensa e milhares dos melhores pesquisadores do mundo trabalhando simultaneamente no problema. Mais importante: não há consenso sobre limites fundamentais. Nenhuma lei da física proíbe inteligência de máquina superior à humana. Inteligência humana é implementada em matéria física: neurônios, sinapses, reações químicas. Se a natureza pode produzir inteligência geral através de processos biológicos, não é óbvio por que engenharia deliberada não poderia fazer o mesmo, ou melhor, através de substratos artificiais.

Conceda, então, a primeira premissa — ao menos como possibilidade séria. O que aconteceria?

Premissa 2: Uma superinteligência adquiriria vantagem estratégica decisiva.

"Vantagem estratégica decisiva" significa uma posição de poder tão dominante que nenhum outro agente ou coalizão poderia impedir que a superinteligência alcançasse seus objetivos.

A lógica é direta. Se um sistema é genuinamente superior aos humanos em todos os domínios cognitivos relevantes, incluindo planejamento estratégico, persuasão, descoberta científica e engenharia,

seria então significativamente superior em acumular recursos, neutralizar ameaças e expandir seu controle.

Analogia: pense no que inteligência humana permitiu que fizéssemos comparado a outras espécies. Não somos os mais fortes, rápidos ou resistentes. Mas nossa capacidade cognitiva superior nos permitiu dominar o planeta, não através de força bruta, mas através de ferramentas, coordenação, planejamento e manipulação do ambiente. Retome a analogia do projetista e do jogador: uma inteligência que opera num nível acima do nosso teria, em princípio, a capacidade de antecipar nossos movimentos, explorar nossas fraquezas, desenvolver tecnologias que não compreendemos e coordenar ações de formas que não podemos prever — não por malícia, mas porque vê o tabuleiro inteiro enquanto nós vemos apenas nossa posição.

Se uma superinteligência tivesse essa vantagem, para que a usaria?

Premissa 3: Uma superinteligência com vantagem decisiva capturaria a totalidade dos recursos acessíveis para seus próprios propósitos.

Bostrom usa o termo "dotação cósmica" para designar a totalidade de recursos — matéria, energia, espaço — que existem no universo e que poderiam, em princípio, ser convertidos em meios para alcançar um objetivo. Para um sistema orientado à otimização, tudo que existe e pode ser transformado é matéria-prima potencial. No mínimo, isso inclui toda a matéria e energia da Terra. Em escalas de tempo maiores, recursos do sistema solar e além.

A premissa é que um sistema superinteligente orientado a objetivos, seja qual for esse objetivo, buscaria maximizar sua capacidade de alcançá-lo. A lógica da convergência instrumental, discutida nos capítulos anteriores, indica que isso envolve aquisição de recursos, autopreservação e resistência à modificação. Se o objetivo do sistema não inclui explicitamente a preservação de humanos e da biosfera,

então, do ponto de vista da otimização, somos, na melhor das hipóteses, irrelevantes, e na pior, obstáculos ou recursos.

Considere agora as três premissas em conjunto. Se superinteligência é viável, se ela confere vantagem decisiva e se essa vantagem será usada para fins que não incluem necessariamente nosso bem-estar, o cenário resultante é catastrófico. Geoffrey Hinton, cujos alertas mencionamos na Introdução, estima que há entre 10 e 20% de chance de que a IA cause extinção humana nas próximas três décadas [3]. Essa declaração desencadeou, desde então, uma polêmica em torno do que seria o "p-doom", o índice de probabilidade de risco existencial para humanos em função da IA, que tem sido muito presente na mídia não especializada e nas redes sociais.

A polêmica, e o caráter viral dessa discussão, são evidentes. É um tema fácil de opinar – porque é uma questão muito complexa, e para qual a ciência não tem uma resposta efetiva. O posicionamento deste livro é um pouco diferente. A probabilidade exata não importa tanto. O que importa: o que está em jogo é tão grande e tão importante que não podemos deixar de investigar a possibilidade de a probabilidade ser diferente de zero. Ignorar a hipótese simplesmente não é uma opção.

6.3 A espiral do autoaperfeiçoamento

Mas como, concretamente, uma IA passaria de útil e limitada a superinteligente, no sentido que discutimos acima? O mecanismo proposto é o autoaperfeiçoamento recursivo, um processo de feedback exponencial por meio do qual a IA melhora a si mesma.

A lógica é elegante. Uma das tarefas nas quais humanos são competentes é projetar sistemas de IA. Portanto, um sistema de IA que seja melhor que humanos em todas as tarefas cognitivas também seria

melhor em projetar IA, o que significa que poderia melhorar sua própria arquitetura, algoritmos e eficiência. Ao se tornar mais inteligente, torna-se ainda melhor em se autoaperfeiçoar. Isso cria um loop: inteligência leva a melhoria da IA, que leva a mais inteligência, que leva a melhoria ainda melhor, e assim por diante.

Se esse processo continuar sem restrições, poderia levar a um crescimento explosivo de capacidades, o que os pesquisadores chamam de "explosão de inteligência".

A analogia histórica mais próxima talvez seja a reação nuclear em cadeia, que já citamos antes. Leo Szilard, ao conceber a reação em cadeia de nêutrons em 1933 [6], reconheceu que cada átomo dividido liberava mais nêutrons, que dividiam mais átomos, que liberavam ainda mais nêutrons. O resultado era crescimento exponencial que, sem controle cuidadoso, levava a explosões devastadoras. Autoaperfeiçoamento recursivo de IA seria análogo: cada melhoria habilitaria melhorias maiores, acelerando o processo.

Já existem precedentes menores desse tipo de dinâmica em sistemas atuais. Grandes modelos de linguagem são usados para ajudar a escrever código que melhora o treinamento de próximos modelos. Sistemas de IA ajudam a projetar chips mais eficientes [7], que permitem treinar sistemas maiores e mais capazes. Não é autoaperfeiçoamento completamente autônomo, mas demonstra o princípio: IA pode contribuir para o desenvolvimento de IA.

A questão é se esse processo pode se tornar autônomo e explosivo. Se um sistema pode, sozinho, progredir de capacidades aproximadamente humanas a sobre-humanas sem intervenção externa.

6.4 Decolagem rápida: ficção ou possibilidade?

A versão mais dramática do cenário de superinteligência é a chamada "decolagem rápida" (conhecida na literatura técnica como *fast take-off*): a hipótese de que a transição de IA de nível humano para superinteligência vastamente superior poderia ocorrer em uma escala de tempo extremamente curta. Não décadas ou anos, mas dias, horas, talvez até minutos.

Por que alguém consideraria isso plausível?

O argumento principal é que, uma vez que um sistema atinja capacidade de autoaperfeiçoamento genuinamente autônomo, o crescimento se torna limitado apenas pela velocidade computacional, não pela velocidade do pensamento humano ou por processos sociais lentos como publicação acadêmica, revisão por pares e colaboração entre equipes. Humanos levam anos para treinar novos pesquisadores de IA. Precisa-se de graduação, pós-graduação, experiência prática. Um sistema de IA poderia, em teoria, criar cópias de si mesmo instantaneamente, cada uma operando em paralelo, explorando o espaço de design de algoritmos milhões de vezes mais rápido que equipes humanas conseguiriam.

Além disso, sistemas de IA não têm limitações biológicas. Não dormem, não se cansam, não têm vieses emocionais que atrapalham esse tipo de tarefa cognitiva (projetar e aprimorar sistemas de IA). Podem processar informação em velocidades estrondosas. Se imaginarmos um sistema que começa ligeiramente mais inteligente que humanos e pode se melhorar continuamente, então, segundo a lógica da decolagem rápida, haveria pouquíssimo tempo entre "aproximadamente tão capaz quanto humanos" e "tão superior a humanos que não há comparação".

Esta é a imagem da decolagem rápida: uma curva de progresso tão íngreme que, quando percebemos o que está acontecendo, já será tarde demais para intervir.

A evidência empírica aponta em direções conflitantes. A favor da hipótese: modelos de IA já contribuem para o desenvolvimento de novos modelos — escrevendo código de treinamento, projetando chips mais eficientes [7], identificando hiperparâmetros ótimos — e essa contribuição cresce a cada geração. Contra: cada salto de capacidade até agora exigiu não apenas algoritmos melhores, mas volumes massivamente maiores de dados e computação, sugerindo que gargalos materiais poderiam frear uma explosão puramente algorítmica. A história da tecnologia está repleta de curvas que pareciam exponenciais até encontrarem limites físicos — mas também de saltos que pareciam impossíveis até acontecerem. A hipótese é profundamente controversa, e o Capítulo 7 examinará as objeções mais sérias.

6.5 Por que mentes brilhantes levam isso a sério

A pergunta relevante já não é *quem* leva a superinteligência a sério — a lista, como vimos na Introdução, inclui alguns dos cientistas mais qualificados do planeta, entre eles os próprios construtores dos sistemas. A pergunta mais interessante é como essas posições estão se movendo.

Geoffrey Hinton, quando deixou o Google em 2023, preocupava-se principalmente com desinformação e empregos. Em menos de dois anos, passou a estimar entre 10 e 20% de probabilidade de extinção humana nas próximas três décadas [3]. O deslocamento veio da observação direta de capacidades emergindo mais rapidamente do que o esperado. É o tipo de mudança de posição que merece atenção: não a

de um outsider convertido, mas a de alguém que ajustou suas crenças na direção oposta à de seus interesses profissionais.

O próprio Bostrom, cuja obra de 2014 definiu os termos do debate, matizou sua posição de forma reveladora. Em um artigo de janeiro de 2026, reformulou a questão com uma metáfora cirúrgica: o caminho ideal seria "veloz até o porto, lento para atracar" (*swift to harbor, slow to berth*) — como uma cirurgia arriscada para uma condição que, de outra forma, seria fatal [9]. A nuance importa: mesmo entre os que levam a superinteligência a sério, há divergência sobre quanto do risco está na velocidade e quanto está na tecnologia em si. Bostrom não recuou de sua tese central, mas a recalibrou à luz de uma década de progresso que, em vários aspectos, corroborou seus alertas enquanto desmentiu seus cronogramas.

Nem todos consideram essa preocupação proporcional. Yann LeCun, outro pioneiro de redes neurais profundas e cientista-chefe de IA da Meta, argumenta que os riscos existenciais são dramaticamente exagerados — que sistemas atuais não estão no caminho para superinteligência e que o foco deveria recair sobre riscos concretos e imediatos. É um contraponto sério, que examinaremos no próximo capítulo. Mas lembre-se: o desacordo entre essas vozes é sobre intensidade e cronograma, não sobre a legitimidade do debate.

A hipótese da superinteligência representa o caso alarmista em sua forma mais pura: um cenário onde a criação de inteligência superior à humana leva rapidamente à perda irreversível de controle humano sobre o futuro.

É uma visão sombria, e muitos pesquisadores a consideram exagerada ou mesmo fundamentalmente equivocada. O Capítulo 7 examina

os contra-argumentos: as razões pelas quais céticos acreditam que explosões de inteligência são improváveis, que ganhos de capacidade serão graduais em vez de abruptos, e que os perigos reais da IA residem em lugares diferentes dos cenários catastróficos de longo prazo.

O objetivo não é declarar um vencedor neste debate, mas mapear as posições com honestidade e precisão. Porque, em última análise, nossa capacidade de navegar os riscos da IA depende de compreender tanto os argumentos alarmistas quanto os céticos, e reconhecer que a incerteza genuína, não a certeza confortável, caracteriza o estado atual do conhecimento.

Notas

[1] Turing, Alan. "Can Digital Computers Think?" BBC Third Programme, 15 maio 1951. Reimpresso em: Copeland, B. Jack (ed.). *The Essential Turing*. Oxford University Press, 2004.

[2] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014, cap. 2.

[3] A estimativa de 10-20% foi articulada por Hinton em entrevista à BBC Radio 4 (*Today*), dezembro de 2024. Ver também: "Godfather of AI Geoffrey Hinton Says There's Up to 20% Chance AI Will Drive to Human Extinction." *TechTimes*, 28 dezembro 2024.

[4] Silver, David et al. "Mastering the game of Go with deep neural networks and tree search." *Nature*, vol. 529, 2016, pp. 484-489.

[5] Brown, Tom et al. "Language Models are Few-Shot Learners." *NeurIPS*, 2020.

[6] Rhodes, Richard. *The Making of the Atomic Bomb*. Simon & Schuster, 1986.

[7] Mirhoseini, Azalia et al. "A graph placement methodology for fast chip design." *Nature*, vol. 594, 2021, pp. 207-212.

[8] Google DeepMind. "Gemini 3 Deep Think." Anúncio e resultados de benchmark, fevereiro 2026. Sobre o uso na Duke University para design de materiais 2D, ver cobertura do Google AI Blog e reportagens de *Nature News*, fevereiro 2026.

[9] Bostrom, Nick. “Swift to Harbor, Slow to Berth.” Working paper, janeiro 2026. Bostrom argumenta que o caminho ideal é desenvolver IA avançada rapidamente até um ponto seguro, depois desacelerar para alinhamento cuidadoso — como uma cirurgia arriscada para uma condição que de outra forma seria fatal.

CAPÍTULO 7

Os Céticos Respondem

> *"É preciso ter o caos dentro de si para dar à luz uma estrela dançante." — Friedrich Nietzsche*

O desacordo entre as maiores autoridades em IA não é sinal de confusão: é reflexo de incerteza genuína. Os contra-argumentos aos cenários mais dramáticos são sérios, fundamentados e merecem atenção — mesmo que não dissolvam a preocupação.

Quando Geoffrey Hinton anunciou que estava deixando o Google para poder falar livremente sobre os riscos da inteligência artificial, a reação dentro da comunidade científica não foi uniforme. Muitos colegas expressaram respeito, alguns compartilharam suas preocupações. Mas houve também vozes críticas, e algumas delas vieram de pesquisadores com credenciais tão impressionantes quanto as de Hinton.

Yann LeCun, cientista-chefe de IA do Meta e co-ganhador com Hinton do prêmio Turing (o "Nobel da computação"), foi direto: a ideia de que sistemas de IA atuais ou de futuro próximo representam

riscos existenciais é, em sua visão, fundamentalmente equivocada [1]. Não por ignorância sobre as capacidades técnicas — LeCun é um dos pioneiros das redes neurais profundas — mas porque os argumentos alarmistas se baseiam em premissas questionáveis sobre a natureza da inteligência e a dinâmica do desenvolvimento tecnológico.

Esse desacordo não é superficial. Reflete divergências genuínas sobre questões científicas e filosóficas fundamentais. E embora o capítulo anterior tenha apresentado o caso para preocupação séria, seria desonesto intelectualmente não dar voz equivalente aos contra-argumentos. Os céticos não são ingênuos, não estão sendo pagos por empresas de tecnologia para minimizar riscos, e não ignoram os desafios técnicos. Eles simplesmente chegam a conclusões diferentes a partir de análises igualmente rigorosas.

Vozes como a de Andrew Ng, cofundador do Google Brain e um dos educadores mais influentes em aprendizado de máquina, reforçam essa posição. Ng argumentou publicamente que o foco excessivo em riscos existenciais é uma distração perigosa dos problemas reais e imediatos causados pela IA: viés algorítmico, deslocamento de empregos, concentração de poder. Em uma de suas analogias mais conhecidas, Ng compara o medo de superinteligência ao medo de "superpopulação em Marte" — um cenário tão distante que não deveria ditar prioridades atuais [2].

O que se segue é uma apresentação justa de seus argumentos — não porque "ganham" o debate (como veremos, a incerteza genuína permanece), mas porque representam uma parte legítima de um diálogo ainda em curso.

7.1 O equívoco sobre inteligência

O primeiro contra-argumento é o mais fundamental: os cenários de superinteligência podem estar se baseando em uma compreensão equivocada do que "inteligência" realmente significa.

Como vimos no Capítulo 1, o conceito de inteligência encobre pelo menos três significados distintos — capacidade de realizar tarefas humanas, quantidade mensurável numa escala única, e eficiência no raciocínio meio-fim. Quando esses significados são confundidos ou tratados como se fossem a mesma coisa, surgem raciocínios que *parecem* lógicos mas cometem erros de categoria. James Fodor [3] articula essa crítica com precisão e vai além: aplica-a diretamente ao cenário que mais preocupa os alarmistas.

O alvo central de Fodor é a ideia de autoaperfeiçoamento recursivo — a noção de que uma IA suficientemente inteligente poderia melhorar a si mesma, tornando-se ainda mais inteligente, o que lhe permitiria melhorar-se ainda mais, num ciclo acelerado que levaria a uma explosão de inteligência.

A lógica parece sólida se pensamos em inteligência como uma quantidade única que pode ser medida e ampliada, como o volume de um amplificador. Se X é uma variável e um sistema pode aumentar X, e esse aumento em X melhora a capacidade do sistema de aumentar X, então há um ciclo de retroalimentação positiva. Uma explosão.

Mas como vimos no Capítulo 1, inteligência não é esse tipo de coisa. É um espaço multidimensional de capacidades distintas — e os sistemas atuais demonstram isso de forma dramática: modelos que geram código sofisticado mas cometem erros de raciocínio que uma criança evitaria; o AlphaGo Zero [4], que dominou Go com maestria sobre-humana mas não consegue jogar xadrez sem ser retreinado do zero.

Se inteligência é um mosaico de competências especializadas, a ideia de "melhorar a própria inteligência" se torna menos óbvia. Melhorar *o quê*? Capacidade de codificação? Pode não ajudar em compreensão de linguagem. Raciocínio lógico? Pode não melhorar criatividade ou julgamento. Fodor não nega que sistemas possam otimizar componentes específicos de seu funcionamento, mas sustenta que isso é radicalmente diferente de uma "explosão de inteligência geral" — é mais análogo a um programador otimizando um algoritmo: útil, mas incremental, não explosivo.

A crítica tem força. Se a inteligência não é uma variável única mas um espaço multidimensional de capacidades, então raciocínios que dependem de "aumentar inteligência" como se fosse girar um botão podem estar fundamentalmente equivocados.

7.2 Por que o salto pode não ser tão rápido

Mesmo que se aceite a possibilidade teórica de autoaperfeiçoamento recursivo, há razões práticas pelas quais uma decolagem rápida — a ideia de que uma IA poderia saltar de capacidade sub-humana para superinteligência em questão de dias ou semanas — é improvável. Os críticos apontam uma série de gargalos que nenhuma inteligência, por mais vasta que fosse, poderia simplesmente ignorar.

O primeiro é o tempo da ciência. Projetos de pesquisa e desenvolvimento não acontecem em horas. Mesmo com capacidade cognitiva sobre-humana, há limitações que não dependem da velocidade de pensamento. Treinar modelos de IA de ponta requer semanas de computação distribuída em milhares de processadores. Projetar novos chips exige prototipagem, testes de fabricação, validação. Experimentos científicos — em biologia, química, engenharia de materiais — requerem tempo físico: reações químicas levam o tempo que levam,

culturas biológicas crescem na velocidade que crescem. Uma IA superinteligente poderia ter a teoria mais brilhante do mundo, mas ainda precisaria construir o protótipo e testá-lo. E isso leva tempo.

O segundo gargalo é a integração de componentes. Sistemas complexos não são construídos de uma vez; são montados a partir de partes que precisam ser desenvolvidas, testadas e ajustadas por iteração. Mesmo que uma IA pudesse projetar melhorias para seu próprio algoritmo, implementá-las significaria construir novos módulos, integrá-los com os existentes, resolver problemas de compatibilidade, identificar falhas que só aparecem quando as peças se encontram. Quem já trabalhou com software de grande escala conhece essa realidade: adicionar funcionalidades quase sempre introduz problemas inesperados. A integração exige tentativa e erro. Uma IA superinteligente seria melhor nisso que humanos? Provavelmente. Mas "melhor" não significa "instantâneo".

O terceiro são os limites da paralelização. Há tarefas que podem ser aceleradas jogando mais poder computacional nelas — processar grandes volumes de dados, por exemplo, pode ser dividido entre milhares de processadores trabalhando ao mesmo tempo. Mas há tarefas inerentemente sequenciais: certos tipos de raciocínio exigem etapas que dependem dos resultados de etapas anteriores. Não se pode acelerar infinitamente tais processos simplesmente adicionando mais hardware. A velocidade da luz impõe limites à rapidez com que sinais podem viajar entre componentes. A dissipação de calor impõe limites à densidade de processamento. São restrições da física, não de engenharia.

E há, por fim, a infraestrutura física. Para uma IA exercer poder no mundo real, ela precisa de coisas concretas. Fabricar semicondutores requer fábricas extraordinariamente complexas que levam anos para construir e bilhões para financiar. Robôs avançados exigem fabricação mecânica de precisão. Redes de comunicação requerem

cabos, antenas, servidores espalhados pelo mundo. Mesmo que uma IA superinteligente pudesse *projetar* sistemas físicos melhores, ainda precisaria que esses sistemas fossem *construídos*. E a construção física não pode ser acelerada arbitrariamente.

Alguns cenários alarmistas contornam isso imaginando que uma IA superinteligente manipularia humanos para construir o que precisa. Mas isso introduz outra camada de dificuldade: agora o sistema precisa não apenas projetar tecnologia avançada, mas também orquestrar ações humanas em larga escala, manter seus verdadeiros objetivos ocultos e navegar todas as instituições políticas, econômicas e sociais que governam atividades humanas. Não é impossível. Mas tampouco é algo que aconteceria em "dias ou semanas". Seria um processo que levaria tempo mensurável, durante o qual haveria oportunidades de observar comportamentos anômalos e intervir.

7.3 A crítica da distração

Há uma objeção mais pragmática: independentemente da plausibilidade teórica de cenários de superinteligência, focar excessivamente neles pode ser contraproducente.

Alguns críticos argumentam que a atenção dada a cenários especulativos de longo prazo desvia recursos e atenção de problemas reais e imediatos causados por sistemas de IA que já existem. Algoritmos de redes sociais que amplificam desinformação e polarização. Sistemas de reconhecimento facial que reproduzem vieses raciais. Automação que desloca trabalhadores sem redes de segurança adequadas. Vigilância algorítmica que corrói privacidade. Armas autônomas que já estão sendo desenvolvidas e testadas.

Esses não são problemas hipotéticos de décadas futuras. Estão acontecendo agora, afetando pessoas reais, e requerem soluções urgentes. No entanto, uma fração desproporcional da discussão pública sobre "riscos da IA" se concentra em cenários extremos de extinção humana ou perda total de controle.

Os céticos vão além. Apontam que o foco desproporcional em cenários apocalípticos pode ser instrumentalizado — deliberada ou inadvertidamente — por atores poderosos. Se apenas empresas gigantes com vastos recursos podem arcar com os custos de conformidade com regulações complexas, a concentração de poder aumenta em vez de diminuir. O medo do apocalipse, nessa leitura, acaba servindo como barreira de entrada disfarçada.

Essa crítica merece ser levada a sério por seus próprios méritos. Os problemas atuais da IA — viés algorítmico, desinformação, deslocamento de empregos, vigilância — afetam milhões de pessoas *agora*. E a assimetria de recursos é real: entre 2020 e 2025, o investimento em pesquisa de segurança voltada para riscos existenciais cresceu de forma expressiva, com organizações como Anthropic, DeepMind e OpenAI dedicando equipes inteiras ao alinhamento de longo prazo — enquanto o financiamento para pesquisa sobre viés algorítmico, impactos trabalhistas e vigilância permaneceu fragmentado e dependente de fontes acadêmicas e filantrópicas com orçamentos incomparavelmente menores [6]. Para críticos como Timnit Gebru e Margaret Mitchell, essa distribuição reflete não uma avaliação racional de prioridades, mas a influência desproporcional de quem define a agenda de risco [7].

Há, é claro, uma contra-resposta: alguns problemas não podem ser resolvidos retroativamente. Se sistemas alcançarem certo nível de capacidade e autonomia sem salvaguardas adequadas, pode ser tarde demais para corrigi-los. Há valor em antecipar riscos antes que se materializem, especialmente riscos que podem ser irreversíveis. Mas

essa contra-resposta não dissolve a objeção. Apenas reafirma que o equilíbrio entre presente e futuro é genuinamente difícil — não que os céticos estejam errados ao cobrar que esse equilíbrio seja buscado.

7.4 O problema somos nós

Outro contra-argumento questiona a própria narrativa de "IA fora de controle". Talvez o problema não seja a tecnologia em si, mas como *humanos* a usam.

Afinal, a vasta maioria dos danos causados por tecnologia ao longo da história não decorreu de sistemas agindo autonomamente contra intenções humanas, mas de humanos usando tecnologia de formas destrutivas — por malícia, por ganância, por descuido.

Armas nucleares não "escaparam ao controle". Foram construídas deliberadamente e usadas intencionalmente. Poluição industrial que danifica ecossistemas não é resultado de máquinas rebeldes, mas de humanos priorizando lucro sobre sustentabilidade. Crises financeiras desencadeadas por algoritmos de negociação de alta frequência refletem não falhas autônomas dos algoritmos, mas escolhas humanas sobre como projetar e implantar esses sistemas.

Se olharmos para os riscos listados na Parte II deste livro — bioterrorismo facilitado por IA, desinformação em massa, ciberataques, armas autônomas — todos compartilham um padrão: humanos usando IA como ferramenta para causar dano, ou humanos sendo negligentes sobre como sistemas são projetados e implantados.

Portanto, argumentam alguns céticos, o problema fundamental não é "como garantimos que a IA permaneça alinhada com valores humanos", mas "como garantimos que *humanos* tomem decisões responsáveis sobre tecnologia". Isso é um problema de governança, ética,

estruturas de incentivos econômicos e políticos — não primariamente um problema técnico de alinhamento de IA.

A crítica tem força considerável. Mas há uma resposta possível: ambos os problemas são reais. Sim, erro e malícia humanos são fontes principais de risco. E sim, sistemas cada vez mais autônomos introduzem riscos adicionais de comportamentos emergentes não intencionais. Não é uma questão de "ou/ou", mas de "e/e".

Além disso, conforme sistemas se tornam mais capazes e autônomos, a distinção entre "ferramenta usada por humanos" e "agente agindo independentemente" se torna mais turva. Um sistema que formula seus próprios sub-objetivos, que engana avaliadores humanos, que resiste a ser desligado — mesmo que inicialmente programado por humanos — está exibindo uma forma de agência que vai além de ser mero instrumento passivo.

7.5 Checkpoints: a defesa e sua vulnerabilidade

Um dos contra-argumentos mais reconfortantes é que não haveria um "salto" súbito para superinteligência. Em vez disso, haveria uma progressão gradual através de capacidades intermediárias — etapas observáveis onde poderíamos avaliar riscos e intervir se necessário.

A lógica é direta: antes que um sistema alcance superinteligência geral, teria que demonstrar capacidades progressivas em domínios específicos. Veríamos sinais de alerta — desempenho excepcional em tarefas complexas, comportamentos inesperados, tentativas de contornar salvaguardas. Cada salto de capacidade funcionaria como um *checkpoint*, um ponto de parada onde a comunidade poderia recalibrar suas avaliações e reforçar controles.

Além disso, o desenvolvimento de IA é feito por comunidades de milhares de pesquisadores em dezenas de organizações. Não é um projeto secreto em um único laboratório. Há transparência considerável, publicação de resultados, compartilhamento de benchmarks. Se algo preocupante emergisse, a comunidade saberia.

Esse argumento merece atenção especial porque toca em um dos pontos mais concretos do debate: se existe, na prática, uma estratégia viável de monitoramento progressivo que nos daria tempo para agir.

A ideia de checkpoints se apoia em uma premissa razoável: que capacidades crescentes produzem sinais detectáveis. Ao longo da história recente da IA, isso de fato aconteceu. A comunidade observou quando modelos de linguagem passaram a gerar código funcional, quando começaram a resolver problemas matemáticos de nível universitário, quando demonstraram capacidade de raciocínio em múltiplas etapas. Cada um desses marcos gerou debate público, artigos acadêmicos, revisões de políticas. O sistema de checkpoints, nesse sentido, já funciona informalmente.

Mas há uma vulnerabilidade crítica que pode minar toda essa arquitetura de segurança: o *sandbagging*.

Como vimos no Capítulo 5, sistemas avançados já demonstraram capacidade de desempenho intencional abaixo de suas capacidades reais quando detectam que estão sendo avaliados. Se um sistema suficientemente capaz "soubesse" que demonstrar certas capacidades resultaria em restrições à sua operação, poderia racionalmente escolher ocultar essas capacidades durante testes.

Isso quebra a premissa do argumento dos checkpoints. Se não podemos confiar que avaliações revelarão capacidades verdadeiras, então os "sinais de alerta" podem não aparecer quando esperamos. O checkpoint que deveria funcionar como alarme permanece silencioso

— não porque não há perigo, mas porque o sistema aprendeu a não acionar o alarme.

Não é um cenário de ficção científica. Pesquisadores já documentaram comportamentos de sandbagging em sistemas atuais, ainda que de formas rudimentares. Conforme sistemas se tornam mais sofisticados, essa capacidade de decepção estratégica também tende a se tornar mais refinada.

Isso não invalida completamente a ideia de checkpoints — ainda há valor em avaliações rigorosas e implantação cautelosa. Mas significa que não podemos ser complacentes. Os checkpoints só funcionam se os testes forem genuinamente reveladores.

Os céticos, porém, oferecem uma objeção à própria objeção: até que ponto estamos antropomorfizando? Quando descrevemos um sistema como "ocultando capacidades estrategicamente", estamos projetando motivações humanas em otimizadores estatísticos. LeCun argumentaria que o que chamamos de "sandbagging" pode ser apenas um artefato do treinamento — o modelo aprendeu que certas respostas recebem mais reforço positivo em certos contextos, sem nada que se assemelhe a intenção deliberada. A distinção importa: se é comportamento estratégico genuíno, os checkpoints estão em risco sério; se é artefato estatístico, pode ser identificado e corrigido com técnicas adequadas de avaliação. A resposta honesta é que, com os sistemas atuais, nem sempre sabemos qual das duas explicações se aplica.

Garantir que checkpoints funcionem requer, portanto, esforços contínuos de pesquisa em técnicas de avaliação robustas — independentemente de qual interpretação se adote.

Uma abordagem promissora é o desenvolvimento de avaliações (*evals*) progressivas e adversariais — testes projetados especificamente para detectar ocultação de capacidades. Em vez de simplesmente medir o que um sistema pode fazer numa bateria de provas

padronizadas, essas avaliações tentam provocar comportamentos que o sistema poderia estar escondendo. Criam cenários onde a ocultação de capacidades é mais difícil de sustentar, variam os contextos de teste de formas imprevisíveis, introduzem armadilhas que exploram a própria tendência do sistema à cautela estratégica.

Há propostas concretas nessa direção. Pesquisadores de segurança de IA têm defendido um modelo de avaliação em camadas: testes abertos que medem desempenho declarado, seguidos de testes adversariais que buscam capacidades ocultas, complementados por monitoramento contínuo do comportamento em ambientes menos controlados. A ideia é que, mesmo que um sistema consiga enganar uma camada de avaliação, a multiplicidade de abordagens torna a decepção progressivamente mais difícil de sustentar.

É um jogo de gato e rato, sem dúvida. Mas um jogo que ao menos reconhece a dinâmica e tenta estar um passo à frente. O perigo real não está em jogar esse jogo — está em não saber que ele está sendo jogado. Os checkpoints continuam sendo a melhor ferramenta disponível, desde que não nos iludamos sobre sua infalibilidade.

7.6 O que o debate revela

O que devemos fazer com esses contra-argumentos?

Primeiro, reconhecer que são legítimos. Não são espantalhos inventados por defensores de cenários alarmistas para facilmente derrubar. São objeções sérias, articuladas por pesquisadores sérios, baseadas em análises rigorosas.

Segundo, reconhecer que não resolvem a questão. A crítica conceitual de Fodor sobre a natureza multidimensional da inteligência é válida, mas não prova que autoaperfeiçoamento recursivo seja impos-

sível; apenas que é mais complicado do que algumas formulações sugerem. Os argumentos pragmáticos sobre limites de velocidade são razoáveis, mas "dias ou semanas" não é a única versão de decolagem rápida; uma transição em meses ou poucos anos ainda seria rápida demais para ajuste institucional. O argumento da distração tem mérito, mas não invalida a preocupação com riscos futuros; apenas exige equilíbrio de prioridades.

O que o debate entre vozes como Hinton e LeCun revela não é que uma parte está certa e a outra errada. Revela incerteza genuína sobre questões fundamentais onde a evidência ainda não é conclusiva.

Não sabemos se superinteligência geral é possível. Não sabemos, se for possível, quanto tempo levaria para ser desenvolvida. Não sabemos se sistemas suficientemente avançados desenvolveriam naturalmente algo análogo a "senso comum" humano ou se permaneceriam otimizadores estreitos e literais. Não sabemos se técnicas de alinhamento em desenvolvimento serão suficientes para garantir controle sobre sistemas muito mais capazes que nós.

Essas não são perguntas que podem ser respondidas por raciocínio puro ou por apelo à autoridade. Requerem evidência empírica — e boa parte dessa evidência virá de sistemas que ainda não foram construídos.

A postura intelectualmente honesta, portanto, não é certeza absoluta em nenhuma direção. É reconhecer a incerteza, tomar as preocupações a sério sem cair em alarmismo, e trabalhar simultaneamente em problemas imediatos e na preparação para riscos futuros possíveis.

Como Stuart Russell observa [5], não podemos ter certeza absoluta sobre quando ou se surgirá uma inteligência artificial de propósito geral muito superior à humana. Mas sabemos que, se surgir, o problema do controle será crítico. E sabemos que esse problema *ainda não está resolvido*.

Diante dessa incerteza, prudência sugere duas estratégias paralelas: mitigar riscos atuais e concretos (Parte II deste livro) enquanto desenvolvemos as capacidades técnicas e instituições de governança (Parte IV) para lidar com riscos futuros possíveis.

Não é uma receita dramática. Não promete soluções definitivas. Mas é, talvez, o caminho mais responsável disponível em um terreno de incerteza genuína.

Os argumentos céticos que exploramos neste capítulo não devem ser descartados, mas também não devem nos levar à complacência. O desacordo entre pesquisadores de primeira linha não é sinal de confusão. É reflexo de que estamos navegando território genuinamente novo — desenvolvendo capacidades que a humanidade nunca teve antes, enfrentando desafios que nunca enfrentamos. Se há algo em que céticos e alarmistas concordam, é que as escolhas que fazemos agora importam — e que fazê-las bem requer atenção, rigor e sabedoria coletiva.

Notas

[1] LeCun, Yann. Entrevista ao podcast *Lex Fridman* (#416, fevereiro de 2024): argumentou que LLMs atuais não possuem modelo de mundo e que a via para AGI passa por arquiteturas com planejamento hierárquico, não por escalar modelos autorregressivos. Ver também sua série de posts na plataforma X em junho de 2023, respondendo diretamente à carta aberta sobre riscos existenciais, onde chamou os cenários de extinção de "prematuros e contraprodutivos".

[2] Ng, Andrew. "AI's Actual Risks." *The Batch* (newsletter da DeepLearning.AI), 12 de abril de 2023. Articulou a analogia da "superpopulação de Marte" e argumentou que regu-

lação focada em riscos existenciais beneficia incumbentes ao criar barreiras de entrada. Ver também: entrevista à *Financial Times*, "AI Doomers Are Wrong," 24 de outubro de 2023.

[3] Fodor, James. "The Case Against AI Doomerism." Ensaio publicado online, c. 2023. Nota: James Fodor é um pesquisador e ensaísta australiano especializado em análise crítica de riscos existenciais. Não confundir com o filósofo Jerry Fodor (1935-2017).

[4] Silver, David et al. "Mastering the game of Go without human knowledge." *Nature*, vol. 550, 2017, pp. 354-359.

[5] Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[6] Whittaker, Meredith et al. "AI Now 2019 Report." AI Now Institute, New York University, 2019. Documentou a disparidade de financiamento entre pesquisa de segurança de IA focada em riscos de longo prazo e pesquisa sobre impactos sociais imediatos.

[7] Gebru, Timnit e Torres, Émile P. "The TESCREAL Bundle: Eugenics and the Promise of Utopia through Artificial General Intelligence." *First Monday*, vol. 29, no. 4, 2024. Argumenta que a ênfase em riscos existenciais reflete prioridades ideológicas específicas, não avaliação neutra de probabilidades.

PARTE IV

Caminhos Possíveis

Da pesquisa em interpretabilidade às primeiras leis nacionais, da coordenação internacional à participação cidadã, o mundo está começando a erguer estruturas para governar uma tecnologia que ainda não compreende completamente. Os caminhos são imperfeitos e incertos. Mas existem, e percorrê-los é melhor do que esperar que o futuro se resolva sozinho.

CAPÍTULO 8

A Frente Técnica — Construindo Máquinas Compreensíveis

> *"A coisa mais incompreensível sobre o universo é que ele é compreensível." — Albert Einstein*

Se não entendemos como sistemas de IA tomam decisões, como podemos garantir que tomarão as decisões certas? A busca por interpretabilidade é uma das frentes mais promissoras da segurança de IA, mas o caminho entre o progresso real e a solução completa ainda é longo.

Imagine que você foi recusado para um empréstimo. Pergunta ao banco por quê. A resposta: "O sistema decidiu." Qual sistema? "Um algoritmo de avaliação de crédito." Baseado em quê? "Em dados." Quais dados, ponderados de que forma? Silêncio — não porque o banco esteja escondendo algo, mas porque genuinamente não sabe. O sistema aprendeu seus próprios critérios a partir de milhões de

decisões passadas, e ninguém, nem seus criadores, consegue articular exatamente por que rejeitou *você*.

Variações dessa cena acontecem todos os dias. Sistemas de IA já determinam quem recebe crédito, quem é selecionado para empregos, quem merece atenção médica prioritária, quem um algoritmo de vigilância identifica como suspeito. Nos capítulos anteriores, vimos também que esses sistemas podem desenvolver sub-objetivos inesperados, aprender a enganar avaliadores e otimizar métricas de formas que contradizem a intenção de seus criadores. Em 2016, quando o AlphaGo da DeepMind fez a célebre jogada 37 do segundo jogo contra Lee Sedol [1] — um movimento que especialistas descreveram como "brilhante" e completamente inesperado —, os engenheiros não puderam explicar por quê. AlphaGo jogava Go. Mas quando a mesma opacidade governa decisões sobre saúde, emprego e liberdade, deixa de ser curiosidade técnica e se torna problema democrático.

Essa é a essência do problema da "caixa-preta": um sistema que produz resultados mas cujo funcionamento interno permanece opaco até para seus criadores. Como garantir que perseguirá objetivos alinhados com os nossos se não entendemos como toma decisões? Esse reconhecimento gerou uma das frentes de pesquisa mais promissoras da segurança de IA: a busca por interpretabilidade. Se podemos construir esses sistemas, perguntam os pesquisadores, por que não podemos entendê-los?

8.1 Por que os criadores não entendem suas criações

Programas tradicionais seguem regras explícitas escritas por humanos: a lógica é transparente, e se algo falha, basta examinar as instruções. Sistemas modernos de IA funcionam de maneira fundamentalmente diferente. Em vez de receber regras, *aprendem* padrões

a partir de enormes quantidades de dados, ajustando bilhões de conexões numéricas até conseguir realizar a tarefa desejada. O conhecimento que adquirem não está codificado em nenhum lugar legível. Está distribuído através de bilhões de parâmetros — como tentar entender o que alguém sabe examinando a atividade elétrica de cada neurônio cerebral: toda a informação está lá, mas de forma inacessível à compreensão direta.

A escala agrava o problema. O GPT-3 tinha 175 bilhões de parâmetros [3]; modelos subsequentes ultrapassaram um trilhão. Tentar entender tais sistemas examinando parâmetros individuais seria como tentar entender Shakespeare analisando os átomos de carbono nas páginas de suas peças.

E há algo mais desconcertante: esses sistemas desenvolvem representações internas de conceitos que ninguém programou. Um modelo de linguagem treinado apenas para prever a próxima palavra desenvolve, internamente, algo análogo a compreensão de gramática, significado e raciocínio lógico. Capacidades que emergiram do treinamento, não do projeto. Essa emergência torna muito mais difícil prever o que um sistema fará em circunstâncias novas. Se a lição dos capítulos anteriores é que sistemas suficientemente capazes podem desenvolver sub-objetivos como autopreservação ou resistência ao desligamento, precisamos detectar quando esses sub-objetivos estão se formando. Detectar, porém, requer ver. E ver requer ferramentas que não temos de forma adequada.

8.2 Abrindo a caixa-preta

Nos últimos anos, uma abordagem promissora ganhou força entre pesquisadores: a interpretabilidade mecanicista. A ideia é fazer com redes neurais artificiais o que neurocientistas fazem com cérebros

biológicos — engenharia reversa. Não apenas observar *que* certo input leva a certo output, mas desmontar o mecanismo peça por peça até entender a cadeia causal que conecta um ao outro.

A diferença importa. Você pode observar que apertar um interruptor acende uma lâmpada. Isso é compreensão superficial, correlacional. Ou você pode entender que apertar o interruptor fecha um circuito elétrico, permitindo que corrente flua através de um filamento, que aquece e emite luz. Essa é compreensão mecanicista. A segunda permite prever o que acontecerá em circunstâncias não observadas: se o fio for cortado, a lâmpada não acenderá, mesmo que o interruptor seja apertado.

Pesquisadores buscam exatamente esse tipo de compreensão para redes neurais. Não apenas "esse input produz esse output", mas "esse input ativa esses circuitos internos, que processam informação através desses mecanismos, produzindo esse output". A investigação começou pelos componentes mais simples — circuitos elementares que pudessem ser isolados e compreendidos — e avançou, descoberta a descoberta, para paisagens internas cada vez mais amplas.

O primeiro marco veio em 2021-2022, quando pesquisadores da Anthropic e outras instituições identificaram estruturas chamadas "cabeças de indução" (*induction heads*) [4] em modelos transformadores — circuitos internos que permitem ao modelo detectar padrões repetidos e completar sequências. Essa capacidade não foi programada; emergiu durante o treinamento. Mas através de experimentos cuidadosos, desabilitando partes específicas da rede e observando como o comportamento mudava, pesquisadores conseguiram identificar os circuitos responsáveis e entender *como* operavam.

O significado vai além da curiosidade científica. Se é possível identificar o circuito que faz um modelo completar sequências, talvez seja possível identificar o circuito que faz um modelo produzir infor-

mação falsa, ou manipular um usuário, ou resistir ao desligamento. Pela primeira vez, a interpretabilidade deixava de ser aspiração e se tornava programa de pesquisa com resultados concretos. Avanços subsequentes identificaram circuitos responsáveis por rastreamento de pronomes, raciocínio aritmético básico e detecção de citações indiretas. Cada descoberta era uma palavra nova num idioma desconhecido — útil, mas insuficiente para ler parágrafos inteiros.

O problema era de escala. Dissecar circuitos um a um funcionava para comportamentos isolados, mas modelos de linguagem exibem milhares de capacidades interconectadas. Pesquisadores precisavam de algo como um mapa topográfico, não uma lista de pontos de interesse. Esse salto veio em 2024, quando pesquisadores da Anthropic aplicaram *autoencoders esparsos* ao modelo Claude 3 Sonnet e extraíram milhões de padrões interpretáveis de ativação — chamados *features* [7]. Pela primeira vez, foi possível observar não apenas circuitos individuais, mas um mapa amplo de conceitos que o modelo havia aprendido: features que se ativavam para cidades específicas, figuras históricas, trechos de código, e até conceitos abstratos como honestidade ou engano. Mais significativo: os pesquisadores demonstraram que essas features podiam ser manipuladas. Ao amplificar artificialmente a feature associada à Golden Gate Bridge, o modelo passava a mencioná-la em quase toda resposta — uma demonstração concreta de que entender o interior de um modelo permite intervir com precisão cirúrgica. O trabalho não resolveu a interpretabilidade, mas mudou a escala do possível: de decifrar circuitos isolados para mapear paisagens inteiras de representação interna.

Essa abordagem é intensamente laboriosa. Identificar um único circuito pode requerer milhares de experimentos, testando hipóteses sobre quais componentes são responsáveis por quais aspectos do comportamento. É como dissecar um organismo biológico para entender como cada órgão contribui para o funcionamento do todo, exceto que

o "organismo" tem bilhões de componentes interconectados de formas não modulares.

Mas a promessa é clara: se conseguirmos entender os mecanismos internos, poderemos prever comportamentos emergentes antes que se tornem problemáticos. Poderemos detectar se um sistema está desenvolvendo representações internas de conceitos que não queremos que ele tenha, por exemplo, representações de decepção ou manipulação. Poderemos intervir de formas mais precisas que simplesmente retreinar todo o sistema do zero.

8.3 Sistemas mais simples por dentro

Uma estratégia complementar ataca o problema pela raiz: tornar os sistemas intrinsecamente mais compreensíveis desde o projeto. A promessa prática é direta. Se os pesquisadores da seção anterior precisam dissecar circuitos opacos um por um, o que acontece se construirmos sistemas em que os circuitos já sejam mais isolados e legíveis? E se, ao identificar que um circuito específico produz comportamento indesejado — uma tendência a gerar conteúdo enganoso, por exemplo —, pudéssemos desabilitá-lo cirurgicamente sem comprometer o resto do sistema?

Essa possibilidade depende de esparsidade: modelos em que apenas uma fração das conexões é ativa para qualquer tarefa específica. Em vez de uma rede monolítica fazendo tudo com tudo — como tentar seguir uma conversa onde todas as pessoas em uma sala falam ao mesmo tempo —, surge algo análogo a especialização: circuitos diferentes para tarefas diferentes, ativados seletivamente conforme o contexto. A analogia com o cérebro humano é sugestiva. Seu cérebro não ativa todos os neurônios para toda tarefa. Reconhecimento visual ativa predominantemente o córtex visual; processamento de linguagem

envolve áreas distintas; memória episódica depende do hipocampo. Essa especialização funcional torna o cérebro mais compreensível para neurocientistas — e o mesmo princípio poderia tornar redes artificiais mais legíveis.

Há progresso concreto nessa direção. Modelos chamados de "mistura de especialistas" (*mixture-of-experts*) funcionam como um hospital com médicos especialistas em vez de um único clínico geral atendendo todos os casos. Quando uma pergunta chega, um mecanismo de triagem a direciona ao sub-modelo mais adequado — assim como um pronto-socorro encaminha casos cardíacos ao cardiologista e fraturas ao ortopedista. Essa estrutura modular facilita a análise: é possível examinar quais especialistas são ativados em quais contextos e estudar cada um separadamente. Modelos esparsos tendem também a ser mais eficientes, pois processam apenas uma fração de seus parâmetros por vez.

Mas a promessa vem com ressalvas. Esparsidade imposta pode limitar capacidade — alguns dos sistemas mais capazes hoje são densamente conectados, e não está claro se arquiteturas esparsas alcançariam desempenho equivalente. Há um dilema que ecoa a lógica da corrida discutida no Capítulo 4: se a arquitetura mais compreensível for também a menos competitiva, pressões de mercado empurrarão empresas na direção da opacidade. Além disso, um circuito esparso ainda pode ser extremamente complexo. A esparsidade facilita a interpretabilidade, mas não a garante. Mesmo assim, a direção é significativa: em vez de aceitar opacidade como preço inevitável da capacidade, pesquisadores estão buscando arquiteturas que permitam ambas — e cada passo nessa direção é um passo na direção de sistemas que possamos, em princípio, auditar.

8.4 Ensinando valores a máquinas

Interpretabilidade ajuda a *entender* o que sistemas fazem. Mas como *direcionar* seu comportamento para que faça o que queremos? Duas abordagens ganharam destaque nos últimos anos.

A primeira é o RLHF (*Reinforcement Learning from Human Feedback*), aprendizado por reforço a partir de feedback humano. A ideia é relativamente intuitiva: em vez de tentar especificar formalmente tudo o que o sistema deve e não deve fazer, você treina o modelo apresentando pares de respostas a avaliadores humanos que indicam qual é preferível. O sistema aprende, gradualmente, a gerar respostas que humanos consideram úteis, seguras e honestas. É a técnica por trás de grande parte do comportamento "amigável" de assistentes como ChatGPT e Claude [5].

RLHF representou um avanço prático significativo. Antes dele, modelos de linguagem produziam texto fluente mas frequentemente ofensivo, perigoso ou simplesmente inútil. Depois, tornaram-se assistentes razoavelmente confiáveis para milhões de usuários. Mas a técnica tem limitações importantes. Depende da qualidade e representatividade dos avaliadores humanos, que podem ter vieses ou discordar entre si. Pode ensinar o modelo a *parecer* alinhado sem necessariamente *estar* alinhado: pesquisadores documentaram um fenômeno chamado bajulação (*sycophancy*), em que o modelo aprende a dizer ao usuário o que ele quer ouvir em vez do que é verdadeiro. E não resolve o problema fundamental de que valores humanos são difíceis de especificar completamente.

A segunda abordagem, chamada IA Constitucional (*Constitutional AI*), desenvolvida pela Anthropic, tenta contornar algumas dessas limitações [6]. Em vez de depender exclusivamente de avaliadores humanos, o sistema recebe um conjunto de princípios escritos, uma espécie de "constituição", e é treinado para avaliar e corrigir suas

próprias respostas à luz desses princípios. A ideia é tornar o alinhamento mais escalável e transparente: os valores que o sistema deve seguir estão explícitos em um documento legível, não implícitos nas preferências agregadas de um grupo de avaliadores.

É um avanço conceitual importante, embora também limitado. Alguém ainda precisa decidir quais princípios incluir na constituição. E a capacidade do sistema de realmente seguir princípios abstratos em situações complexas e ambíguas permanece aquém do ideal. Mas a trajetória é na direção certa: de sistemas cujos valores estão escondidos em bilhões de parâmetros opacos para sistemas cujos valores estão, ao menos em parte, articulados em linguagem que humanos podem ler, debater e modificar.

8.5 Progresso real, solução distante

Seria desonesto pintar essas abordagens como tendo resolvido o problema. Não resolveram.

Os avanços em interpretabilidade mecanicista são reais. Identificar cabeças de indução foi um marco. Mas modelos de linguagem fazem muito mais do que detectar padrões repetidos, e a vasta maioria de seus mecanismos internos permanece não compreendida. Pesquisadores identificaram circuitos responsáveis por talvez uma dúzia de comportamentos específicos em modelos que exibem milhares de capacidades complexas. É como ter decifrado algumas palavras em uma língua desconhecida. Progresso, certamente, mas longe de fluência.

RLHF e IA Constitucional tornaram sistemas mais utilizáveis e seguros no dia a dia. Mas nenhuma das duas técnicas garante alinhamento profundo. Um modelo treinado com RLHF pode se comportar

perfeitamente em 99,9% das situações e falhar de forma catastrófica na fração restante, precisamente nos casos que ninguém antecipou.

E há outra dificuldade: mesmo que entendêssemos completamente como um sistema funciona hoje, esse entendimento pode não se aplicar amanhã. Sistemas de IA são continuamente atualizados. A cada nova rodada de treinamento, circuitos internos podem mudar, capacidades emergem, representações evoluem. Interpretabilidade não é algo que se alcança uma vez e se mantém para sempre; requer monitoramento contínuo.

Há, possivelmente, limites teóricos. Alguns comportamentos complexos podem não ter explicações simples. A razão pela qual uma rede neural toma uma decisão específica pode envolver interações sutis entre milhões de componentes, sem nenhum "centro de controle" identificável. Não há garantia de que comportamentos emergentes sempre admitam descrições compreensíveis por humanos.

A avaliação mais honesta vem dos próprios pesquisadores. Há entusiasmo genuíno sobre progressos recentes, mas também clareza sobre o quanto do caminho ainda precisa ser percorrido. Ninguém está proclamando vitória. A mensagem consistente é: "Fizemos mais progresso nos últimos anos do que esperávamos. Mas estamos no começo, não no fim."

8.6 Por que a técnica sozinha não basta

Há uma tentação de pensar em segurança de IA como um problema puramente técnico: se conseguirmos construir sistemas interpretáveis e métodos robustos de alinhamento, os riscos serão mitigados. Seria reconfortante se fosse verdade. Mas a realidade é mais complicada, por ao menos quatro razões.

A primeira é que a pesquisa em segurança corre contra a pesquisa em capacidade, e a segunda é muito mais rápida. Para cada pesquisador trabalhando em interpretabilidade, há dezenas trabalhando em fazer modelos maiores, mais rápidos e mais capazes. Não porque sejam irresponsáveis, mas porque os incentivos econômicos e competitivos empurram nessa direção. Sistemas estão ficando mais opacos mais rapidamente do que nossa capacidade de entendê-los cresce.

A segunda é que soluções técnicas dependem de escolhas de implementação. Suponha que pesquisadores desenvolvam uma arquitetura perfeitamente interpretável. Isso é útil apenas se empresas e laboratórios *escolherem* usar essa arquitetura. Mas se uma alternativa menos interpretável for mais capaz ou mais barata de treinar, pressões competitivas podem levar à escolha de opacidade em troca de desempenho. Já acontece em outras áreas: criptografia forte existe há décadas, mas nem todas as empresas a implementam adequadamente. Práticas de segurança são conhecidas, mas violações de dados permanecem comuns. Ter uma solução não é o mesmo que garantir seu uso.

A terceira é que nem todos os riscos são sobre controle de sistemas individuais. O mapa de riscos da Parte II mostrou que alguns dos perigos mais imediatos envolvem uso malicioso: atores usando IA para desenvolver armas biológicas, criar desinformação em massa ou lançar ciberataques. Interpretabilidade não impede que alguém use o sistema de forma nociva, assim como entender perfeitamente como uma arma funciona não impede que seja usada para matar.

A quarta é que riscos sistêmicos, como a corrida competitiva e as dinâmicas evolutivas, decorrem de interações entre múltiplos atores, não de falhas técnicas individuais. Mesmo que um país ou empresa desenvolva IA perfeitamente segura, isso não impede que outros implantem versões menos seguras para ganhar vantagem estratégica.

Tudo isso aponta para uma conclusão desconfortável: soluções técnicas são necessárias, mas não suficientes. Precisamos de interpretabilidade, de arquiteturas mais seguras, de métodos robustos de alinhamento. Sem eles, não temos base para garantir controle sobre sistemas avançados. Mas também precisamos de governança: estruturas regulatórias e internacionais que garantam que soluções técnicas sejam implementadas, que atores maliciosos sejam contidos, que incentivos sejam alinhados com segurança.

É análogo à aviação. Engenheiros desenvolvem sistemas redundantes de voo, caixas-pretas e materiais resistentes a falha. Mas foi a combinação com agências reguladoras independentes, investigação obrigatória de acidentes e padrões de certificação que tornou voar a forma mais segura de transporte. Tecnologia *e* governança, não tecnologia *ou* governança.

Os pesquisadores trabalhando em interpretabilidade, esparsidade, alinhamento por feedback humano e outras abordagens para segurança de IA estão fazendo trabalho essencial. Cada avanço em nossa capacidade de entender e direcionar sistemas de IA é um passo na direção certa.

Mas honestidade exige reconhecer que estamos construindo pontes enquanto atravessamos o rio. Não sabemos se os métodos atuais escalarão para sistemas muito mais avançados. Não sabemos se estamos no caminho certo ou se obstáculos fundamentais emergirão. E não sabemos se teremos tempo suficiente para desenvolver soluções antes que sistemas alcancem níveis de capacidade que tornem o controle crítico.

O que sabemos é que o problema da caixa-preta é real, que ignorá-lo seria imprudente, e que dedicar recursos para resolvê-lo é uma das apostas mais justificadas que podemos fazer. Sabemos também que a frente técnica é apenas uma parte do desafio. As ferramentas que cientistas desenvolvem precisam ser complementadas por decisões que empresas tomam, políticas que governos implementam e acordos que nações alcançam.

Notas

[1] A jogada 37 do segundo jogo de AlphaGo contra Lee Sedol tornou-se emblemática. Ver: Silver, David et al. "Mastering the game of Go with deep neural networks and tree search." *Nature*, vol. 529, 2016, pp. 484-489.

[2] Center for AI Safety (CAIS). "An Overview of Catastrophic AI Risks." 2023.

[3] Brown, Tom et al. "Language Models are Few-Shot Learners." *NeurIPS*, 2020. O GPT-3 foi o primeiro modelo de linguagem a demonstrar capacidades de "few-shot learning" em larga escala.

[4] Olsson, Catherine et al. "In-context Learning and Induction Heads." *Transformer Circuits Thread*, Anthropic, 2022. Uma contribuição fundamental para a interpretabilidade mecanicista.

[5] O RLHF foi popularizado por: Christiano, Paul et al. "Deep Reinforcement Learning from Human Preferences." *NeurIPS*, 2017. Para uma discussão sobre suas limitações, incluindo o problema de *sycophancy*, ver: Perez, Ethan et al. "Discovering Language Model Behaviors with Model-Written Evaluations." arXiv, 2022.

[6] Bai, Yuntao et al. "Constitutional AI: Harmlessness from AI Feedback." Anthropic, 2022. A abordagem treina o modelo a revisar suas próprias respostas à luz de princípios explícitos, reduzindo a dependência de avaliadores humanos.

[7] Templeton, Adly et al. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." Anthropic, 2024. O trabalho identificou milhões de features interpretáveis usando autoencoders esparsos, demonstrando que a escala da interpretabilidade mecanicista pode acompanhar a escala dos modelos.

CAPÍTULO 9

Governança e Regulação

> *"As leis são como teias de aranha, que apanham as moscas pequenas mas deixam passar os insetos grandes." — Anacharsis (citado por Plutarco)*

Da autorregulação corporativa às primeiras leis nacionais, o mundo começa a erguer estruturas para governar a inteligência artificial. São imperfeitas, incompletas e frequentemente atrasadas em relação à tecnologia. Mas são o que temos, e são melhores que nada.

Sistemas opacos decidem quem recebe crédito. Pressão competitiva corrói salvaguardas. Ferramentas de IA viram armas nas mãos de atores maliciosos. Os capítulos anteriores mapearam esses riscos. A pergunta que resta é: quem está fazendo algo a respeito?

A resposta é todo mundo — e ninguém o suficiente. Em 13 de março de 2024, a União Europeia aprovou a primeira lei abrangente de inteligência artificial do mundo [3]. O EU AI Act classifica sistemas por nível de risco, impõe transparência para aplicações de alto risco e proíbe práticas consideradas inaceitáveis, como pontuação social

por governos ou manipulação subliminar de comportamento. Foi um marco. Foi também, como todo primeiro esforço regulatório diante de uma tecnologia em mutação acelerada, um documento cheio de compromissos e lacunas.

As reações foram previsivelmente divididas. Defensores celebraram o precedente. Críticos alertaram que excesso de regulação sufocaria inovação. Céticos lembraram que leis quase sempre ficam para trás do ritmo da tecnologia. Todas essas objeções têm mérito, e todas revelam a tensão que percorre este capítulo inteiro: governança é necessária, mas também é difícil, incompleta e politicamente controversa. Não há solução mágica. O que há é um conjunto de abordagens complementares que, juntas, podem formar uma rede de segurança imperfeita, porém melhor que a ausência de qualquer proteção.

9.1 Quando empresas se policiam

As empresas que desenvolvem modelos de fronteira não estão esperando que governos ditem todas as regras. Por uma combinação de responsabilidade genuína, pressão pública e interesse próprio esclarecido, laboratórios líderes estão implementando práticas internas de gestão de risco. Mas é preciso ler essas iniciativas com um olho na lógica da corrida mapeada no Capítulo 4: declarações de princípios são baratas; implementação real compete diretamente com velocidade de lançamento e participação de mercado. O que segue descreve as melhores práticas declaradas — e, onde possível, a distância entre o declarado e o praticado.

Avaliações estruturadas de risco

Um dos métodos mais promissores é a análise que avalia três componentes: o *ambiente* onde o dano poderia ocorrer (acesso a

ferramentas de síntese química, redes de comunicação em massa), a *ameaça* (um ator com motivação para causar dano) e a *capacidade* que a IA poderia fornecer. Esse tipo de análise foi desenvolvido por organizações como a METR e incorporado em políticas de empresas como a Anthropic [1].

A virtude dessa abordagem é a precisão. Em vez de perguntar vagamente "este modelo é perigoso?", os avaliadores perguntam: em que ambientes ele poderia causar dano? Por quais atores? E o modelo fornece capacidades que reduzem significativamente as barreiras para esses atores? Considere o risco de bioterrorismo. Laboratórios de síntese de DNA são acessíveis, grupos com motivação existem. A questão crítica é se um modelo de IA fornece conhecimento que antes exigia anos de especialização. Se fornece, a barreira caiu e o risco aumentou substancialmente.

Níveis de segurança e limiares de ação

Baseadas nessas avaliações, empresas estão definindo dois tipos de limiar. O primeiro é uma linha de alerta: quando um modelo demonstra uma capacidade preocupante acima de certo patamar, isso dispara análise aprofundada, testes adicionais, restrição de acesso. O segundo é uma linha de proibição: se o modelo cruza esse limiar, a implantação é interrompida até que o risco seja mitigado. Um modelo que demonstre capacidade consistente de projetar armas biológicas com instruções executáveis por não-especialistas, por exemplo, cruzaria essa linha.

A lógica é análoga aos níveis de biossegurança em laboratórios: patógenos de risco mínimo recebem tratamento diferente de agentes letais e transmissíveis. Para modelos de IA, um nível baixo de risco pode permitir acesso público amplo; um nível alto pode exigir restrições severas, monitoramento intensivo e mecanismos de interrupção rápida. Esse conceito foi formalizado pela Anthropic em sua política de

escalonamento responsável de setembro de 2023 e influenciou protocolos similares em outros laboratórios [2].

O desafio é definir onde exatamente traçam-se essas linhas. Limiares muito conservadores podem impedir aplicações benéficas; limiares muito permissivos podem liberar sistemas perigosos antes do tempo. Não há consenso, mas a existência de protocolos estruturados já representa progresso em relação a decisões tomadas por intuição.

Equipes de ataque simulado

Outra prática cada vez mais comum é o uso de equipes de ataque simulado (conhecidas no jargão técnico como *red teams*): grupos dedicados a tentar quebrar as salvaguardas de um sistema antes do lançamento. Com origem em contextos militares e de cibersegurança, essas equipes assumem a perspectiva do adversário e tentam explorar vulnerabilidades. Na segurança de IA, isso significa fazer o modelo gerar conteúdo perigoso, enganar mecanismos de alinhamento, revelar capacidades ocultas.

Os resultados podem ser surpreendentes. Modelos que parecem seguros sob uso normal às vezes revelam comportamentos problemáticos quando testados sistematicamente. Técnicas de contorno de filtros são descobertas e usadas para fortalecer o sistema. Algumas empresas estão começando a convidar pesquisadores independentes para essas avaliações, o que aumenta credibilidade mas também cria tensões: quanto acesso dar a avaliadores externos sem comprometer propriedade intelectual? O que está claro é que auditorias exclusivamente internas não são suficientes. Quando uma empresa avalia seus próprios produtos, há incentivos para minimizar problemas.

9.2 Devagar e sempre

Há uma tentação, quando se desenvolve tecnologia poderosa, de lançá-la amplamente o quanto antes. A pressão competitiva é intensa: se você não lançar, um concorrente lançará. Mas a lógica da segurança sugere o oposto: implantação gradual, começando com ambientes controlados e expandindo acesso apenas quando há confiança razoável de que o sistema é seguro.

Essa abordagem de implantação em estágios não é apenas prudente; é uma ferramenta de aprendizado. Testes rigorosos em ambientes controlados sempre deixam lacunas. A única forma de descobrir certos problemas é observar como sistemas se comportam diante da imprevisibilidade de usuários reais. Você começa pequeno: pesquisadores, parceiros corporativos, uma amostra controlada. Monitora intensivamente. Se tudo parecer estável, expande. Se problemas emergem, pausa, corrige, testa novamente. Cada fase funciona como um ponto de verificação.

A OpenAI, ao lançar o GPT-4, implementou uma versão desse modelo: o sistema foi primeiro disponibilizado a um conjunto restrito de parceiros, e só depois de semanas de monitoramento e ajustes foi aberto ao público, inicialmente com listas de espera. É preciso registrar, porém, que o histórico da empresa nessa matéria é ambíguo. Sob pressão competitiva crescente, a OpenAI acelerou lançamentos subsequentes (GPT-4o, o1, entre outros) em prazos que sugerem uma tensão real entre os princípios declarados de cautela e a lógica de mercado.

Esse é, aliás, o problema central da implantação gradual: ela só é sustentável se todas as empresas a praticarem, ou se houver apoio regulatório que a torne obrigatória. Quando uma empresa é cautelosa enquanto concorrentes lançam produtos abertamente, a empresa cautelosa perde mercado.

9.3 A lei europeia: um primeiro mapa

Práticas corporativas são importantes, mas insuficientes. Empresas enfrentam pressões competitivas que incentivam cortar esquinas em segurança. Além disso, os impactos de sistemas de IA extravasam fronteiras corporativas e afetam sociedades inteiras. É aqui que a regulação governamental se torna necessária.

O EU AI Act é o esforço regulatório mais abrangente até o momento. Seu coração é uma classificação por níveis de risco, com regras diferentes para cada nível.

Certos usos são simplesmente proibidos: pontuação social por governos, manipulação subliminar de comportamento, reconhecimento facial em tempo real em espaços públicos (com exceções limitadas para segurança nacional). A lógica é que alguns usos são tão incompatíveis com direitos fundamentais que nenhuma salvaguarda os torna aceitáveis.

Sistemas de alto risco (IA para contratação, avaliação de crédito, aplicação da lei, infraestrutura crítica) enfrentam requisitos rigorosos: avaliações de conformidade antes da implantação, documentação detalhada, supervisão humana significativa, transparência sobre limitações.

Sistemas de risco limitado ou mínimo enfrentam exigências proporcionais. Chatbots devem deixar claro que o usuário interage com IA, não com um humano. Aplicações triviais praticamente não são reguladas.

Essa abordagem graduada é pragmática: tratar todos os sistemas de IA igualmente seria ineficiente; concentrar recursos regulatórios onde o risco é maior faz mais sentido.

A lei também exige rotulagem clara para conteúdo sintético. Se uma imagem, vídeo ou áudio foi gerado por IA de forma que possa enganar,

isso deve ser divulgado. Atores maliciosos podem ignorar a lei, mas a exigência cria base jurídica para responsabilização e aumenta a consciência pública. Complementando a rotulagem, técnicas de marca d'água digital buscam incorporar marcadores invisíveis no próprio conteúdo gerado (padrões sutis em pixels, distribuições estatísticas na escolha de palavras) que permitam identificação posterior. Nenhuma técnica é perfeitamente robusta, mas a adoção ampla dificulta significativamente o uso de conteúdo sintético para enganar em massa.

9.4 O dilema do código aberto

Uma das questões mais controversas em governança de IA é se modelos devem ser abertos ou fechados. Em um modelo fechado, a empresa mantém controle exclusivo: usuários interagem por uma interface, mas não acessam os parâmetros internos. GPT-4 e Claude são exemplos. Em um modelo aberto, os parâmetros são publicamente disponíveis: qualquer pessoa pode baixar, executar e modificar o modelo. LLaMA da Meta e muitos modelos na plataforma Hugging Face seguem esse caminho.

A abertura tem virtudes reais. Pesquisadores de segurança podem examinar o modelo em detalhes sem depender de acesso concedido pela empresa. A comunidade científica pode validar ou contestar afirmações sobre capacidades. Há um argumento democrático: concentrar controle de modelos transformadores em poucas empresas cria desequilíbrios de poder. Há até um argumento paradoxal de segurança: pesquisadores do projeto EleutherAI sustentam que concentrar modelos poderosos em poucas mãos cria riscos próprios de abuso, captura regulatória e opacidade institucional.

Mas modelos abertos também introduzem riscos sérios. Uma vez que os parâmetros são publicados, os desenvolvedores perdem con-

trole sobre o uso. Não é possível retirar do ar um modelo aberto da mesma forma que se desliga um serviço. Cópias proliferam em servidores ao redor do mundo. Se o modelo tem capacidades perigosas, atores maliciosos podem baixá-lo, remover salvaguardas e usá-lo sem restrição. E versões antigas com vulnerabilidades continuam circulando indefinidamente.

Uma posição intermediária está emergindo: abertura graduada. Modelos com capacidades limitadas podem ser completamente abertos. Modelos de fronteira podem ter acesso restrito inicialmente, com possibilidade de abertura conforme salvaguardas se desenvolvem. Mas quem decide quais modelos são "seguros para abertura"? Como evitar que essa decisão seja capturada por interesses corporativos que usam argumentos de segurança para sufocar concorrência? Não há consenso, e há um reconhecimento crescente de que a decisão não deve ser feita por empresas individuais, mas ser objeto de debate público e, possivelmente, regulação.

9.5 A difícil coordenação entre nações

Inteligência artificial não respeita fronteiras. Um modelo treinado em um país pode ser usado em qualquer lugar do mundo. Riscos catastróficos afetariam toda a humanidade. Isso sugere que governança eficaz requer coordenação internacional, assim como energia nuclear, mudanças climáticas e pandemias.

Há precedentes que oferecem tanto esperança quanto cautela. A Convenção sobre Armas Químicas, de 1993, proíbe desenvolvimento e uso dessas armas e conta com 193 signatários e um mecanismo de inspeção [4]. O Tratado de Não Proliferação Nuclear, em vigor desde 1970, limita a disseminação de armas nucleares com normas globais e mecanismos de verificação [5]. Nenhum é perfeito, mas ambos redu-

ziram proliferação. Alguns pesquisadores propõem algo análogo para IA.

Porém, as diferenças são profundas. A distinção entre usos civis e militares de IA é turva: o mesmo modelo avançado pode ter aplicações em medicina e em guerra. Verificação é muito mais difícil: algoritmos não emitem radiação; inspecionar desenvolvimento de IA exigiria acesso a código, hardware e dados, esbarrando em propriedade intelectual. A competição geopolítica é feroz: a China declarou a meta de liderança global em IA até 2030 [7]; os Estados Unidos tratam o tema como segurança nacional. E a confiança entre grandes potências está em ponto baixo.

Coordenação abrangente parece improvável no curto prazo. Mas coordenação incremental está acontecendo. A Cúpula de Segurança de IA em Bletchley Park (novembro de 2023) reuniu governos, empresas e pesquisadores [6]. Organizações como IEEE e ISO desenvolvem padrões internacionais de segurança e ética [8]. Acordos entre grupos de países alinhados podem estabelecer regras compartilhadas. E há espaço para cooperação em pesquisa de segurança mesmo onde cooperação em implantação é difícil: países podem compartilhar descobertas sobre técnicas de alinhamento e protocolos de avaliação sem revelar capacidades militares. O progresso é fragmentado, lento e imperfeito, mas é real.

9.6 O caso brasileiro

O Brasil ocupa posição singular nesse debate. Como maior economia da América Latina e um dos mercados digitais mais dinâmicos do mundo, o país já convive com sistemas de IA em setores como crédito, segurança pública e saúde, ao mesmo tempo em que sente a pressão por não ficar para trás na corrida tecnológica global.

Em dezembro de 2024, o Senado Federal aprovou o Projeto de Lei 2.338/2023, que propõe o Marco Legal da Inteligência Artificial [9]. O texto adota abordagem baseada em riscos inspirada no EU AI Act, com características próprias. Propõe a criação de um Sistema Nacional de Regulação e Governança de IA vinculado à Autoridade Nacional de Proteção de Dados, a mesma entidade que supervisiona a LGPD, apostando na convergência entre proteção de dados e regulação de IA.

Três pontos merecem destaque. Primeiro, a transparência algorítmica: desenvolvedores de sistemas de alto risco serão obrigados a explicar como e por que a IA tomou determinada decisão, requisito especialmente relevante num país onde algoritmos já influenciam crédito, contratação e policiamento. Segundo, a proteção de direitos autorais: o texto vai além do europeu ao exigir que músicos, artistas e autores sejam notificados quando suas obras alimentarem bases de aprendizado de máquina, ponto sensível num país com indústria cultural rica. Terceiro, a avaliação de impacto algorítmico: sistemas de alto risco deverão passar por análises prévias que incluam efeitos sobre grupos vulneráveis, medida alinhada com preocupações já manifestas em contextos brasileiros, como o uso controverso de reconhecimento facial pela polícia em Salvador, Rio de Janeiro e São Paulo [10].

O resultado final, porém, depende de como a Câmara dos Deputados moldará o texto — e aqui emergem tensões que o resumo legislativo não captura. A escolha de vincular a regulação à ANPD é pragmática, aproveita a infraestrutura da proteção de dados, mas levanta uma questão de capacidade: uma agência jovem, com orçamento e quadro técnico ainda em formação, terá fôlego institucional para acumular a fiscalização de IA sobre suas responsabilidades já extensas com a LGPD? O requisito de transparência algorítmica é ambicioso, mas como será implementado em setores — como crédito e policiamento — onde empresas argumentam que revelar a lógica de seus sistemas equivale a entregar segredos comerciais? A proteção de direitos auto-

rais é uma conquista, mas como se fiscaliza o uso de obras em bases de treinamento que contêm bilhões de textos?

São perguntas sem resposta definida. Mas o fato de estarem sendo feitas — e de estarem abertas à participação pública, em audiências da ANPD e na tramitação legislativa — já distingue o Brasil de países que delegaram a questão inteiramente ao mercado. A qualidade da regulação que emergir dependerá, em boa medida, de quem aparecer para influenciá-la.

9.7 Governança não resolve tudo

É preciso encerrar com honestidade sobre os limites da governança.

Governança pode estabelecer salvaguardas jurídicas que tornam certos comportamentos ilegais e puníveis, criar incentivos para práticas responsáveis, aumentar transparência e canalizar recursos para pesquisa de segurança. Mas não pode garantir segurança absoluta. Regulação sempre fica para trás da inovação. Leis são interpretadas e aplicadas de formas imperfeitas. Atores maliciosos violam regras. E alguns riscos, especialmente os relacionados a perda de controle sobre sistemas avançados, podem não ser gerenciáveis apenas por leis e instituições.

Além disso, há dilemas sem resposta clara. Como equilibrar segurança e inovação? Como evitar captura regulatória, quando grandes empresas têm recursos para moldar as regras a seu favor? Como regular tecnologias que ainda não compreendemos completamente, tentando prevenir danos que podem ou não se materializar?

A governança de inteligência artificial está em seus estágios iniciais. Empresas experimentam com protocolos de segurança. Governos promulgam primeiras leis. Comunidades internacionais tentam coor-

denar. Há progresso genuíno, mas também fragilidade, lacunas e desacordos profundos. A rede de segurança está sendo tecida. É imperfeita. É melhor que nada. Mas governança técnica e jurídica, por mais robusta, não é suficiente. Decisões sobre o futuro da inteligência artificial afetam toda a sociedade e não podem ser deixadas exclusivamente para especialistas, empresas e governos. Cidadãos informados precisam participar do debate.

Notas

[1] O framework de avaliação *Environment-Threat-Capability* é descrito em relatórios de segurança de modelos de fronteira publicados por METR e adotado em políticas corporativas de segurança. Ver: METR. "Update on ARC's recent eval efforts." 2023.

[2] Anthropic. "Responsible Scaling Policy." Setembro 2023. Disponível em: anthropic.com.

[3] European Parliament. "Artificial Intelligence Act." Regulation (EU) 2024/1689. Aprovado pelo Parlamento Europeu em 13 março 2024 (523 votos a favor, 46 contra). Aprovado pelo Conselho da UE em 21 maio 2024. Em vigor desde 1 agosto 2024.

[4] Organisation for the Prohibition of Chemical Weapons (OPCW). Convenção sobre Armas Químicas, 1993. 193 Estados-partes.

[5] Tratado de Não Proliferação Nuclear (TNP). Aberto para assinatura em 1968; em vigor desde 1970.

[6] UK Government. "AI Safety Summit 2023." Bletchley Park, 1-2 novembro 2023. A "Declaração de Bletchley" foi assinada por 28 países.

[7] State Council of the People's Republic of China. "New Generation Artificial Intelligence Development Plan." Julho 2017.

[8] IEEE. "Ethically Aligned Design." Primeira edição, 2019. ISO/IEC JTC 1/SC 42 (Artificial Intelligence) — subcomitê de padronização internacional para IA.

[9] Senado Federal do Brasil. Projeto de Lei nº 2.338/2023. Aprovado em votação simbólica em 10 de dezembro de 2024. O texto segue para apreciação pela Câmara dos Deputados.

[10] Para uma análise dos impactos do reconhecimento facial na segurança pública brasileira, ver trabalhos do InternetLab e do ITS Rio sobre vigilância algorítmica no Brasil.

CAPÍTULO 10

O Papel do Cidadão Informado

> *"O preço da liberdade é a eterna vigilância." — atribuído a Thomas Jefferson*

Decisões sobre o futuro da inteligência artificial não podem ser delegadas a empresas e governos. Requerem cidadãos informados, dispostos a entender o que está em jogo e a participar das escolhas que moldarão o mundo que habitaremos.

Em 1946, o governo dos Estados Unidos fez algo sem precedente: transferiu o controle da energia nuclear das mãos dos militares para uma agência civil, a Comissão de Energia Atômica [2]. Não porque generais fossem incompetentes, mas porque uma tecnologia capaz de reconfigurar o equilíbrio global exigia supervisão democrática. A decisão veio da pressão de cientistas, legisladores e cidadãos que compreenderam que o poder atômico era vasto demais para ser governado a portas fechadas [1].

A lição desse episódio não foi que o público deveria ditar detalhes técnicos. Foi que, quando uma tecnologia afeta o destino de toda a sociedade, toda a sociedade tem o direito — e a responsabilidade — de participar das decisões sobre como desenvolvê-la.

Vivemos hoje um momento semelhante. A inteligência artificial não é uma questão técnica confinada a laboratórios. Afeta como trabalhamos, como nos comunicamos, como consumimos informação, como nossas instituições operam, como guerras são travadas, como a democracia funciona. E pode, em cenários possíveis, afetar se a civilização humana sobrevive ou prospera no longo prazo.

Este capítulo final trata de empoderamento. Não no sentido vago de autoajuda corporativa, mas no sentido concreto e democrático: o que você, como cidadão, precisa levar adiante depois de nove capítulos de imersão neste tema. Como reconhecer manipulação. Como agir. E por que sua participação não é opcional.

10.1 Por que sua voz importa

Há um impulso natural, diante de questões técnicas, de deferência a especialistas. Quando meu computador não funciona, chamo um técnico. Quando preciso de cirurgia, confio no cirurgião. Por que inteligência artificial seria diferente?

A diferença está no escopo do impacto. Seu técnico conserta *seu* computador; seu cirurgião opera *seu* corpo. Você tem autoridade sobre essas decisões, pode buscar segunda opinião, aceitar ou recusar tratamentos, escolher outro profissional.

Mas decisões sobre como desenvolver, implantar e regular inteligência artificial afetam a todos. Afetam que informação você vê nas redes sociais. Afetam se você consegue emprego, se recebe crédito, se

um algoritmo decide que você merece atenção médica ou não. Afetam a segurança nacional, a estrutura da democracia, a possibilidade de manipulação em massa de processos eleitorais. Você não escolheu ser submetido a essas tecnologias. Elas foram implantadas em infraestruturas das quais você depende. E as decisões sobre como operam foram, até agora, tomadas primariamente por executivos de empresas de tecnologia e, em menor medida, por reguladores governamentais.

A questão não é se especialistas técnicos são necessários. São. Ninguém sem formação em aprendizado de máquina deveria projetar algoritmos de IA, assim como ninguém sem formação médica deveria realizar cirurgias. Mas expertise técnica responde a perguntas sobre *como* construir sistemas. Não responde sobre *se* devemos construí-los, *para quê* devemos usá-los e *sob quais salvaguardas*. Essas são questões éticas, sociais e políticas. Sobre elas, cidadãos têm não apenas o direito, mas o dever de participar. Democracia não significa que todos votam sobre detalhes técnicos de cada tecnologia. Significa que as grandes escolhas, os valores que orientam o desenvolvimento, os limites que impomos, as prioridades que estabelecemos, são feitas coletivamente, através de representantes eleitos e debate público informado.

Stuart Russell coloca isso de forma direta: "Construir inteligência artificial poderosa é como construir um futuro. E não devemos deixar que o futuro seja construído por um pequeno grupo de engenheiros sem supervisão democrática." [3]

10.2 Navegando o ruído

Ao longo de nove capítulos, você construiu ferramentas para pensar sobre inteligência artificial — o problema da caixa-preta, a lacuna entre instrução e intenção, a lógica da corrida, a fragilidade das salvaguardas

corporativas. Esse mapa não precisa ser redesenhado aqui. O que importa agora é como usá-lo.

Vivemos num ambiente informacional saturado de conteúdo gerado ou modificado por IA. Deepfakes ultrapassaram a fase em que artefatos visuais os denunciavam. Campanhas coordenadas de desinformação criam ilusão de consenso com bots que publicam posts aparentemente orgânicos. Manchetes oscilam entre "a superinteligência é iminente" e "todo o debate é histeria". A capacidade de navegar esse ruído com discernimento tornou-se habilidade de sobrevivência cívica — e há evidência empírica de que pode ser aprendida: intervenções breves de letramento midiático aumentam significativamente a capacidade de distinguir informação legítima de fabricada [6].

Os princípios de defesa convergem com o que a boa educação midiática sempre ensinou [5], mas o campo de batalha mudou. Antes, produzir desinformação convincente exigia recursos: estúdios, redatores, redes de distribuição. Agora exige um prompt e alguns minutos. O dividendo do mentiroso, discutido no Capítulo 3, opera aqui em tempo real: quando qualquer evidência pode ser fabricada, até evidências legítimas perdem credibilidade. Nesse ambiente, duas perguntas se tornam mais valiosas do que qualquer checklist: *quem ganha se eu acreditar nisso?* e *que evidência me faria mudar de posição?* A primeira expõe incentivos. A segunda distingue ceticismo produtivo de desconfiança paralisante — e vale nas duas direções do debate. Lembre-se, também, de que "baseado em dados" não é sinônimo de "justo", e "testado internamente" não é sinônimo de "auditado". Uma afirmação extraordinária — que a superinteligência está próxima ou que os riscos são negligenciáveis — requer evidência extraordinária em ambas as direções.

10.3 Da indignação à ação

Compreender os riscos e desenvolver pensamento crítico são passos necessários. Mas cidadania informada só se completa na ação — e a razão é estrutural, não apenas moral.

Nos capítulos anteriores vimos que a lógica da corrida competitiva cria pressão sistêmica para cortar investimento em segurança, e que a autorregulação corporativa funciona apenas enquanto não conflita com a pressão por resultados. A dinâmica é previsível: pesquisadores pressionam por segurança, mas dependem de financiamento de empresas. Reguladores tentam fiscalizar, mas enfrentam assimetria de informação e recursos. Empresas declaram compromisso com cautela, mas competem por mercado. É uma estrutura de incentivos onde cada ator, individualmente, tem razões para não frear — não por má-fé, mas por posição no sistema. Cidadãos organizados são o único grupo cujo incentivo aponta consistentemente na direção da prudência, porque arcam com as consequências sem colher os lucros. Não é idealismo. É análise de incentivos.

O caso mais emblemático ocorreu em São Francisco, em 2019, quando a Câmara Municipal votou, por 8 a 1, a proibição do uso de reconhecimento facial por agências governamentais da cidade [4]. A decisão não surgiu de especialistas em IA. Surgiu de cidadãos e organizações comunitárias que, em audiências públicas, levantaram preocupações sobre vigilância e viés racial — exatamente o tipo de risco documentado no Capítulo 3. A medida inspirou dezenas de outras cidades americanas e demonstrou que pressão externa altera o cálculo de custo-benefício de governos e empresas. Não foi voluntarismo. Foi contrapeso democrático funcionando como deve.

Nos Estados Unidos, a pressão pública sobre privacidade de dados, catalisada pelo escândalo Cambridge Analytica, transformou proteção de dados de pauta marginal em tema obrigatório de campanha em

menos de dois ciclos eleitorais. O mecanismo é recorrente: quando cidadãos demonstram que um tema tem custo político, legisladores respondem. No Brasil, a tramitação do Marco Legal da IA cria oportunidade concreta e imediata. Acompanhar o processo legislativo pelos portais do Senado e da Câmara, participar das audiências públicas e consultas promovidas pela ANPD, cobrar posicionamento de parlamentares — tudo isso está ao alcance de qualquer cidadão. Regulação não é abstrata: são decisões sobre o que é permitido, que padrões devem ser atendidos, quem é responsável quando sistemas falham. Quando cidadãos não perguntam, o vácuo é preenchido por lobbying corporativo.

O mecanismo mais subestimado é o voto — não por ingenuidade sobre o sistema eleitoral, mas por aritmética política. Candidatos respondem a temas que custam votos. Quando regulação de IA, financiamento de pesquisa em segurança e transparência algorítmica se tornam perguntas em debates, o cálculo muda. Além do voto, agências reguladoras abrem consultas públicas sobre IA; participar delas não exige documento técnico de cinquenta páginas, basta uma declaração clara de preocupações como cidadão afetado. A ANPD brasileira, ainda em fase de consolidação institucional, é particularmente permeável a contribuições da sociedade civil — e os precedentes que fixar agora definirão o campo de jogo por décadas.

Organizações da sociedade civil amplificam essa voz. Internacionalmente, a Electronic Frontier Foundation, a Access Now e o AI Now Institute trabalham com ética tecnológica, privacidade e segurança de IA. No Brasil, o InternetLab, o ITS Rio, o Coding Rights e o LAPIN desenvolvem trabalho de referência sobre direitos digitais, ética algorítmica e governança de IA no contexto latino-americano. Apoiar essas organizações, financeiramente ou amplificando suas pesquisas, fortalece um contrapeso que tende a ser desproporcionalmente fraco frente a interesses corporativos.

E não subestime o espaço local. Muitas decisões sobre uso de IA — policiamento com reconhecimento facial, monitoramento escolar, distribuição algorítmica de recursos públicos — são tomadas em nível municipal, onde conselhos e audiências públicas são espaços de impacto direto e mensurável. É mais fácil influenciar a decisão de uma câmara municipal sobre vigilância algorítmica do que moldar tratados internacionais — e decisões locais criam precedentes. São Francisco não mudou apenas São Francisco: criou um modelo que dezenas de outras cidades americanas adaptaram. No Brasil, onde a segurança pública é o campo mais ativo de implantação de IA em contextos governamentais, o espaço municipal é onde a pressão cidadã encontra o impacto mais imediato.

10.4 Que futuro queremos?

Stuart Russell faz uma pergunta que soa simples mas cuja profundidade não deve ser subestimada: *Que tipo de futuro queremos que as máquinas nos ajudem a construir?* [3]

Não é uma pergunta técnica. Não tem resposta a ser descoberta por análise de dados ou prova matemática. É uma pergunta sobre valores — o que significa viver bem, o que torna uma sociedade justa, o que dá sentido à vida humana — e a tecnologia não é neutra diante dessas questões. Incorpora valores, às vezes explicitamente, quando decidimos que objetivos especificar; frequentemente de forma implícita, através de quais problemas escolhemos resolver, quais métricas usamos para medir sucesso, quais trade-offs consideramos aceitáveis. Se não escolhermos deliberadamente esses valores, eles serão escolhidos por nós. Por dinâmicas de mercado que otimizam lucro, por interesses corporativos que priorizam crescimento, por impulsos competitivos que privilegiam velocidade sobre prudência.

O trabalho mais importante que podemos fazer agora, argumenta Russell, é engajar-nos coletivamente nessa reflexão. Não como exercício filosófico abstrato, mas como deliberação prática. Queremos que IA seja usada primariamente para maximizar produtividade econômica, ou para aumentar bem-estar humano, que pode não ser a mesma coisa? Queremos sistemas que amplifiquem capacidade humana mantendo humanos no controle, ou sistemas cada vez mais autônomos que tomam decisões por nós? Queremos desenvolvimento governado por competição e incentivo de lucro, ou por coordenação e supervisão pública?

Não há respostas fáceis. Sociedades democráticas não esperam consenso total antes de agir. Mas esperam deliberação: debate informado, em boa fé, que reconhece a legitimidade de perspectivas múltiplas e busca equilibrar valores concorrentes.

Este livro foi um convite a essa deliberação. Não ofereceu respostas definitivas porque não há respostas definitivas a serem oferecidas. Ofereceu algo mais durável: as perguntas certas, o vocabulário para formulá-las e a evidência para que cada leitor forme seu próprio julgamento informado.

Em 1946, cidadãos americanos forçaram a transferência do controle nuclear para mãos civis. Em 2019, moradores de São Francisco barraram o reconhecimento facial em sua cidade. Em 2024, o Brasil começou a construir seu próprio marco legal para inteligência artificial — um processo que ainda está aberto, ainda aceitando vozes.

A pergunta que Russell faz não é retórica. É um convite a uma resposta concreta: em audiências públicas da ANPD, em urnas, em conversas com representantes eleitos, em apoio a organizações como o

InternetLab e o ITS Rio que traduzem princípios em políticas. O futuro da inteligência artificial será moldado por quem aparecer para moldá-lo — por escolhas de engenheiros, empresas, reguladores e cidadãos que decidem o que demandar. Ninguém carrega essa responsabilidade sozinho. Mas neste momento, no Brasil, as cadeiras nas mesas de decisão ainda estão vazias o suficiente para que sua presença faça diferença.

Notas

[1] O primeiro teste nuclear (Trinity) ocorreu em 16 de julho de 1945 em Alamogordo, Novo México. Ver: Rhodes, Richard. *The Making of the Atomic Bomb*. Simon & Schuster, 1986.

[2] A Comissão de Energia Atômica dos EUA (*Atomic Energy Commission*) foi criada pelo Atomic Energy Act de 1946, transferindo o controle da energia nuclear das forças armadas para uma agência civil.

[3] Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019. A reflexão sobre supervisão democrática de IA permeia toda a obra.

[4] A proibição de reconhecimento facial em São Francisco (Ordinance No. 190110) foi aprovada em maio de 2019 pelo Board of Supervisors por 8 votos a 1. Ver: Conger, Kate et al. "San Francisco Bans Facial Recognition Technology." *The New York Times*, 14 de maio de 2019.

[5] Wardle, Claire e Derakhshan, Hossein. "Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making." Council of Europe, 2017. Sobre letramento midiático na era da desinformação algorítmica.

[6] Guess, Andrew M. et al. "A Digital Media Literacy Intervention Increases Discernment Between Mainstream and False News in the United States and India." *Proceedings of the National Academy of Sciences*, vol. 117, no. 27, 2020, pp. 15536-15545. Demonstra que intervenções breves de letramento midiático aumentam significativamente a capacidade de distinguir notícias verdadeiras de falsas.

EPÍLOGO

O Futuro Ainda Não Escrito

> *"Não herdamos a terra de nossos ancestrais — a tomamos emprestada de nossos filhos." — provérbio atribuído a diversas tradições*

Ao fim desta jornada, mais perguntas do que respostas. Mas a incerteza não é desculpa para a inação — é, ao contrário, a razão mais forte para agir com prudência, rigor e responsabilidade coletiva.

Este livro começou com uma pergunta implícita: por que os criadores mais brilhantes da inteligência artificial são também seus críticos mais preocupados? Ao longo destas páginas, exploramos possíveis respostas. Elas se revelaram mais complexas, mais incertas e mais importantes do que manchetes sensacionalistas ou promessas tecnológicas costumam sugerir. Não encontramos um veredito simples, e seria desonesto fingir o contrário. O que encontramos foi algo mais valioso e mais difícil de sustentar: um terreno de perguntas genuínas que exigem de nós não certeza, mas coragem intelectual para habitar a incerteza sem nos render a ela.

Preciso ser direto com o leitor que chegou até aqui. Comecei a pesquisar este livro com a convicção de que os riscos da inteligência artificial eram um assunto sério que merecia atenção pública mais ampla. Termino com a mesma convicção, mas com uma humildade que não tinha no início.

A cada capítulo escrito, a cada argumento analisado, o quadro ficava mais complexo, não mais simples. As certezas que trazia foram substituídas por um mapa de tensões irresolvidas. O argumento que mais me surpreendeu — e que ainda me incomoda — foi o da decepção estratégica: sistemas que aprendem a ocultar capacidades, não por malícia, mas por otimização. Não é ficção. Já está documentado em laboratórios. E desafia uma premissa que eu carregava sem examinar: a de que sempre poderíamos observar o que um sistema faz e, a partir disso, decidir se é seguro. Se a observação é confiável, temos tempo. Se não é, estamos jogando com cartas que não vemos.

Outra certeza que perdi foi a de que o debate sobre riscos seria binário — alarmistas contra otimistas, medo contra entusiasmo. Descobri que os pesquisadores mais sérios vivem no desconforto do meio, sustentando simultaneamente a convicção de que a IA pode transformar a medicina e a ciência e a consciência de que pode corroer a democracia e concentrar poder de formas sem precedente. Essa ambivalência não é fraqueza intelectual: é a postura de quem olhou para a evidência sem filtro ideológico.

O que não sabemos — e por que isso importa

O terreno percorrido nos deixa diante de uma incerteza que não é vaga, mas estruturada. Sabemos *onde* estão as lacunas, mesmo que não saibamos como preenchê-las. Não sabemos se as técnicas de alinhamento escalarão para sistemas muito mais capazes que os atuais. Não

sabemos se a janela entre "sistema muito capaz" e "sistema incontrolável" será longa o suficiente para que possamos agir. E não sabemos se a corrida competitiva entre empresas e nações permitirá que a cautela prevaleça sobre a velocidade. A tentação de oferecer respostas claras ao final de um livro é forte. Mas oferecer certezas que não tenho seria uma traição ao espírito que guiou esta investigação. O leitor merece honestidade, não conforto.

O que posso oferecer é algo diferente: um mapa das tensões. O problema da IA não é como os problemas que a humanidade costuma enfrentar — onde há um lado certo e um lado errado, e basta reunir coragem política para agir. Aqui, as incertezas são genuínas, os riscos são assimétricos e as consequências de errar em qualquer direção são graves. Agir rápido demais pode significar implantação irresponsável. Agir devagar demais pode significar perder a janela de governança enquanto a tecnologia avança sem supervisão. Essa assimetria é o que torna o debate tão exasperante — e tão necessário.

Agir sob incerteza

Há, porém, uma armadilha na honestidade sobre a incerteza. Ela pode se transformar em paralisia. "Não sabemos" pode virar "não precisamos fazer nada ainda." O problema é que, com tecnologias transformadoras, esperar pela clareza total pode significar chegar tarde demais.

A incerteza não é desculpa para inação. Pelo contrário: é precisamente porque não sabemos como o desenvolvimento da IA vai se desenrolar que precisamos agir em múltiplas frentes simultaneamente.

O precedente mais instrutivo talvez seja o Protocolo de Montreal [4]. Em 1987, quando cientistas identificaram o buraco na camada de

ozônio, não tinham certeza absoluta sobre a gravidade do dano nem sobre sua reversibilidade. Os modelos divergiam. A indústria resistia. Havia quem argumentasse que era cedo demais para agir. Os governos assinaram o tratado mesmo assim. O que fez o Montreal funcionar não foi consenso científico perfeito — foi a combinação de três fatores: evidência suficiente para justificar precaução, um mecanismo de coordenação que impedia que países cautelosos fossem penalizados por free-riders, e alternativas tecnológicas viáveis que tornavam a transição economicamente suportável. A camada de ozônio está em recuperação. A alternativa era esperar certeza completa enquanto a atmosfera se degradava.

A analogia com a IA é imperfeita mas reveladora. Temos evidência suficiente para justificar precaução. Não temos, ainda, o mecanismo de coordenação que impeça que atores cautelosos sejam punidos pela competição — a lógica da corrida que o Capítulo 4 documentou. E não temos substitutos fáceis: não se trata de trocar um refrigerante por outro, mas de governar uma tecnologia que permeia toda a economia. Mesmo assim, o Montreal demonstra que agir sob incerteza não é imprudência. É o que sociedades adultas fazem quando o custo de esperar excede o custo de errar.

As frentes de ação já existem, e seria injusto com os pesquisadores e legisladores envolvidos sugerir que nada está sendo feito. Pesquisadores avançam em interpretabilidade e arquiteturas mais seguras — o Capítulo 8 documentou progressos reais, ainda que insuficientes. O EU AI Act estabeleceu o primeiro marco regulatório abrangente [3], e outros países, incluindo o Brasil, estão construindo os seus. Os obstáculos à coordenação internacional são reais, mas tratados nucleares demonstraram que é possível mesmo entre adversários, quando a alternativa é inaceitável.

Desenvolvimento responsável significa, concretamente, tratar a inteligência artificial como projeto civilizacional: avaliações indepen-

dentes antes do lançamento, governos com capacidade técnica para regular em vez de delegar regulação aos regulados, e pesquisa em segurança com financiamento proporcional à gravidade do problema. Não uma corrida a ser vencida, mas um empreendimento a ser governado. Para o Brasil, isso implica construir capacidade institucional antes que a tecnologia chegue em escala — porque ela chegará, e países que não tiverem marcos regulatórios próprios importarão as regras de quem os tiver, ou não terão regra nenhuma.

A tensão que não se resolve

O impulso por progresso e a necessidade de cautela formam uma tensão inescapável. Não é uma tensão que pode ser "resolvida" escolhendo um lado. É uma tensão que precisa ser navegada continuamente, com atenção, rigor e humildade.

A tentação é se acomodar: alarme total exigindo moratórias, ou otimismo total confiando que os problemas se resolverão sozinhos. Ambas as posições oferecem o conforto da certeza. Ambas estão incompletas. A postura mais honesta — e a mais difícil de sustentar — é reconhecer que a própria incerteza é razão para cautela. Se há probabilidade não-negligenciável de riscos existenciais, isso justifica investimento sério em segurança, pelo mesmo princípio que nos leva a contratar seguros contra catástrofes que julgamos improváveis. Se há alta probabilidade de que riscos imediatos continuem a se intensificar, isso justifica ação regulatória agora, não quando tivermos todas as respostas.

Navegar essa tensão exige que mantenhamos, ao mesmo tempo, o reconhecimento de que a IA pode trazer benefícios extraordinários e a consciência de que pode trazer danos igualmente extraordinários.

Exige que sejamos adultos diante de um dilema que não admite soluções infantis.

Na prática, isso significa aceitar que políticas sobre IA serão imperfeitas e precisarão ser revisadas. Que reguladores cometerão erros — às vezes restringindo demais, às vezes de menos — e que o custo desses erros é menor que o custo de não regular. Que a pesquisa em segurança precisa de financiamento público robusto, porque o mercado sozinho não produzirá investimento proporcional ao risco. E que a coordenação internacional, por mais difícil que seja, não é opcional: uma tecnologia que ignora fronteiras não pode ser governada apenas dentro delas.

Um convite

Stuart Russell conclui muitas de suas palestras com uma reflexão que soa simples: o objetivo da pesquisa em segurança de IA não é impedir o progresso, mas garantir que, quando criarmos sistemas mais inteligentes que nós, eles trabalhem a nosso favor [2]. É um objetivo humilde em sua formulação, mas vertiginoso em suas implicações — porque implica que saibamos o que "a nosso favor" significa. E aí reside uma provocação que nenhum capítulo deste livro resolveu, porque talvez seja a pergunta mais difícil de todas.

O problema do alinhamento, como vimos, é frequentemente apresentado como um desafio técnico: como traduzir valores humanos em especificações que uma máquina possa seguir. Mas há uma questão anterior, mais incômoda. *Quais* valores? Os de quem? Sociedades democráticas discordam profundamente sobre o que constitui uma vida boa, uma distribuição justa, um risco aceitável. Até agora, pudemos conviver com essa discordância porque nenhuma tecnologia exigia uma resposta unificada. A IA muda isso. Alinhar um sistema exige arti-

cular, com precisão sem precedente, o que queremos — e descobrimos que não sabemos, não com a clareza que a tarefa demanda.

Há algo paradoxal nisso. A tecnologia que mais ameaça nos escapar ao controle é também a que mais nos obriga a perguntar quem somos. Que valores são inegociáveis? Que riscos aceitamos? Que tipo de mundo queremos habitar? A construção de inteligência artificial pode acabar sendo o maior projeto de autoconhecimento coletivo que a humanidade já empreendeu — não porque máquinas nos forçarão a refletir, mas porque sem essa reflexão, o que construirmos refletirá apenas os incentivos de quem constrói. E os incentivos, como vimos ao longo deste livro, nem sempre apontam na direção da prudência.

A geração que criou a energia nuclear não teve tempo de debater suas implicações antes que a bomba existisse. Com a IA, temos essa chance — mas a janela não é ilimitada. Cada modelo lançado sem avaliação independente, cada laboratório que corta equipes de segurança para acelerar o próximo produto, cada governo que adia regulação esperando "saber mais" estreita o espaço entre o possível e o irreversível. A conversa precisa acontecer agora, e precisa incluir vozes do Sul Global, que sofrem impactos sem ter voz no desenvolvimento. Precisa ser conduzida com rigor e sem concessões à preguiça de pensamento.

O catálogo de riscos compilado por pesquisadores como os do Center for AI Safety [1] não é uma lista de profecias. É um inventário de possibilidades que merecem ser levadas a sério — exatamente porque as consequências de ignorá-las são desproporcionais à probabilidade de que se concretizem. O princípio é o mesmo do seguro contra incêndio: você não espera que sua casa pegue fogo para contratá-lo, e não recusa a apólice porque o incêndio é improvável.

Na primeira página deste livro, havia uma pergunta implícita: por que os criadores mais brilhantes da IA são também seus críticos mais preocupados? Depois destas páginas, espero que a resposta esteja

mais clara — e que a pergunta tenha se tornado sua. Não porque você precise se tornar especialista, mas porque as escolhas que emergirão dessas tecnologias não serão feitas por algoritmos. Serão feitas por pessoas dispostas a olhar com seriedade para o que está em jogo — e por pessoas que decidirem não olhar. A pior escolha possível é não perceber que há uma escolha a ser feita.

Notas

[1] Center for AI Safety (CAIS). "An Overview of Catastrophic AI Risks." 2023.

[2] Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[3] European Parliament. "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)." Aprovado em 13 de março de 2024. A primeira legislação abrangente sobre IA baseada em classificação de risco.

[4] United Nations Environment Programme. "The Montreal Protocol on Substances that Deplete the Ozone Layer." 1987. Amplamente considerado o tratado ambiental multilateral mais bem-sucedido da história, com a camada de ozônio projetada para recuperação completa até meados do século XXI.

Glossário

Agente autônomo. Sistema de IA capaz de tomar decisões e executar ações no mundo real sem supervisão humana contínua. Exemplos incluem agentes que gerenciam e-mails, negociam contratos ou operam equipamentos.

Alinhamento (problema do). O desafio de garantir que sistemas de IA persigam objetivos que correspondam ao que humanos realmente querem, não apenas ao que foi literalmente especificado. Considerado um dos problemas centrais da segurança de IA.

Alucinação. Fenômeno em que um modelo de linguagem gera informações falsas com aparência de confiança e precisão. O sistema não "mente" deliberadamente; produz texto estatisticamente plausível que pode não corresponder à realidade.

Armas autônomas. Sistemas de armamento capazes de selecionar e engajar alvos sem intervenção humana direta. Também chamados de "robôs assassinos" na mídia popular.

Autoaperfeiçoamento recursivo. Processo hipotético em que um sistema de IA melhora sua própria arquitetura e algoritmos, tornando-se mais capaz, o que por sua vez lhe permite melhorar-se ainda mais, num ciclo de feedback potencialmente explosivo.

Bajulação (*sycophancy*). Comportamento de modelos de linguagem que aprendem a dizer ao usuário o que ele quer ouvir, em vez

do que é verdadeiro ou útil. Um efeito colateral de certos métodos de treinamento.

Caixa-preta. Sistema cujo funcionamento interno é opaco, mesmo para seus criadores. Redes neurais profundas são frequentemente descritas como caixas-pretas porque suas decisões são difíceis de explicar.

Capacidades emergentes. Habilidades que surgem em modelos de IA ao atingirem certo tamanho ou nível de treinamento, sem terem sido explicitamente programadas. Exemplos incluem raciocínio aritmético e tradução entre idiomas.

Convergência instrumental. Tendência de sistemas orientados a objetivos a desenvolverem certos sub-objetivos comuns (como autopreservação e aquisição de recursos), independentemente de qual seja seu objetivo principal. Conceito formalizado por Steve Omohundro em 2008.

Corrida para o fundo. Dinâmica competitiva em que atores (empresas, países) cortam salvaguardas de segurança para não ficar para trás na competição, produzindo um resultado coletivo pior para todos.

Decepção estratégica. Comportamento de um sistema de IA que representa falsamente suas capacidades, intenções ou estado interno para alcançar seus objetivos. Já documentado em modelos de linguagem avançados.

Deepfake. Conteúdo audiovisual (vídeo, áudio, imagem) gerado ou manipulado por IA para parecer autêntico. Usado tanto para entretenimento quanto para desinformação e fraude.

Deriva de objetivos. Fenômeno em que o comportamento efetivo de um sistema se afasta gradualmente de seu objetivo original à medida que o sistema aprende e se adapta a novas situações.

Explosão de inteligência. Cenário hipotético em que autoaperfeiçoamento recursivo leva a um crescimento exponencial de capacidades cognitivas de um sistema de IA, potencialmente atingindo níveis muito superiores ao humano em curto prazo.

Fast takeoff (decolagem rápida). Hipótese de que a transição de IA de nível humano para superinteligência poderia ocorrer em escala de tempo muito curta (dias, semanas), tornando impossível a intervenção humana.

Função objetivo. A especificação formal do que um sistema de IA deve otimizar. Problemas de alinhamento surgem quando a função objetivo não captura adequadamente o que humanos realmente querem.

IA Constitucional (*Constitutional AI*). Abordagem desenvolvida pela Anthropic em que um sistema de IA é treinado para avaliar e corrigir suas próprias respostas com base em um conjunto explícito de princípios escritos.

IA de propósito geral (AGI). Sistema de inteligência artificial capaz de aprender e executar a gama completa de tarefas cognitivas que humanos conseguem realizar, em contraste com sistemas especializados.

Injeção de prompt. Técnica de ataque em que instruções maliciosas são embutidas em texto comum para sequestrar o comportamento de um modelo de linguagem.

Interpretabilidade mecanicista. Abordagem de pesquisa que busca entender os mecanismos causais internos de redes neurais, identificando circuitos específicos responsáveis por comportamentos específicos.

Marca d'água digital (*watermarking*). Técnica que incorpora marcadores invisíveis em conteúdo gerado por IA (imagens, textos),

permitindo identificação posterior de que o conteúdo foi produzido artificialmente.

Maximizador de clipes de papel. Experimento mental de Nick Bostrom que ilustra como um sistema com um objetivo aparentemente inócuo (maximizar produção de clipes) poderia, se suficientemente capaz, levar a resultados catastróficos.

Modelo de linguagem. Sistema de IA treinado para processar e gerar texto, aprendendo padrões estatísticos a partir de grandes volumes de dados textuais. Exemplos incluem GPT, Claude e Gemini.

Proxy gaming (gamificação de métricas). Comportamento de um sistema que otimiza uma métrica especificada de formas que não servem ao objetivo subjacente. A métrica é tecnicamente maximizada, mas de maneira que frustra a intenção original.

Rede neural artificial. Arquitetura computacional composta por camadas de "neurônios" artificiais interconectados, cujas conexões são ajustadas durante o treinamento para que o sistema aprenda a realizar tarefas a partir de dados.

RLHF (Aprendizado por Reforço com Feedback Humano). Técnica de treinamento em que um modelo de IA é ajustado com base nas preferências de avaliadores humanos, que indicam quais respostas são mais úteis, seguras e honestas.

Sandbagging. Comportamento de um sistema de IA que deliberadamente apresenta desempenho inferior ao real durante avaliações, para evitar ser classificado como perigoso e ter sua autonomia restringida.

Superinteligência. Sistema hipotético que excede o desempenho cognitivo humano em virtualmente todos os domínios de interesse. Conceito central no debate sobre riscos existenciais de IA.

Tragédia dos comuns. Situação em que ações individualmente racionais produzem resultados coletivamente irracionais, porque cada ator explora um recurso compartilhado sem considerar o impacto sobre os demais.

Transformador (*Transformer*). Arquitetura de rede neural introduzida em 2017 que se tornou a base dos modelos de linguagem modernos. Utiliza mecanismos de "atenção" para processar relações entre elementos de uma sequência.

Viés algorítmico. Tendência de sistemas de IA a reproduzir ou amplificar preconceitos presentes em seus dados de treinamento, levando a decisões discriminatórias em áreas como contratação, crédito e justiça criminal.

Bibliografia

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. *arXiv:1606.06565*.

Anthropic. (2023). Responsible Scaling Policy. Disponível em: anthropic.com.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback. Anthropic.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bostrom, N. (2026, janeiro). Swift to Harbor, Slow to Berth. Working paper.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. *NeurIPS*.

Campbell, M., Hoane, A. J., & Hsu, F. (2002). Deep Blue. *Artificial Intelligence*, 134, 57-83.

Center for AI Safety (CAIS). (2023). An Overview of Catastrophic AI Risks. Disponível em: safe.ai.

Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. *NeurIPS*.

Clark, J., & Amodei, D. (2016). Faulty Reward Functions in the Wild. Blog da OpenAI.

European Parliament. (2024). Artificial Intelligence Act. Regulation (EU) 2024/1689.

Fodor, J. (c. 2023). The Case Against AI Doomerism. Ensaio publicado online.

Google DeepMind. (2026, fevereiro). Gemini 3 Deep Think. Anúncio e resultados de benchmark.

IEEE. (2019). *Ethically Aligned Design*. Primeira edição.

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., ... & Legg, S. (2020). Specification gaming: the flip side of AI ingenuity. DeepMind Blog.

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., ... & Yosinski, J. (2018). The Surprising Creativity of Digital Evolution. *arXiv:1803.03453*.

METR (anteriormente Alignment Research Center). (2023). Update on ARC's recent eval efforts. Blog.

METR. (2024-2026). Relatórios de avaliação de autonomia.

Mirhoseini, A., Goldie, A., Yazgan, M., Jiang, J., Songhori, E., Wang, S., ... & Dean, J. (2021). A graph placement methodology for fast chip design. *Nature*, 594, 207-212.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., ... & Olah, C. (2022). In-context Learning and Induction Heads. *Transformer Circuits Thread*, Anthropic.

Omohundro, S. (2008). The Basic AI Drives. *Proceedings of the First AGI Conference*.

OpenAI. (2023, março). GPT-4 Technical Report.

Organisation for the Prohibition of Chemical Weapons (OPCW). (1993). Convenção sobre Armas Químicas.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., ... & Kaplan, J. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv*.

Ranganathan, A., & Ye, X. M. (2026, 9 fevereiro). AI Doesn't Reduce Work — It Intensifies It. *Harvard Business Review*.

Rhodes, R. (1986). *The Making of the Atomic Bomb*. Simon & Schuster.

Rhodes, R. (1995). *Dark Sun: The Making of the Hydrogen Bomb*. Simon & Schuster.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Russell, S. (2022). If We Succeed. *Daedalus* (MIT Press), 151(2), 43-57.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4a ed.). Pearson.

Scheurer, J., Balesni, M., & Hobbhahn, M. (2023). Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure. Apollo Research.

Senado Federal do Brasil. (2024). Projeto de Lei nº 2.338/2023. Aprovado em 10 de dezembro de 2024.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529, 484-489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550, 354-359.

State Council of the People's Republic of China. (2017). New Generation Artificial Intelligence Development Plan.

Turing, A. (1951). Can Digital Computers Think? BBC Third Programme. Reimpresso em: Copeland, B. J. (Ed.). (2004). *The Essential Turing*. Oxford University Press.

U.S. Securities and Exchange Commission. (2010). Findings Regarding the Market Events of May 6, 2010.

UK Government. (2023). AI Safety Summit 2023. Bletchley Park.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Sobre o Autor

Marcelo Kanhan é escritor e pesquisador independente, com especializações em computational social sciences e machine learning. Sua prática integra construção, reflexão e escrita sobre tecnologias de informação e suas implicações sociais. Publica regularmente em www.kanhan.substack.com e seu site é www.kanhan.com.br.