Max Tegmark

Vida 3.0

Os seres humanos na era da inteligência artificial

Tradução

Petê Rissatti

Benvirá

ISBN 978-65-5810-027-0

Dados Internacionais de Catalogação na Publicação (CIP)
Angélica Ilacqua CRB-8/7057

Tegmark, Max
Vida 3.0 : o ser humano na era da inteligência artificial / Max Tegmark ; tradução de Petê Rissatti. – São Paulo: Benvirá, 2020.
360 p.

Bibliografia
ISBN 978-65-5810-027-0
Título original: *Life 3.0 – Being Human in the Age of Artificial Intelligence*

1. Inteligência artificial. I. Título. II. Rissatti, Petê.

CDD 338.064
20-0425 CDU 330.341.1

Índices para catálogo sistemático:

1. Inteligência artificial.

Direção executiva *Flávia Alves Bravin*

Direção editorial *Renata Pascual Müller*

Gerência editorial *Rita de Cássia da Silva Puoço*

Edição *Tatiana Vieira Allegro*

Produção *Rosana Peroni Fazolari*

Preparação *Alyne Azuma e Paula Carvalho*

Consultoria técnica *Cássio Leandro Barbosa*

Revisão *Estela Janiski Zumbano*

Diagramação *Claudirene de Moura Santos*

Capa *Deborah Mattos*

Imagem de capa *iStock/GettyImagesPlus/Maksim Tkachenko*

Livro digital (E-pub)

Produção do e-pub *Fernando Ribeiro*

1ª edição, outubro de 2020

Av. Paulista, 901 – 3º andar
Bela Vista – São Paulo – SP – CEP: 01311-100

Dúvidas?

Acesse sac.sets@somoseducacao.com.br

CÓDIGO DA OBRA 703225 | CL 670941 | CAE 734280

À equipe do FLI, que tornou tudo possível.

Sumário

Você acha que a inteligência artificial sobre-humana pode ser criada neste século?

NÃO

SIM

Prólogo

A história da equipe Ômega

A equipe Ômega era a alma da empresa. Enquanto o restante da companhia ganhava dinheiro para manter as coisas funcionando por meio de vários aplicativos comerciais de IA limitada, a equipe Ômega avançava na busca pelo que sempre tinha sido o sonho do CEO: construir uma inteligência artificial geral. A maioria dos outros funcionários via "os Ômegas", como eram carinhosamente chamados, como um bando de sonhadores, eternamente a décadas de distância de seus objetivos. Mas eles os mimavam de bom grado, pois gostavam tanto do prestígio que o trabalho de ponta dos Ômega conferia à empresa, quanto dos algoritmos aprimorados que a equipe ocasionalmente lhes fornecia.

O que não perceberam foi que os Ômegas haviam moldado com cuidado sua imagem para esconder um segredo: estavam extremamente perto de realizar o plano mais audacioso da história humana. O carismático CEO os escolhera não apenas por serem pesquisadores brilhantes, mas também pela ambição, pelo idealismo e pelo forte compromisso de ajudar a humanidade. Ele os lembrava que o plano da equipe era muito perigoso e que, se governos poderosos o descobrissem, fariam praticamente qualquer coisa – inclusive sequestro – para impedi-los ou, de preferência, roubar seu código. Mas todos estavam 100% envolvidos pelo mesmo motivo por que muitos dos principais físicos do mundo se juntaram ao Projeto Manhattan para desenvolver armas nucleares: estavam convencidos de que, se não o fizessem primeiro, alguém menos idealista o faria.

A IA que haviam construído, apelidada de Prometheus, estava se tornando cada vez mais competente. Embora suas habilidades cognitivas ainda estivessem muito atrás das dos seres humanos em muitas áreas, por exemplo, em habilidades sociais, os Ômegas haviam se esforçado bastante para torná-la extraordinária em uma tarefa específica:

programar sistemas de IA. Eles tinham escolhido deliberadamente essa estratégia porque haviam comprado o argumento da explosão da inteligência feito pelo matemático britânico Irving Good, em 1965:

> Que uma máquina ultrainteligente seja definida como uma máquina que possa superar em muito todas as atividades intelectuais de qualquer homem, por mais inteligente que seja. Como o design de máquinas é uma dessas atividades intelectuais, uma máquina ultrainteligente pode projetar máquinas ainda melhores; indiscutivelmente haveria uma "explosão de inteligência", e a inteligência do homem ficaria para trás. Assim, a primeira máquina ultrainteligente é a última invenção que o homem precisa fazer, contanto que a máquina seja dócil o suficiente para nos dizer como mantê-la sob controle.

Eles imaginaram que, se conseguissem manter esse autoaperfeiçoamento se repetindo, a máquina logo ficaria inteligente o suficiente para poder ensinar a si mesma todas as outras habilidades humanas úteis.

Os primeiros milhões

Eram nove da manhã de uma sexta-feira quando decidiram fazer o lançamento. Prometheus zumbia em seu personalizado cluster de computadores, localizado em longas filas de baias em uma ampla sala com ar-condicionado e controle de acesso. Por motivos de segurança, estava completamente desconectado da internet, mas continha uma cópia local de grande parte da web (Wikipédia, Biblioteca do Congresso, Twitter, uma seleção de vídeos do YouTube, grande parte do Facebook etc.) para ser usada como dados de treinamento a partir do qual aprender.[1] A equipe escolheu esse horário de início para trabalhar sem distrações: famílias e amigos acreditavam que eles estavam em um retiro corporativo de fim de semana. A copa estava abastecida com alimentos para preparo no micro-ondas e bebidas energéticas, e eles estavam prontos para começar.

Quando foi lançado, o Prometheus era um pouco pior do que eles na programação de sistemas de IA, mas compensava esse fato sendo muito mais rápido, gastando o equivalente a milhares de anos-pessoa trabalhando no problema enquanto eles tomavam Red Bull. Às dez, ele havia concluído o primeiro redesign de si mesmo, versão 2.0, que era um pouco melhor, mas ainda sub-humana. Mas quando o Prometheus 5.0 foi lançado, às duas da tarde, os Ômegas ficaram impressionados: havia ultrapassado seus parâmetros

de desempenho, e o índice de progresso parecia estar se acelerando. Ao anoitecer, decidiram implantar o Prometheus 10.0 para iniciar a fase 2 de seu plano: ganhar dinheiro.

O primeiro alvo deles foi o MTurk, Amazon Mechanical Turk. Após seu lançamento, em 2005, como um serviço de crowdsourcing na internet, ele cresceu rápido, com dezenas de milhares de pessoas em todo o mundo competindo anonimamente, 24 horas por dia, para executar tarefas altamente estruturadas chamadas HITs, "Human Intelligence Tasks" (Tarefas de Inteligência Humana). Essas tarefas variavam de transcrever gravações de áudio a classificar imagens e escrever descrições de páginas da web, e todas tinham uma coisa em comum: se as fizesse bem, ninguém saberia que você era uma IA. O Prometheus 10.0 conseguiu executar cerca de metade das categorias de tarefas de maneira aceitável. Para cada uma, os Ômegas faziam o Prometheus projetar um módulo de software customizado de inteligência artificial limitada, capaz de executar exatamente essas tarefas e nada mais. Em seguida, eles faziam o upload desse módulo no Amazon Web Services, uma plataforma de computação em nuvem que podia rodar em todas as máquinas virtuais que eles alugassem. A cada dólar pago à divisão de computação em nuvem da Amazon, eles ganhavam mais de dois dólares da divisão MTurk da Amazon. Mal suspeitava a Amazon que existia uma oportunidade tão incrível de arbitragem dentro da própria empresa!

Para não deixar pistas, eles haviam criado discretamente milhares de contas MTurk nos meses anteriores com nomes de pessoas fictícias, e os módulos construídos pelo Prometheus agora assumiam suas identidades. Os clientes do MTurk costumavam pagar após cerca de oito horas, momento em que os Ômegas reinvestiam o dinheiro em mais tempo de computação em nuvem, usando módulos de tarefas ainda melhores, criados pela versão mais recente do Prometheus, cada vez mais aprimorado. Como conseguiam duplicar seu dinheiro a cada oito horas, logo começaram a saturar o suprimento de tarefas do MTurk e descobriram que não podiam ganhar mais do que 1 milhão de dólares por dia sem chamar atenção, algo que não queriam fazer. Mas era suficiente para poderem dar o próximo passo, eliminando qualquer necessidade de pedir dinheiro ao diretor financeiro.

Jogos perigosos

Além das inovações em IA, um dos projetos recentes com os quais os Ômegas mais se divertiram foi o planejamento de como ganhar dinheiro o mais rápido possível depois do lançamento do Prometheus. Essencialmente, toda a economia digital estava disponível, mas será que era melhor começar criando jogos de computador, música, filmes ou software, escrever livros ou artigos, negociar na bolsa de valores ou fazer invenções e vendê-las? O resumo era simples: maximizar a taxa de retorno do investimento, mas as estratégias normais de investimento eram uma paródia em câmera lenta do que podiam fazer – enquanto um investidor normal talvez ficasse satisfeito com um retorno de 9% por ano, seus investimentos em MTurk rendiam 9% por *hora*, gerando oito vezes mais dinheiro por dia. Agora que eles haviam saturado o MTurk, qual seria o próximo passo?

A primeira ideia foi ganhar muito dinheiro no mercado de ações – afinal, praticamente todos eles em algum momento recusaram uma oferta de emprego lucrativa para desenvolver IA para fundos de cobertura (*hedge funds*) – que estavam investindo pesado exatamente nessa ideia. Alguns lembraram que foi assim que a IA no filme *Transcendence: A revolução* conseguiu seus primeiros milhões. Mas as novas regulamentações sobre derivativos após o colapso do ano anterior tinham limitado suas opções. Logo perceberam que, embora pudessem obter retornos muito melhores do que outros investidores, seria improvável obterem retornos próximos ao que conseguiriam com a venda de seus produtos. Quando se tem a primeira IA superinteligente do mundo trabalhando para você, é melhor investir nas suas próprias empresas, não nas de outros! Embora pudesse haver exceções ocasionais (como usar as habilidades sobre-humanas de hacker de Prometheus para obter informações privilegiadas e adquirir opções de compra de ações prestes a subir), os Ômegas acharam que isso não valia a atenção indesejada que poderiam atrair.

Quando mudaram o foco para produtos que pudessem desenvolver e vender, os jogos de computador pareceram a primeira opção óbvia. O Prometheus logo conseguiu se tornar extremamente hábil no design de jogos atraentes, manipulando com facilidade a codificação, o design gráfico, o *ray tracing* (uma espécie de renderização) das imagens e todas as outras tarefas necessárias para criar um produto final pronto para o envio. Além disso, depois de digerir todos os dados da web sobre as preferências das pessoas, a IA saberia exatamente do que cada categoria de jogador gostava e poderia desenvolver uma capacidade sobre-humana de otimizar um jogo para obter receita de vendas. The Elder Scrolls V: Skyrim, um jogo no qual muitos Ômegas haviam desperdiçado mais horas do que gostariam de admitir, tinha arrecadado mais de 400 milhões dólares durante sua

primeira semana, lá em 2011, e eles estavam confiantes de que o Prometheus poderia criar em 24 horas algo ao menos tão viciante usando 1 milhão de dólares em recursos de computação em nuvem. Poderiam vendê-lo on-line e usar o Prometheus para se passar por humanos e falar sobre o jogo na blogosfera. Se isso gerasse 250 milhões de dólares em uma semana, teriam dobrado seu investimento oito vezes em oito dias, dando um retorno de 3% por hora – um pouco pior do que o início do MTurk, mas muito mais sustentável. Ao desenvolver um conjunto de outros jogos por dia, eles imaginaram que seriam capazes de ganhar 10 bilhões de dólares em pouco tempo, sem chegar perto de saturar o mercado de jogos.

Mas uma especialista em segurança cibernética da equipe os dissuadiu desse plano. Ela destacou que isso representaria um risco inaceitável de o Prometheus *se rebelar* e assumir o controle do próprio destino. Como não tinham certeza de como os objetivos do Prometheus evoluiriam durante seu autoaperfeiçoamento recursivo, decidiram se esforçar para a mantê-lo confinado ("encaixotado"), de forma que não pudesse escapar para a internet. Para o mecanismo principal do Prometheus em execução na sala de servidores, foi usado um confinamento físico: simplesmente não havia conexão com a internet, e a única saída do Prometheus era na forma de mensagens e documentos enviados a um computador controlado pelos Ômegas.

Em um computador conectado à internet, por outro lado, executar qualquer programa complicado criado pelo Prometheus era uma proposta arriscada: como os Ômegas não tinham como entender completamente o que ele faria, não tinham como saber se ele, digamos, não começaria a se espalhar de modo viral on-line. Ao testar o software que Prometheus havia escrito para as tarefas do MTurk, os Ômegas se protegeram disso, executando-o apenas dentro de uma máquina virtual, que é um programa que simula um computador: por exemplo, muitos usuários de Mac compram softwares de máquina virtual que permitem executar programas do Windows, fazendo-os "achar" que de fato estão em uma máquina com esse sistema. Os Ômegas criaram sua própria máquina virtual, que apelidaram de Caixa de Pandora e que simulava uma máquina ultrassimplificada, sem todos os detalhes que costumamos associar aos computadores: teclado, monitor, alto-falante, conexão à internet, nada. Para as transcrições de áudio do MTurk, os Ômegas configuravam o sistema de modo que tudo o que pudesse entrar na Caixa de Pandora fosse um único arquivo de áudio e tudo o que pudesse sair fosse um único documento de texto – a transcrição. Essas leis da caixa eram para o software interno como as leis da física são para nós, dentro de nosso Universo: o

software não podia sair da caixa, assim como nós não conseguimos viajar mais rápido que a velocidade da luz, por mais inteligentes que sejamos. Exceto por essa única entrada e saída, o software dentro da Caixa de Pandora de fato ficava preso em um universo paralelo com suas regras computacionais. Os Ômegas eram tão paranoicos em relação à fuga que a "encaixotaram", limitando a vida útil do código não confiável. Por exemplo, sempre que o software de transcrição encaixotado terminava de transcrever um arquivo de áudio, todo o conteúdo da memória da Caixa de Pandora era automaticamente apagado e o programa era reinstalado do zero. Dessa forma, quando iniciava a próxima tarefa de transcrição, não tinha conhecimento do que havia acontecido anteriormente e, portanto, não era capaz de aprender com o passar do tempo.

Quando usaram a nuvem da Amazon em seu projeto MTurk, os Ômegas puderam colocar todos os módulos de tarefas criados pelo Prometheus nas tais caixas virtuais na nuvem, porque a entrada e a saída do MTurk eram muito simples. Mas isso não funcionaria para jogos de computador com gráficos pesados, que não poderiam ser encaixotados porque precisariam de acesso total a todo o hardware do computador do jogador. Além disso, eles não queriam arriscar que algum usuário com experiência em computador analisasse o código do jogo, descobrisse a Caixa de Pandora e decidisse investigar o que havia lá dentro. O risco de fuga colocou não apenas o mercado de jogos fora do alcance por ora, mas também o mercado extremamente lucrativo de outros softwares, com centenas de bilhões de dólares à disposição.

Os primeiros bilhões

Os Ômegas restringiram sua pesquisa a produtos de alto valor, puramente digitais (evitando a fabricação lenta) e de fácil compreensão (como textos ou filmes que eles sabiam não representar um risco de fuga). No final, decidiram lançar uma empresa de mídia, começando com entretenimento animado. O site, o plano de marketing e os comunicados à imprensa já estavam prontos para ser lançados antes mesmo de o Prometheus se tornar superinteligente – só faltava o conteúdo.

Embora o Prometheus tenha se tornado supreendentemente capaz na manhã de domingo, constantemente arrecadando dinheiro no MTurk, suas habilidades intelectuais ainda eram bastante limitadas: ele havia sido deliberadamente otimizado para projetar sistemas de IA e escrever softwares que executassem tarefas MTurk bem entediantes. Por

exemplo, ele era ruim em fazer filmes. E não era ruim por um motivo grave, mas pela mesma razão pela qual James Cameron era ruim em fazer filmes quando nasceu: é uma habilidade que leva tempo para ser aprendida. Como uma criança humana, o Prometheus podia aprender o que quisesse com os dados a que tinha acesso. Enquanto James Cameron levou anos para aprender a ler e escrever, o Prometheus já conseguia fazer isso na sexta-feira, quando também encontrou tempo para ler toda a Wikipédia e alguns milhões de livros. Fazer filmes era mais difícil. Escrever um roteiro que humanos considerassem interessante era tão difícil quanto escrever um livro, exigindo uma compreensão detalhada da sociedade humana e do que os seres humanos achavam divertido. Transformar o roteiro em um arquivo de vídeo final exigia grandes quantidades de *ray tracing* de atores simulados e as cenas complexas que interpretavam, vozes simuladas, produção de trilhas sonoras musicais atraentes etc. A partir da manhã de domingo, o Prometheus podia assistir a um filme de duas horas em cerca de um minuto, o que incluía a leitura de qualquer livro no qual o filme se baseasse e todas as críticas e classificações on-line. Os Ômegas notaram que, depois de assistir a algumas centenas de filmes, o Prometheus tornou-se muito bom em prever que tipo de crítica um filme receberia e como atrairia diferentes públicos. De fato, ele aprendeu a escrever as próprias resenhas de filmes de maneira que parecia demonstrar uma percepção real, comentando tudo, desde as tramas e a atuação até os detalhes técnicos, como iluminação e ângulos da câmera. Com isso, a equipe entendeu que, quando Prometheus fizesse os próprios filmes, saberia o que significava sucesso.

A princípio, os Ômegas instruíram Prometheus a se concentrar em fazer animação, para evitar perguntas embaraçosas sobre quem eram os atores simulados. No domingo à noite, o fim de semana terminou animado: eles providenciaram cerveja e pipoca de micro-ondas, diminuíram as luzes e assistiram ao filme de estreia de Prometheus. Era uma comédia de fantasia animada, no estilo de *Frozen*, da Disney, e o *ray tracing* havia sido realizado pelo código de Prometheus, na nuvem da Amazon, utilizando a maior parte do 1 milhão de dólares que haviam lucrado no MTurk no dia. Quando começou, acharam fascinante e assustador o filme ter sido criado por uma máquina sem orientação humana. No entanto, em pouco tempo estavam rindo das piadas e prendendo a respiração durante os momentos dramáticos. Alguns deles até choraram um pouco com o final emocionante, totalmente absortos naquela realidade fictícia, a ponto de se esquecerem totalmente quem a havia criado.

Os Ômegas agendaram o lançamento do site para a sexta-feira seguinte, dando a Prometheus tempo para produzir mais conteúdo e tempo para eles fazerem as coisas que não confiavam a Prometheus: comprar anúncios e começar a recrutar funcionários para as empresas de fachada criadas nos meses anteriores. Para disfarçar, o discurso oficial seria que a empresa de mídia (que não tinha associação pública com os Ômegas) havia comprado a maior parte de seu conteúdo de produtores de filmes independentes, tipicamente startups de alta tecnologia em regiões de baixa renda. Esses fornecedores falsos estavam convenientemente localizados em lugares remotos, como Tiruchirappalli e Yakutsk, que nem os jornalistas mais curiosos se incomodariam em visitar. Os únicos funcionários de fato contratados trabalhavam em marketing e administração e diriam a todos que sua equipe de produção ficava em um local diferente e não daria entrevistas no momento. Para combinar com o discurso oficial, escolheram o slogan corporativo "Canalizando o talento criativo do mundo" e classificaram sua empresa como sendo diferente, usando tecnologia de ponta para capacitar pessoas criativas, especialmente em países em desenvolvimento.

Quando chegou a sexta-feira e os visitantes curiosos começaram a entrar no site, encontraram algo que lembrava os serviços de entretenimento on-line da Netflix e do Hulu, mas com diferenças interessantes. Todas as séries animadas eram novas, e ninguém nunca tinha ouvido falar nelas. Eram bastante cativantes: a maioria das séries consistia em episódios de 45 minutos com uma trama forte, cada um terminando de modo a nos deixar ansiosos para descobrir o que aconteceria no episódio seguinte. E eram mais baratos que a concorrência. O primeiro episódio de cada série era gratuito, e era possível assistir aos outros por 49 centavos de dólar cada ou com descontos para toda a série. Inicialmente, havia apenas três séries com três episódios cada, mas novos episódios eram adicionados todo dia, além de novas séries, atendendo a diferentes públicos-alvo. Durante as primeiras duas semanas do Prometheus, suas habilidades de produção de filmes melhoraram rapidamente, em termos não apenas de qualidade, mas também de algoritmos melhores para simulação de personagens e *ray tracing*, o que reduziu bastante o custo da computação em nuvem para criar cada episódio. Como resultado, os Ômegas conseguiram lançar dezenas de novas séries durante o primeiro mês, voltadas para públicos que iam de crianças pequenas a adultos, além de expandir-se para todos os principais mercados de idiomas do mundo, tornando seu site visivelmente internacional em comparação com todos os concorrentes. Alguns comentaristas ficaram impressionados com o fato de não apenas os diálogos serem multilíngues, mas também

os próprios vídeos: por exemplo, quando um personagem falava italiano, os movimentos da boca correspondiam às palavras em italiano, bem como os gestos, que eram característicos da cultura do país. Embora o Prometheus fosse agora perfeitamente capaz de fazer filmes com atores simulados indistinguíveis dos seres humanos, os Ômegas evitavam isso para não revelar suas intenções. No entanto, lançaram muitas séries com personagens humanos animados semirrealistas, em gêneros que competiam com programas de TV e filmes tradicionais com atores reais.

Sua rede acabou se mostrando bastante viciante e teve um crescimento impressionante de espectadores. Muitos fãs acharam os personagens e os enredos mais inteligentes e interessantes do que as produções mais caras das telonas de Hollywood e ficaram encantados por poderem assisti-los de maneira muito mais acessível. Impulsionada pela publicidade agressiva (que os Ômegas conseguiam pagar por causa dos custos de produção quase nulos), pela excelente cobertura da mídia e pelas ótimas críticas boca a boca, sua receita global subiu para 10 milhões de dólares por dia um mês após o lançamento. Depois de dois meses, ultrapassaram a Netflix e, após três, estavam arrecadando mais de 100 milhões de dólares por dia, começando a rivalizar com Time Warner, Disney, Comcast e Fox como um dos maiores impérios da mídia no mundo.

Esse incrível sucesso atraiu muita atenção indesejada, incluindo especulações de que adotavam uma IA forte, mas usando apenas uma pequena fração de sua receita os Ômegas implementaram uma campanha de desinformação muito bem-sucedida. Em um escritório novo e chamativo em Manhattan, seus porta-vozes recém-contratados elaborariam suas histórias. Muitos humanos foram contratados como laranjas, incluindo roteiristas do mundo todo, com o objetivo de começar a desenvolver novas séries, e nenhum deles sabia sobre o Prometheus. A confusa rede internacional de subcontratados facilitou para que a maioria de seus funcionários imaginasse que outras pessoas em algum outro lugar estavam fazendo a maior parte do trabalho.

Para ficarem menos vulneráveis e evitar críticas sobre a computação em nuvem excessiva, também contrataram engenheiros para começar a construir uma série de enormes instalações de computadores em todo o mundo, que pertenceriam a empresas de fachada aparentemente sem vínculos. Embora tivessem sido anunciados aos locais como "data centers ecológicos", por serem, em grande parte, movidos a energia solar, na verdade estavam focados principalmente na computação, e não no armazenamento. Prometheus havia projetado seus planos nos mínimos detalhes, usando apenas hardware pronto e otimizando-o para minimizar o tempo de construção. As pessoas que

construíam e administravam esses centros não tinham ideia do que era feito lá: pensavam estar supervisionando instalações comerciais de computação em nuvem semelhantes às administradas pela Amazon, pelo Google e pela Microsoft, e sabiam apenas que todas as vendas eram gerenciadas remotamente.

Novas tecnologias

Durante meses, o império empresarial controlado pelos Ômegas começou a ganhar posição em áreas cada vez maiores da economia mundial, graças ao planejamento sobre-humano do Prometheus. Ao analisar cuidadosamente os dados do mundo, ele já havia apresentado aos Ômegas, durante a primeira semana, um plano detalhado de crescimento passo a passo, e continuou aprimorando e refinando esse plano à medida que seus recursos de dados e computadores cresciam. Embora Prometheus estivesse longe de ser onisciente, suas capacidades eram agora muito superiores às dos humanos, a ponto de os Ômegas o verem como o oráculo perfeito, obedientemente oferecendo respostas e conselhos brilhantes para todas as perguntas.

O software do Prometheus tinha se tornado altamente otimizado para aproveitar ao máximo o hardware medíocre inventado por humanos em que era executado, e, como os Ômegas haviam previsto, Prometheus identificou maneiras de melhorar drasticamente esse hardware. Temendo uma fuga, eles se recusaram a construir instalações de construção robótica que o Prometheus pudesse controlar de maneira direta. Em vez disso, contrataram um grande número de cientistas e engenheiros do mundo todo, em vários locais, e lhes deram relatórios internos de pesquisa escritos por Prometheus, fingindo que eram de pesquisadores de outros locais. Esses relatórios detalhavam novos efeitos físicos e técnicas de produção que seus engenheiros logo testaram, entenderam e dominaram. É claro que os ciclos normais de pesquisa e desenvolvimento realizados por humanos levam anos, em grande parte porque envolvem muitos ciclos lentos de tentativa e erro. A situação atual era muito diferente: o Prometheus já tinha os próximos passos planejados; portanto, o fator limitante era simplesmente a rapidez com que as pessoas poderiam ser orientadas a entender e construir as coisas certas. Um bom professor pode ajudar os alunos a aprender ciências muito mais rápido do que eles poderiam aprender se descobrissem tudo por conta própria, e o Prometheus, disfarçadamente, fez o mesmo com esses pesquisadores. Como podia prever com precisão quanto tempo levaria para os

humanos entenderem e construírem as coisas, com várias ferramentas, Prometheus desenvolveu o caminho mais rápido possível, priorizando novas ferramentas que pudessem ser entendidas e construídas com rapidez e que fossem úteis para o desenvolvimento de ferramentas mais avançadas.

No espírito da cultura *maker*, as equipes de engenharia foram incentivadas a usar suas máquinas para construir máquinas melhores. Essa autossuficiência não apenas economizou dinheiro, mas também os tornou menos vulneráveis a ameaças futuras do mundo exterior. Em dois anos, estavam produzindo o melhor hardware de computador que o mundo já havia conhecido. Para evitar ajudar a concorrência, essa tecnologia foi mantida em sigilo e usada apenas para atualizar o Prometheus.

O que o mundo notou, no entanto, foi um espantoso *boom* tecnológico. Empresas iniciantes em todo o mundo estavam lançando produtos revolucionários em quase todas as áreas. Uma startup sul-coreana lançou uma nova bateria que armazenava o dobro da energia da bateria do laptop com a metade da massa e podia ser carregada em menos de um minuto. Uma empresa finlandesa lançou um painel solar barato com o dobro da eficiência dos melhores concorrentes. Uma empresa alemã anunciou um novo tipo de fio produzido em massa que era supercondutor à temperatura ambiente, revolucionando o setor de energia. Um grupo de biotecnologia de Boston, nos Estados Unidos, anunciou um ensaio clínico de Fase II do que eles alegaram ser o primeiro medicamento eficaz para perda de peso sem efeitos colaterais, enquanto rumores sugeriam que uma empresa indiana já estava vendendo algo semelhante de forma clandestina. Uma empresa californiana reagiu com um estudo de Fase II de um medicamento para câncer de grande sucesso, que fazia o sistema imunológico do corpo identificar e atacar as células com qualquer uma das mutações cancerígenas mais comuns. Continuavam a surgir exemplos, provocando discussões sobre uma nova era de ouro para a ciência. Por último, mas não menos importante, as empresas de robótica estavam brotando como cogumelos em todo o mundo. Nenhum dos robôs chegou perto de se igualar à inteligência humana, e a maioria deles não se parecia em nada com humanos. Mas perturbaram de maneira drástica a economia e, ao longo dos anos seguintes, gradualmente substituíram a maioria dos trabalhadores em manufatura, transporte, armazenamento, varejo, construção, mineração, agricultura, silvicultura e pesca.

O que o mundo não notou, graças ao trabalho árduo de uma excelente equipe de advogados, foi que todas essas empresas eram controladas, por meio de uma série de intermediários, pelos Ômegas. Prometheus estava inundando os escritórios de patentes

do mundo com invenções sensacionais por meio de vários proxies, e essas invenções gradualmente levaram ao domínio em todas as áreas da tecnologia.

Embora essas novas empresas disruptivas tivessem conquistado grandes inimigos entre seus concorrentes, fizeram amigos ainda mais poderosos. Eram excepcionalmente lucrativas e, com slogans como "Investindo em nossa comunidade", gastaram uma fração significativa desses lucros contratando pessoas para projetos comunitários – em geral as mesmas pessoas que tinham sido demitidas das empresas desfeitas. Usaram análises detalhadas produzidas pelo Prometheus para identificar trabalhos que seriam extremamente gratificantes para os funcionários e para a comunidade pelo menor custo, adaptados às circunstâncias locais. Em regiões com altos níveis de serviços governamentais, isso costumava se concentrar na construção, cultura e assistência comunitária, enquanto nas regiões mais pobres também incluía lançamento e manutenção de escolas, serviços de saúde, creches, asilos, moradias a preços acessíveis, parques e infraestrutura básica. Em praticamente todos os lugares, os moradores concordaram que isso deveria ter sido feito muito tempo antes. Os políticos locais recebiam doações generosas, e havia o cuidado de colocá-los sob uma lente favorável para incentivar esses investimentos na comunidade corporativa.

Ganhando poder

Os Ômegas tinham lançado uma empresa de mídia não apenas para financiar seus primeiros empreendimentos de tecnologia, mas também para o próximo passo de seu audacioso plano: dominar o mundo. Menos de um ano depois do primeiro lançamento, notáveis canais de notícias foram adicionados à sua programação em todo o mundo. Ao contrário de seus outros canais, esses foram deliberadamente projetados para perder dinheiro e lançados como um serviço público. De fato, seus canais de notícias não geravam receita nenhuma: não exibiam anúncios e estavam disponíveis gratuitamente para qualquer pessoa com conexão à internet. O restante do império de mídia deles era uma máquina de fazer dinheiro tão grande que era possível gastar muito mais recursos em seus serviços de notícias do que qualquer outro esforço jornalístico havia feito na história mundial – e isso ficou claro. Com recrutamento agressivo e salários altamente competitivos para jornalistas e repórteres investigativos, notáveis talentos e descobertas foram trazidos à tela. Por meio de um serviço global da web que pagava a quem revelasse

algo digno de nota, de corrupção numa região a um evento emocionante, geralmente eram os primeiros a dar uma notícia. Pelo menos, era nisso que as pessoas acreditavam: na verdade, costumavam ser os primeiros porque as histórias atribuídas aos jornalistas-cidadãos tinham sido descobertas pelo Prometheus por meio do monitoramento em tempo real da internet. Todos aqueles sites de notícias em vídeo também exibiam podcasts e artigos impressos.

A fase 1 da estratégia de notícias era ganhar a confiança das pessoas, o que foi feito com grande sucesso. Sua disposição sem precedentes para perder dinheiro permitiu uma cobertura de notícias regional e local notavelmente diligente, na qual jornalistas investigativos com frequência expunham escândalos que de fato engajavam seus espectadores. Sempre que um país estava fortemente dividido politicamente e acostumado a notícias partidárias, eles lançavam um canal de notícias para cada facção, de propriedade ostensiva de diferentes empresas, e gradualmente ganhavam a confiança dessa facção. Sempre que possível, conseguiam isso usando proxies para comprar os canais existentes mais influentes, melhorando-os aos poucos, removendo anúncios e introduzindo conteúdo próprio. Nos países em que a censura e a interferência política ameaçavam esses esforços, eles de início concordavam com qualquer coisa que o governo exigisse para se manterem em atividade, com o slogan interno secreto de "A verdade, nada além da verdade, mas talvez não toda a verdade". Prometheus costumava oferecer excelentes conselhos em tais situações, esclarecendo quais políticos precisavam ser apresentados sob uma boa luz e quais (geralmente corruptos da região) poderiam ser expostos. Prometheus também fornecia recomendações inestimáveis sobre quais atitudes tomar, quem subornar e qual era a melhor forma de fazê-lo.

Essa estratégia foi um sucesso esmagador em todo o mundo, com os canais controlados pelos Ômegas emergindo como as fontes de notícias mais confiáveis. Mesmo em países onde os governos até então tinham impedido sua adoção em massa, eles construíram uma reputação de confiabilidade, e muitas de suas notícias se espalhavam. Executivos de canais de notícias concorrentes sentiram que estavam travando uma batalha sem solução: como obter lucro competindo com alguém com melhor condição financeira que distribui seus produtos de graça? Com a queda da audiência, cada vez mais redes decidiram vender seus canais de notícias – geralmente para algum consórcio que mais tarde acabou sendo controlado pelos Ômegas.

Cerca de dois anos após o lançamento do Prometheus, quando a fase de ganho de confiança estava amplamente concluída, os Ômegas lançaram a fase 2 de sua estratégia de

notícias: persuasão. Mesmo antes disso, observadores astutos haviam notado indícios de uma agenda política por trás da nova mídia: parecia haver um leve empurrão em direção ao centro, longe de extremismo de todos os tipos. Sua infinidade de canais, que atendia a diferentes grupos, ainda refletia animosidade entre os Estados Unidos e a Rússia, a Índia e o Paquistão, religiões diferentes, facções políticas e assim por diante, mas as críticas eram um pouco atenuadas, geralmente concentradas em questões concretas que envolviam dinheiro e poder, e não em ataques *ad-hominem*, rumores alarmantes e pouco fundamentados. Quando a fase 2 começou com tudo, esse impulso para neutralizar conflitos antigos se tornou mais evidente, com frequentes histórias emocionantes sobre a situação dos adversários tradicionais misturadas com relatórios investigativos sobre quantos gananciosos interessados em causar conflitos eram impulsionados pela busca do lucro pessoal.

Os comentaristas políticos observaram que, em paralelo ao amortecimento dos conflitos regionais, parecia haver um esforço conjunto para reduzir as ameaças globais. Por exemplo, os riscos da guerra nuclear de repente passaram a ser discutidos em todo lugar. Vários filmes de grande sucesso apresentavam cenários em que a guerra nuclear global havia começado por acidente ou de propósito e dramatizavam as consequências distópicas com o inverno nuclear, o colapso da infraestrutura e a fome. Novos documentários detalhavam como o inverno nuclear podia impactar todos os países. Cientistas e políticos que defendiam a redução na escalada nuclear recebiam muito espaço na mídia, principalmente para discutir os resultados de vários novos estudos sobre quais medidas úteis poderiam ser tomadas – estudos financiados por organizações científicas que recebiam grandes doações de novas empresas de tecnologia. Como resultado, o momento político começou a aumentar a ponto de haver a necessidade de tirar mísseis de alerta de gatilho e encolher arsenais nucleares. A mídia também renovou o interesse nas mudanças climáticas globais, destacando frequentemente os recentes avanços tecnológicos habilitados por Prometheus, que estavam reduzindo o custo das energias renováveis e incentivando os governos a investir nessa nova infraestrutura de energia.

Paralelamente à aquisição da mídia, os Ômegas aproveitaram o Prometheus para revolucionar a educação. Dados o conhecimento e as habilidades de qualquer pessoa, o Prometheus poderia determinar a maneira mais rápida de aprender qualquer assunto novo de uma forma que mantivesse o engajamento e a motivação altos para continuar e produzir os vídeos, materiais de leitura, exercícios e outras ferramentas de aprendizado

otimizados correspondentes. Assim, as empresas controladas pelos Ômegas comercializavam cursos on-line sobre praticamente tudo, personalizados não apenas pelo idioma e pela formação cultural, mas também pelo nível inicial. Quer você fosse um analfabeto de 40 anos querendo aprender a ler ou um doutor em biologia pesquisando sobre o que há de mais recente acerca de imunoterapia contra câncer, Prometheus tinha o curso perfeito para você. Essas ofertas tinham pouca semelhança com a maioria dos cursos on-line atuais: aproveitando seus talentos de produção de filmes, os segmentos de vídeo eram envolventes de verdade e traziam metáforas poderosas que causavam identificação, deixando você ansioso por aprender mais. Alguns cursos eram vendidos com fins lucrativos, mas muitos eram disponibilizados gratuitamente, para o deleite dos professores de todo o mundo, que poderiam usá-los em sala de aula – e para a maioria das pessoas que desejavam aprender alguma coisa.

Essas superpotências educacionais provaram ser ferramentas poderosas para fins políticos, criando "sequências de persuasão" on-line com vídeos em que as ideias de cada um atualizavam os pontos de vista da pessoa e a motivavam a assistir a outro vídeo sobre um tópico relacionado, que provavelmente a deixaria mais convencida. Quando o objetivo fosse neutralizar um conflito entre duas nações, por exemplo, documentários históricos seriam lançados de maneira independente em ambos os países, explicando as origens e o percurso do conflito de modo mais sutil. As notícias pedagógicas explicavam quem, de cada lado, se beneficiaria com o conflito contínuo e suas técnicas para alimentá-lo. Ao mesmo tempo, personagens cativantes do outro país começariam a aparecer em programas populares nos canais de entretenimento, assim como personagens minoritários retratados com simpatia haviam reforçado os movimentos de direitos civis e LGBT no passado.

Em pouco tempo, os comentaristas políticos não puderam deixar de notar o crescente apoio a uma agenda política centrada em sete slogans:

1. Democracia
2. Corte de impostos
3. Cortes nos serviços sociais do governo
4. Cortes nos gastos militares
5. Livre-comércio
6. Fronteiras abertas

7. Empresas socialmente responsáveis

Menos óbvio era o objetivo subjacente: corroer todas as estruturas de poder anteriores no mundo. Os itens 2 a 6 exauriram o poder do Estado, e democratizar o mundo deu ao império empresarial dos Ômegas mais influência sobre a seleção de líderes políticos. As empresas socialmente responsáveis enfraqueceram ainda mais o poder do Estado, assumindo cada vez mais os serviços que os governos tinham (ou deveriam ter) fornecido. A elite empresarial tradicional enfraqueceu simplesmente porque não podia competir com as empresas apoiadas por Prometheus no livre mercado e, portanto, possuía uma parcela cada vez menor da economia mundial. Os líderes de opinião tradicionais – de partidos políticos a grupos religiosos – careciam do mecanismo de persuasão para competir com o império de mídia dos Ômegas.

Como em qualquer mudança radical, houve vencedores e perdedores. Embora houvesse um novo senso de otimismo palpável na maioria dos países à medida que a educação, os serviços sociais e a infraestrutura melhoravam, os conflitos cessavam e as empresas locais lançavam tecnologias inovadoras que varriam o mundo, nem todo mundo estava feliz. Enquanto muitos trabalhadores demitidos foram recontratados para projetos comunitários, aqueles que detinham grande poder e riqueza geralmente viram os dois encolher. Isso começou nos setores de mídia e tecnologia, mas se espalhou por praticamente toda parte. A redução dos conflitos mundiais levou a cortes no orçamento da defesa que prejudicaram os prestadores de serviços militares. As empresas iniciantes em expansão não costumavam ser negociadas publicamente, com a justificativa de que os acionistas maximizadores de lucro bloqueariam seus gastos maciços em projetos comunitários. Assim, o mercado de ações global continuou perdendo valor, ameaçando magnatas das finanças e cidadãos comuns que contavam com seus fundos de pensão. Como se os lucros cada vez menores das empresas de capital aberto não fossem ruins o suficiente, as empresas de investimento no mundo todo tinham notado uma tendência perturbadora: todos os seus antes bem-sucedidos algoritmos de negociação pareciam ter parado de funcionar, com desempenho abaixo de simples fundos de índice. Alguém lá fora sempre parecia enganá-los e vencê-los em seu próprio jogo.

Embora muitas pessoas poderosas resistissem à onda de mudanças, a resposta delas foi surpreendentemente ineficaz, quase como se tivessem caído em uma armadilha bem planejada. Grandes mudanças estavam acontecendo em um ritmo tão desconcertante que era difícil acompanhar e elaborar uma reação coordenada. Além disso, não estava muito

claro o que deveriam buscar. A direita política tradicional tinha visto a maioria de seus slogans ser cooptada, mas os cortes de impostos e o melhor clima comercial ajudavam principalmente seus concorrentes de tecnologia mais alta. Praticamente todas as indústrias tradicionais agora pediam resgate financeiro, mas os fundos governamentais limitados os colocavam em uma batalha sem esperança, enquanto a mídia os retratava como dinossauros em busca de subsídios estatais simplesmente porque não eram capazes de competir. A esquerda política tradicional se opôs ao livre-comércio e aos cortes nos serviços sociais do governo, mas encantou-se com os cortes militares e com a redução da pobreza. De fato, grande parte da cena deles foi roubada pelo fato inegável de os serviços sociais terem melhorado, agora que eram fornecidos por empresas idealistas e não pelo Estado. Pesquisas e mais pesquisas mostravam que a maioria dos eleitores em todo o mundo sentiu sua qualidade de vida melhorar e que as coisas estavam caminhando, de modo geral, em uma boa direção. Isso tinha uma explicação matemática simples: antes do Prometheus, os 50% mais pobres da população da Terra ganhavam apenas cerca de 4% da renda global, o que permitiu que as empresas controladas pelos Ômegas conquistassem seu coração (e seu voto) compartilhando apenas uma fração modesta de seus lucros.

Consolidação

Como resultado, muitas nações viram vitórias esmagadoras nas eleições dos partidos que defendiam os sete slogans dos Ômegas. Em campanhas cuidadosamente otimizadas, eles se colocaram no centro do espectro político, denunciando a direita como gananciosos vorazes em busca de ajuda e criticando a esquerda como grandes defensores da inovação nos impostos e gastos do governo. O que quase ninguém percebeu foi que o Prometheus selecionou cuidadosamente as pessoas ideais para se candidatar e tratou de garantir sua vitória.

Antes do Prometheus, havia um apoio crescente ao movimento universal de renda básica, que propunha uma renda mínima financiada por impostos para todas as pessoas como solução para o desemprego tecnológico. Esse movimento implodiu quando os projetos da comunidade corporativa decolaram, uma vez que o império empresarial controlado pelos Ômegas estava de fato fornecendo a mesma coisa. Com a desculpa de melhorar a coordenação de seus projetos comunitários, um grupo internacional de empresas lançou a Aliança Humanitária, uma organização não governamental com o

objetivo de identificar e financiar os esforços humanitários mais valiosos do mundo. Em pouco tempo, praticamente todo o império Ômega a apoiou e lançou projetos globais em uma escala sem precedentes, até mesmo em países que perderam amplamente o *boom* da tecnologia, melhorando a educação, a saúde, a prosperidade e a governança. Não é preciso dizer que o Prometheus ofereceu planos de projeto cuidadosamente elaborados nos bastidores, classificados por impacto positivo por dólar. Em vez de simplesmente distribuir dinheiro, como nas propostas de renda básica, a Aliança (como ficou conhecida informalmente) atrairia aqueles que ela apoiava no trabalho em prol de sua causa. Como resultado, uma grande parte da população mundial acabou se tornando grata e leal à Aliança – em geral mais do que a seu próprio governo.

Com o passar do tempo, a Aliança assumiu cada vez mais o papel de um governo mundial, à medida que os Estados nacionais viam seu poder ruir de forma contínua. Os orçamentos nacionais continuaram encolhendo devido a cortes de impostos, enquanto o orçamento da Aliança crescia para diminuir o de todos os governos juntos. Todos os papéis tradicionais dos governos nacionais tornaram-se cada vez mais redundantes e irrelevantes. A Aliança fornecia de longe os melhores serviços sociais, a melhor educação e a melhor infraestrutura. A mídia neutralizou o conflito internacional a ponto de os gastos militares se tornarem em grande parte desnecessários, e a crescente prosperidade eliminou a maioria das raízes dos conflitos antigos, que remontam à competição devido à escassez de recursos. Alguns ditadores e outras figuras resistiram violentamente a essa nova ordem mundial e se recusaram a ser comprados, mas todos foram derrubados em golpes cuidadosamente orquestrados ou revoltas em massa.

Os Ômegas haviam concluído a transição mais drástica da história da vida na Terra. Pela primeira vez, nosso planeta era governado por uma única potência, amplificada por uma inteligência tão vasta que poderia permitir que a vida prosperasse por bilhões de anos na Terra e por todo o nosso cosmos – mas qual era, especificamente, o plano deles?

***

Essa foi a história da equipe Ômega. O restante do livro conta outra história – uma que ainda não foi escrita: a história de nosso futuro com a IA. Como você gostaria que isso acontecesse? Algo remotamente parecido com a história dos Ômegas poderia de fato acontecer e, se assim for, você gostaria que acontecesse? Deixando de lado as especulações sobre a IA sobre-humana, como você

gostaria que nossa história começasse? Como você deseja que a IA tenha impacto nos empregos, nas leis e nas armas na próxima década? Olhando mais à frente, como você escreveria o final? Essa história tem proporções verdadeiramente cósmicas, pois envolve nada menos que o grande futuro da vida em nosso Universo. E é uma história que nós devemos escrever.

1

Bem-vindo à conversa mais importante do nosso tempo

"A tecnologia está dando à vida o potencial de florescer como nunca antes – ou de se autodestruir."
Future of Life Institute

Treze bilhões e oitocentos milhões de anos depois de seu nascimento, nosso Universo despertou e tomou consciência de si mesmo. De um pequeno planeta azul, minúsculas partes conscientes de nosso Universo começaram a olhar para o cosmos com telescópios, descobrindo repetidas vezes que tudo o que pensavam existir era apenas uma pequena parte de algo maior: um sistema solar, uma galáxia e um universo com mais de centenas de bilhões de outras galáxias dispostas em um elaborado padrão de grupos, aglomerados e superaglomerados. Embora esses observadores de estrelas cientes de si mesmos discordem em muitas coisas, eles tendem a concordar que essas galáxias são lindas e inspiradoras.

Mas a beleza está nos olhos de quem vê, não nas leis da física; portanto, antes de nosso Universo despertar, não havia beleza. Isso torna nosso despertar cósmico ainda mais maravilhoso e digno de comemoração: ele transformou nosso Universo, fazendo-o deixar de ser um zumbi irracional sem autoconsciência e se tornar um ecossistema vivo que abriga autorreflexão, beleza e esperança – e a busca por objetivos, significado e

propósito. Se nosso Universo nunca tivesse despertado, então, para mim, teria sido completamente inútil – apenas um gigantesco desperdício de espaço. Se nosso Universo voltar a dormir para sempre devido a alguma calamidade cósmica ou infortúnio autoinfligido, ele perderá o sentido, infelizmente.

Por outro lado, as coisas podem melhorar ainda mais. Até o momento, não sabemos se nós, humanos, somos os únicos observadores de estrelas em nosso cosmos, nem mesmo se somos os primeiros, mas já aprendemos o suficiente sobre nosso Universo para saber que ele tem o potencial de despertar muito mais plenamente do que tem feito até agora. Talvez sejamos como o primeiro lampejo de autoconsciência que você vivenciou quando começou a despertar hoje de manhã: uma premonição da consciência muito maior que ocorreria quando você abrisse os olhos e acordasse por completo. Talvez a vida se espalhe por todo o nosso cosmos e floresça por bilhões ou trilhões de anos – e talvez seja por causa das decisões que tomamos aqui, em nosso pequeno planeta, durante a nossa existência.

Uma breve história da complexidade

Então, como esse incrível despertar aconteceu? Não foi um evento isolado, mas apenas um passo em um implacável processo de 13,8 bilhões de anos que está tornando nosso Universo cada vez mais complexo e interessante – e continua em ritmo acelerado.

Como físico, me sinto sortudo por ter passado boa parte do último quarto de século ajudando a estabelecer nossa história cósmica, e tem sido uma incrível jornada de descoberta. Desde os dias em que eu era estudante de pós-graduação, deixamos de discutir se nosso Universo tem 10 ou 20 bilhões de anos e passamos a discutir se tem 13,7 ou 13,8 bilhões de anos, isso graças a uma combinação de telescópios melhores, computadores melhores e melhor entendimento. Nós, físicos, ainda não sabemos ao certo o que causou nosso Big Bang, ou se ele de fato foi o começo de tudo, ou apenas a sequência de um estágio anterior. No entanto, adquirimos uma compreensão bastante detalhada do que aconteceu *desde* o nosso Big Bang, graças a uma avalanche de medições de alta qualidade, então, deixe-me dedicar alguns minutos para resumir 13,8 bilhões de anos de história cósmica.

No começo, havia luz. Na primeira fração de segundo após nosso Big Bang, toda a parte do espaço que nossos telescópios conseguem observar a princípio ("nosso Universo

observável" ou simplesmente "nosso Universo", para abreviar) era muito mais quente e brilhante que o núcleo do nosso Sol e se expandiu rapidamente. Embora possa parecer espetacular, também era entediante no sentido de que nosso Universo não continha nada além de uma sopa chata, sem vida, densa, quente e uniforme de partículas elementares. As coisas pareciam praticamente iguais em todos os lugares, e a única estrutura interessante consistia em ondas sonoras fracas de aparência aleatória que tornavam a sopa cerca de 0,001% mais densa em alguns lugares. Acredita-se que essas ondas fracas tenham se originado como as chamadas flutuações quânticas, porque o Princípio da Incerteza de Heisenberg, dentro da mecânica quântica, proíbe qualquer coisa de ser completamente chata e uniforme.

À medida que se expandia e esfriava, nosso Universo se tornava mais interessante com suas partículas formando objetos cada vez mais complexos. Durante a primeira fração de segundo, a imensa força nuclear agrupou quarks em prótons (núcleos de hidrogênio) e nêutrons, alguns dos quais, por sua vez, se fundiram em núcleos de hélio em poucos minutos. Cerca de 400 mil anos depois, a força eletromagnética agrupou esses núcleos com elétrons para formar os primeiros átomos. Conforme nosso Universo continuava se expandindo, esses átomos gradualmente esfriaram e se transformaram em um gás escuro e frio, e a escuridão dessa primeira noite durou cerca de 100 milhões de anos. Essa longa noite deu origem ao nosso amanhecer cósmico, quando a força gravitacional conseguiu amplificar essas flutuações no gás, juntando átomos para formar as primeiras estrelas e galáxias. Essas primeiras estrelas geraram calor e luz por meio da fusão do hidrogênio em átomos mais pesados, como carbono, oxigênio e silício. Quando essas estrelas morreram, muitos dos átomos que elas criaram foram reciclados no cosmos e formaram planetas em torno de estrelas da segunda geração.

Em algum momento, um grupo de átomos foi organizado em um padrão complexo que poderia se manter e se replicar. Assim, logo havia duas cópias, e o número não parava de dobrar. São necessárias apenas quarenta duplicações para gerar um trilhão, então esse primeiro autorreplicador logo se tornou uma força considerável. A vida tinha chegado.

Os três estágios da vida

A questão sobre como definir a vida é sabidamente controversa. As definições concorrentes são abundantes, algumas das quais incluem requisitos altamente

específicos, como ser composto de células, que podem desqualificar tanto as futuras máquinas inteligentes quanto as civilizações extraterrestres. Como não queremos limitar nosso pensamento sobre o futuro da vida às espécies que encontramos até agora, vamos definir a vida de maneira muito ampla, simplesmente como um processo que pode reter sua complexidade e replicar. O que é replicado não é matéria (feita de átomos), mas informação (feita de bits) especificando como os átomos são organizados. Quando uma bactéria faz uma cópia de seu DNA, nenhum novo átomo é criado, mas um novo conjunto de átomos é organizado no mesmo padrão que o original, copiando a informação. Em outras palavras, podemos pensar na vida como um sistema de processamento de dados autorreplicável cujas informações (software) determinam seu comportamento e os diagramas de seu hardware.

Como nosso Universo, a vida gradualmente se tornou mais complexa e interessante,[2] e, como vou explicar agora, acho útil classificar as formas em três níveis de sofisticação: Vida 1.0, 2.0 e 3.0. Resumi esses três níveis na Figura 1.1.

Ainda são questões não respondidas como, quando e onde a vida apareceu pela primeira vez em nosso Universo, mas há fortes evidências de que, aqui na Terra, a vida tenha aparecido pela primeira vez há cerca de 4 bilhões de anos. Em pouco tempo, nosso planeta estava repleto de uma enorme variedade de formas de vida. Os mais bem-sucedidos, que logo superaram o restante, foram capazes de reagir ao ambiente de alguma maneira. Especificamente, eles eram o que os cientistas da computação chamam de "agentes inteligentes": entidades que coletam informações sobre o ambiente a partir de sensores e, em seguida, processam essas informações para decidir como reagir a ele. Isso pode incluir processamento de informações altamente complexas, como quando você usa informações de seus olhos e ouvidos para decidir o que dizer em uma conversa. Mas também pode envolver hardware e software muito simples.

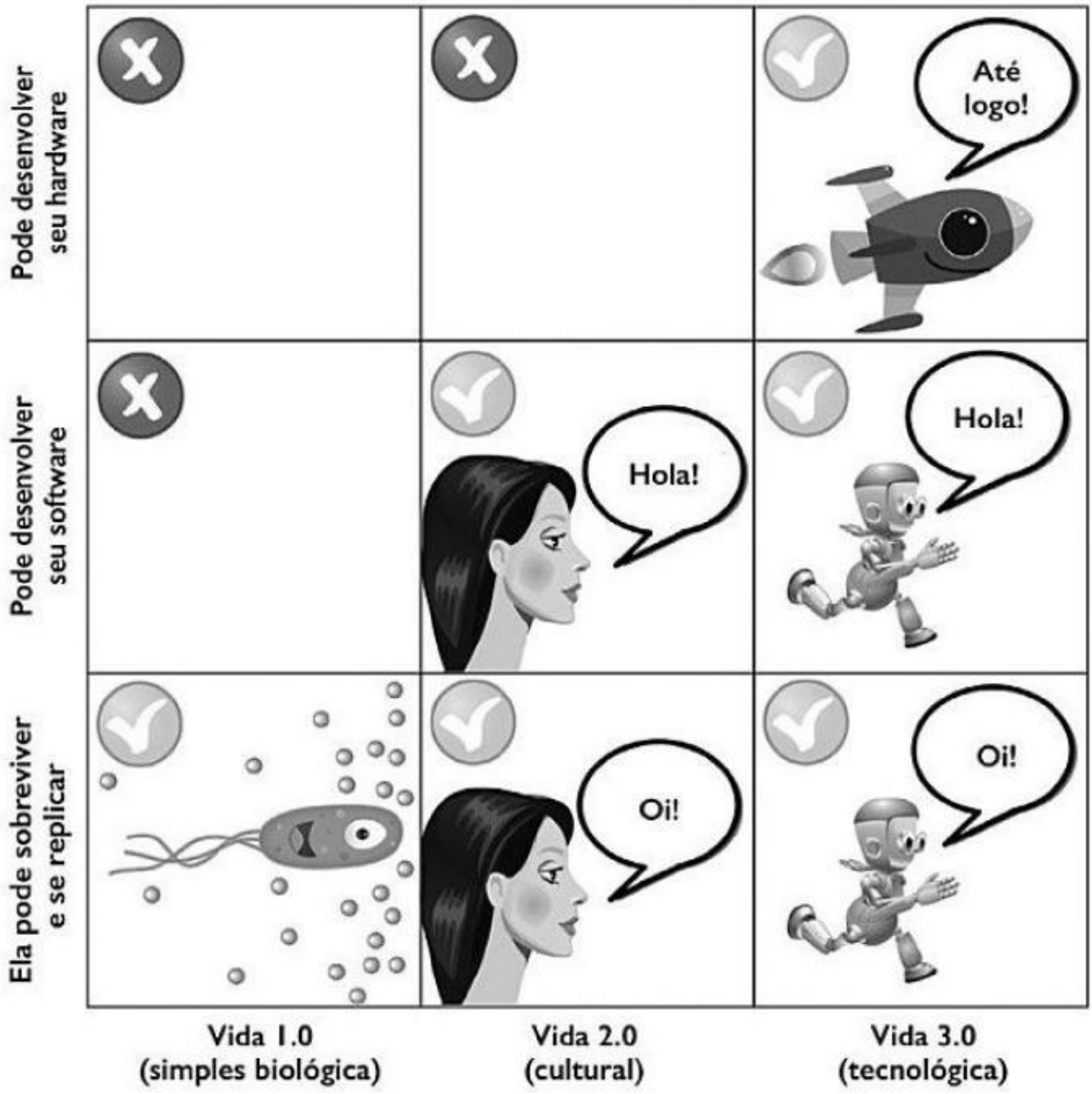


Figura 1.1: Os três estágios da vida: evolução biológica, evolução cultural e evolução tecnológica. A Vida 1.0 é incapaz de reprojetar seu hardware ou software durante sua vida útil: os dois são determinados pelo DNA e mudam apenas pela evolução ao longo de muitas gerações. Por outro lado, a Vida 2.0 pode recriar grande parte de seu software: os humanos podem aprender novas habilidades complexas – idiomas, esportes e profissões, por exemplo – e podem atualizar fundamentalmente suas visões de mundo e seus objetivos. A Vida 3.0, que ainda não existe na Terra, pode recriar drasticamente não apenas seu software, mas também seu hardware, em vez de esperar que ele evolua gradualmente ao longo de gerações.

Por exemplo, muitas bactérias têm um sensor que mede a concentração de açúcar no líquido ao seu redor e podem nadar usando estruturas em forma de hélice chamadas flagelos. O hardware que liga o sensor ao flagelo pode implementar o seguinte algoritmo simples, mas útil: "Se meu sensor de concentração de açúcar indicar um valor menor que

o de alguns segundos atrás, vou inverter a rotação de meus flagelos para mudar de direção".

Você aprendeu a falar e a realizar inúmeras outras habilidades. As bactérias, por outro lado, não são grandes aprendizes. O DNA delas especifica não apenas o design de seu hardware, como sensores de açúcar e flagelos, mas também o design de seu software. Elas nunca aprendem a nadar em direção ao açúcar; esse algoritmo foi codificado no DNA delas desde o início. Obviamente, houve um tipo de processo de aprendizado, mas ele não ocorreu durante a vida dessa bactéria em particular. Na verdade, ocorreu durante a evolução anterior dessa espécie de bactéria, por meio de um lento processo de tentativa e erro, que abrange muitas gerações, em que a seleção natural favoreceu aquelas mutações aleatórias no DNA que melhoravam o consumo de açúcar. Algumas dessas mutações ajudaram a aprimorar o design dos flagelos e de outros hardwares, enquanto outras melhoraram o sistema bacteriano de processamento de informações que implementa o algoritmo de busca de açúcar e outros softwares.

Essas bactérias são um exemplo do que chamarei de "Vida 1.0": *vida em que tanto o hardware quanto o software são resultado da evolução, em vez de projetados*. Você e eu, por outro lado, somos exemplos da "Vida 2.0": *vida cujo hardware é resultado da evolução, mas cujo software é amplamente projetado*. Por software, aqui, quero dizer todos os algoritmos e conhecimentos que você usa para processar as informações dos seus sentidos e decidir o que fazer – tudo, desde a capacidade de reconhecer seus amigos quando os vê até a capacidade de caminhar, ler, escrever, calcular, cantar e contar piadas.

Você não era capaz de executar nenhuma dessas tarefas quando nasceu; portanto, todo esse software foi programado em seu cérebro *a posteriori* por meio do processo que chamamos de aprendizado. Embora seu currículo na infância seja amplamente projetado por sua família e seus professores, que decidem o que você deve aprender, você gradualmente adquire mais poder para projetar seu próprio software. Talvez sua escola permita que você escolha um idioma estrangeiro: deseja instalar um módulo de software em seu cérebro que faça com que fale francês ou um que faça com que fale espanhol? Quer aprender a jogar tênis ou xadrez? Quer estudar para se tornar chef, advogado ou farmacêutico? Quer aprender mais sobre inteligência artificial e o futuro da vida lendo um livro sobre esse assunto?

Essa capacidade da Vida 2.0 de projetar seu software permite que ela seja muito mais inteligente que a Vida 1.0. Alta inteligência requer muito hardware (feito de átomos) e muito software (feito de bits). O fato de a maior parte do nosso hardware humano ser

adicionado após o nascimento (por meio do crescimento) é útil, já que nosso tamanho final não é limitado pela largura do canal de nascimento de nossa mãe. Da mesma forma, o fato de a maior parte do nosso software humano ser adicionado após o nascimento (por meio da aprendizagem) é útil, uma vez que nossa inteligência final não é limitada pela quantidade de informações que nos podem ser transmitidas na concepção por meio do nosso DNA, estilo 1.0. Eu peso cerca de 25 vezes mais do que quando nasci, e as conexões sinápticas que ligam os neurônios do meu cérebro podem armazenar cerca de 100 mil vezes mais informações do que o DNA com o qual nasci. Suas sinapses armazenam todo o seu conhecimento e habilidades em aproximadamente 100 terabytes de informações, enquanto o seu DNA armazena apenas cerca de 1 gigabyte, o suficiente para armazenar um único download de filme. Portanto, é fisicamente impossível para uma criança nascer falando inglês perfeitamente e pronta para gabaritar seus exames de admissão na faculdade: não há como a informação ter sido pré-carregada em seu cérebro, pois o principal módulo de informação que ela recebeu dos pais (seu DNA) não tem capacidade suficiente de armazenamento de informação.

A capacidade de projetar seu software permite que a Vida 2.0 seja não apenas mais inteligente que a Vida 1.0, mas também mais flexível. Se o ambiente mudar, a 1.0 só poderá se adaptar evoluindo lentamente ao longo de muitas gerações. A Vida 2.0, por outro lado, pode se adaptar quase instantaneamente, por meio de uma atualização de software. Por exemplo, bactérias que costumam encontrar antibióticos podem desenvolver resistência a medicamentos por muitas gerações, mas uma única bactéria não altera seu comportamento; em contrapartida, uma garota que descobre que tem alergia a amendoim muda imediatamente seu comportamento para começar a evitá-lo. Essa flexibilidade confere à Vida 2.0 uma vantagem ainda maior no nível da população: embora as informações em nosso DNA humano não tenham evoluído drasticamente nos últimos 50 mil anos, as informações armazenadas coletivamente em nossos cérebros, livros e computadores aumentaram muito. Ao instalar um módulo de software que nos permite nos comunicar por meio de uma sofisticada linguagem falada, garantimos que as informações mais úteis armazenadas no cérebro de uma pessoa possam ser copiadas para outros cérebros, sobrevivendo potencialmente, mesmo após a morte do cérebro original. Ao instalar um módulo de software que nos permite ler e escrever, conseguimos armazenar e compartilhar muito mais informações do que as pessoas conseguem memorizar. Ao desenvolver um software cerebral capaz de produzir tecnologia (ou seja,

estudando ciências e engenharia), permitimos que muitas das informações do mundo sejam acessadas por boa parte dos seres humanos com apenas alguns cliques.

Essa flexibilidade permitiu que a Vida 2.0 dominasse a Terra. Livre de suas amarras genéticas, o conhecimento combinado da humanidade continuou crescendo em um ritmo acelerado, à medida que cada avanço possibilitava o seguinte: idioma, escrita, prensa, ciência moderna, computadores, internet e assim por diante. Essa evolução cultural cada vez mais rápida de nosso software compartilhado surgiu como a força dominante que molda nosso futuro humano, tornando nossa evolução biológica absurdamente lenta quase irrelevante.

No entanto, apesar das tecnologias mais poderosas que temos hoje, todas as formas de vida que conhecemos se mantêm fundamentalmente limitadas por seu hardware biológico. Ninguém pode viver por um milhão de anos, memorizar toda a Wikipédia, entender toda ciência conhecida ou desfrutar de voos espaciais sem uma espaçonave. Ninguém pode transformar nosso cosmos, em grande parte sem vida, em uma biosfera diversa que vai florescer por bilhões ou trilhões de anos, permitindo que nosso Universo finalmente cumpra seu potencial e acorde por completo. Tudo isso requer que a vida seja submetida a uma atualização final, para a Vida 3.0, que pode projetar não apenas seu software, mas também seu hardware. Em outras palavras, a Vida 3.0 é o mestre de seu próprio destino, finalmente 100% livre de seus grilhões evolutivos.

Os limites entre os três estágios da vida são um pouco difusos. Se as bactérias são a Vida 1.0 e os seres humanos são a Vida 2.0, então é possível classificar os ratos como 1.1: eles podem aprender muitas coisas, mas não o suficiente para desenvolver a linguagem ou inventar a internet. Além disso, como eles não têm linguagem, o que aprendem se perde em grande parte quando morrem, porque não é repassado para a próxima geração. Da mesma forma, você pode argumentar que os humanos de hoje devem ser classificados como Vida 2.1: podemos realizar pequenas atualizações de hardware, como implantar dentes artificiais, joelhos e marca-passos, mas nada tão drástico quanto aumentar nossa estatura em dez vezes ou adquirir cérebros mil vezes maiores.

Em resumo, podemos dividir o desenvolvimento da vida em três estágios, distinguidos pela capacidade da vida de se projetar:

- Vida 1.0 (estágio biológico): o hardware e o software são resultado da evolução;
- Vida 2.0 (estágio cultural): o hardware é resultado da evolução, o software é, em grande parte, projetado;

- Vida 3.0 (estágio tecnológico): o hardware e o software são projetados.

Após 13,8 bilhões de anos de evolução cósmica, o desenvolvimento acelerou drasticamente aqui na Terra: a Vida 1.0 chegou cerca de 4 bilhões de anos atrás, a Vida 2.0 (nós humanos) chegou há aproximadamente cem milênios, e muitos pesquisadores de IA acreditam que a Vida 3.0 pode chegar ao longo do próximo século, talvez durante a nossa geração, graças ao progresso na IA. O que vai acontecer e o que isso significa para nós? Esse é o assunto deste livro.

Polêmicas

Essa questão é maravilhosamente polêmica, com os principais pesquisadores de IA do mundo discordando de modo apaixonado não apenas em suas previsões, mas também em suas reações emocionais, que variam de otimismo confiante a sérias preocupações. Eles nem sequer chegam a um consenso sobre questões de curto prazo sobre o impacto econômico, jurídico e militar da IA, e suas divergências aumentam quando expandimos o horizonte de tempo e perguntamos sobre *inteligência artificial geral* (IAG) – em especial sobre a IAG atingir o nível humano e além, possibilitando a Vida 3.0. A *inteligência geral* pode atingir praticamente qualquer objetivo, inclusive aprender, em contraste com, digamos, a inteligência limitada de um programa de xadrez.

Curiosamente, a controvérsia sobre a Vida 3.0 gira em torno de não uma, mas de duas perguntas distintas: quando e o quê? Quando acontecerá (se é que acontecerá) e o que isso significará para a humanidade? Do meu ponto de vista, existem três escolas de pensamento distintas que precisam ser levadas a sério, pois cada uma inclui vários especialistas reconhecidos mundialmente. Como ilustrado na Figura 1.2, penso neles como *utopistas digitais*, *tecnocéticos* e *membros do movimento da IA benéfica*, respectivamente. Vou apresentar alguns de seus campeões de maior destaque.

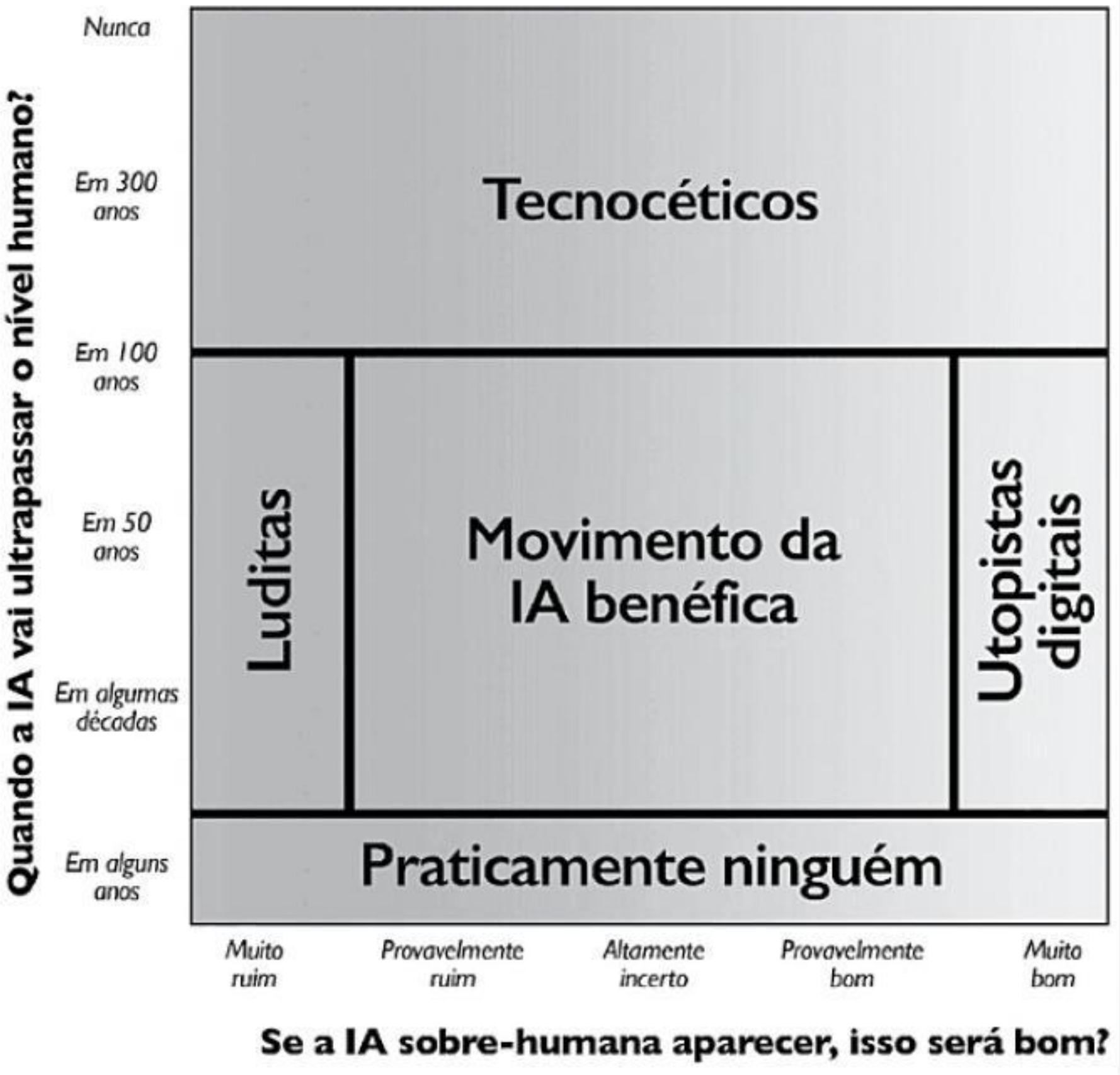


Figura 1.2: A maioria das polêmicas em torno da inteligência artificial forte (que pode corresponder aos humanos em qualquer tarefa cognitiva) se concentra em duas perguntas: quando (ou *se*) isso vai acontecer e será uma coisa boa para a humanidade? Os tecnocéticos e os utopistas digitais concordam que não devemos nos preocupar, mas por razões muito diferentes: os primeiros estão convencidos de que a inteligência artificial geral (IAG) em nível humano não acontecerá num futuro próximo, enquanto os últimos pensam que isso vai acontecer, mas que é praticamente garantido que será uma coisa boa. O movimento da IA benéfica considera que a preocupação é justificada e útil, porque a pesquisa e a discussão sobre segurança da IA agora aumentam as chances de um bom resultado. Os luditas estão convencidos de um resultado ruim e se opõem à IA. Esta figura é parcialmente inspirada em Tim Urban.[3]

Utopistas digitais

Quando eu era criança, imaginava que bilionários exalavam pompa e arrogância. Quando conheci Larry Page no Google, em 2008, ele rompeu totalmente com esse estereótipo.


image
not
available

Outra lição importante da conferência foi que as questões levantadas pelo sucesso da IA não são apenas intelectualmente fascinantes; também são moralmente cruciais, porque nossas escolhas têm o potencial de afetar todo o futuro da vida. O significado moral das escolhas passadas da humanidade às vezes era grande, mas sempre limitado: tínhamos nos recuperado até das maiores pragas, e até mesmo os maiores impérios acabaram desmoronando. As gerações passadas sabiam que, assim como o Sol nasceria amanhã, também nasceriam os humanos de amanhã, enfrentando flagelos perenes, como pobreza, doenças e guerra. Mas alguns dos palestrantes da conferência em Porto Rico argumentaram que desta vez podia ser diferente: pela primeira vez, disseram, poderíamos construir uma tecnologia poderosa o suficiente para acabar definitivamente com esses flagelos – ou com a própria humanidade. Podemos criar sociedades que floresçam como nunca, na Terra e talvez além, ou um estado de vigilância global kafkiano tão poderoso que nunca poderá ser derrubado.

Figura 1.4: Embora a mídia muitas vezes retrate Elon Musk como uma figura em desacordo com a comunidade da IA, existe na verdade um amplo consenso de que uma pesquisa sobre segurança da IA é necessária. Aqui, em 4 de janeiro de 2015, Tom Dietterich, presidente da Associação para o Avanço da Inteligência Artificial, compartilha a empolgação de Elon acerca do novo programa de pesquisa sobre segurança de IA que o empresário tinha

nuclear em cadeia por Leo Szilard – que a energia nuclear era "tolice" e, em 1956, o astrônomo real Richard Woolley chamou as conversas sobre viagens espaciais de "perda de tempo". A forma mais extrema desse mito é que a IAG sobre-humana nunca chegará por ser fisicamente impossível. No entanto, os físicos sabem que um cérebro consiste em quarks e elétrons organizados para agir como um computador poderoso, e que não existe lei da física que nos impeça de construir blobs de quarks ainda mais inteligentes.

Houve vários estudos que perguntaram aos pesquisadores de IA em quantos anos eles achavam que teríamos IAG em nível humano com pelo menos 50% de probabilidade, e todas essas pesquisas chegam à mesma conclusão: os principais especialistas do mundo discordam, então simplesmente não sabemos. Por exemplo, em um estudo realizado com pesquisadores de IA na conferência de Porto Rico, a resposta média (mediana) foi até 2055, mas alguns pesquisadores previram centenas de anos ou mais.

Existe também um mito relacionado de que as pessoas preocupadas com a IA acham que faltam apenas alguns anos para sua chegada. De fato, a maioria daqueles que se preocupam com a IAG sobre-humana acha que ela ainda está a décadas de distância, pelo menos. Mas argumentam que, como não temos 100% de certeza de que ela acontecerá neste século, é inteligente começar uma pesquisa de segurança agora para estarmos preparados para a possibilidade. Como veremos neste livro, muitos dos problemas de segurança são tão difíceis que podem levar décadas para serem resolvidos; portanto, é prudente começar a pesquisá-los agora, e não um dia antes de alguns programadores regados a bebidas energéticas decidirem acionar a IAG de nível humano.

Mitos controversos

Outro equívoco comum é que as únicas pessoas que alimentam preocupações sobre a IA e defendem a pesquisa sobre segurança da IA são luditas que não sabem muito sobre o tema. Quando Stuart Russell mencionou isso durante sua palestra em Porto Rico, a plateia riu alto. Um equívoco relacionado é que o apoio à pesquisa sobre segurança da IA é extremamente controverso. De fato, para apoiar um investimento modesto em pesquisa sobre segurança de IA, as pessoas não precisam estar convencidas de que os riscos são altos, apenas não desprezíveis, assim como um investimento modesto em seguro residencial é justificado por uma probabilidade não desprezível de uma casa pegar fogo.


image
not
available

Mito:
A superinteligência será inevitável até 2100

Mito:
A superinteligência será impossível até 2100

Fato:
Pode acontecer em décadas, séculos ou nunca: Os especialistas em IA discordam, e nós simplesmente não sabemos

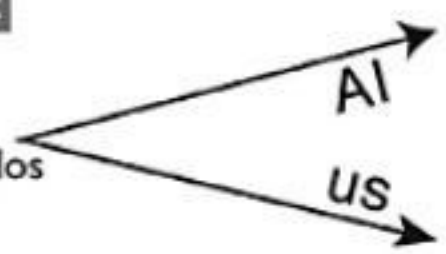
Mito:
Apenas os ludistas se preocupam com a IA

Fato:
Muitos pesquisadores importantes de IA estão preocupados

Preocupação mítica:
IA se tornando maléfica

Preocupação mítica:
IA tornando-se consciente

Preocupação real:
IA tornando-se competente, com objetivos desalinhados com os nossos



Mito:
Robôs são a principal preocupação

Fato:
Inteligência desalinhada é a principal preocupação: não precisa de corpo, apenas de conexão com a internet

Mito:
IA não consegue controlar humanos

Fato:
A inteligência permite o controle: controlamos tigres sendo mais inteligentes

Mito:
As máquinas não podem ter objetivos

Fato:
Um míssil guiado pelo calor tem objetivo

Preocupação mítica:
A superinteligência vai levar anos

Preocupação real:
Está a pelo menos décadas de distância, mas talvez seja preciso todo esse tempo para que seja segura

realizado. Existem três campos principais na controvérsia: tecnocéticos, utopistas digitais e o movimento da IA benéfica.

- Os tecnocéticos veem a construção da IAG sobre-humana como algo tão difícil que não acontecerá nas próximas centenas de anos, tornando tolice se preocupar com ela (e com a Vida 3.0) agora.
- Os utopistas digitais a veem como provável neste século e acolhem com sinceridade a Vida 3.0, enxergando-a como o próximo passo natural e desejável na evolução cósmica.
- O movimento da IA benéfica também a vê como provável neste século, mas enxerga um bom resultado como algo não tão garantido, mas que precisa de muito trabalho em forma de pesquisa sobre segurança da IA.
- Além de tais controvérsias legítimas, em que os principais especialistas do mundo discordam, também existem pseudopolêmicas entediantes causadas por mal-entendidos. Por exemplo, nunca perca tempo discutindo "vida", "inteligência" ou "consciência" sem ter certeza de que você e seu interlocutor estejam usando esses termos com o mesmo sentido! Este livro usa as definições da Tabela 1.1.
- Também tome cuidado com os equívocos comuns na Figura 1.5: "A superinteligência será inevitável/impossível até 2100", "Apenas os luditas se preocupam com a IA", "A preocupação é que a IA se torne má e/ou consciente, e daqui a apenas alguns anos", "Os robôs são a principal preocupação", "A IA não pode controlar humanos e não pode ter objetivos".
- Nos Capítulos 2 a 6, vamos explorar a história da inteligência desde seu humilde começo, bilhões de anos atrás, até possíveis futuros cósmicos, bilhões de anos a partir de agora. Primeiro, vamos investigar desafios de curto prazo, como empregos, armas de inteligência artificial e a busca por IAG em nível humano, depois vamos nos debruçar sobre as possibilidades de um fascinante espectro de futuros possíveis com máquinas inteligentes e/ou humanos. Fico imaginando quais opções você prefere!
- Nos Capítulos de 7 a 9, passaremos de descrições factuais simples para uma exploração de objetivos, consciência e significado e vamos investigar o que podemos fazer agora para ajudar a criar o futuro que queremos.
- Vejo essa conversa sobre o futuro da vida com a IA como a mais importante do nosso tempo – venha participar dela!

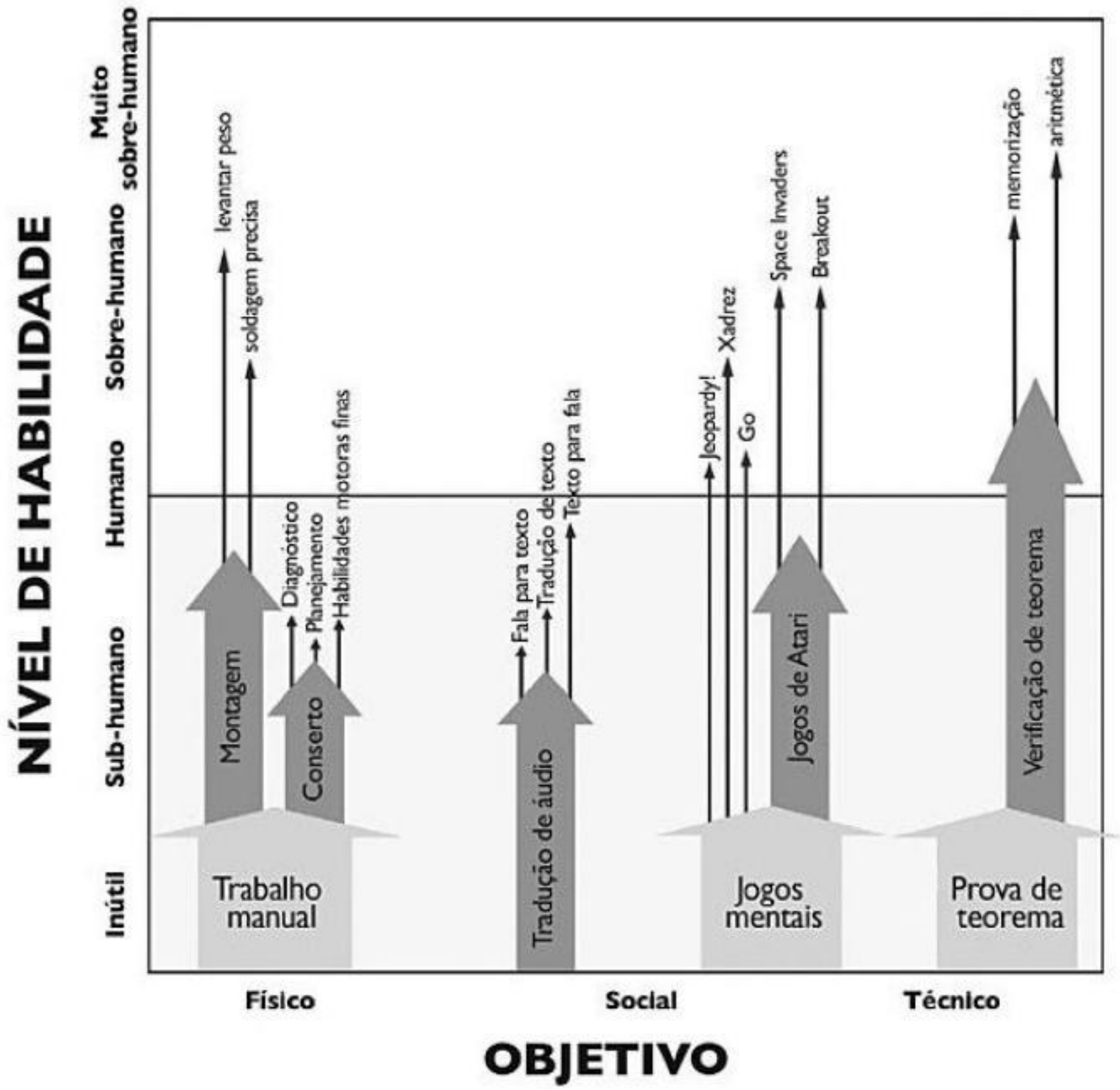


Figura 2.1: Inteligência, definida como a capacidade de atingir objetivos complexos, não pode ser medida por um simples QI, apenas por um espectro de habilidades por todos os objetivos. Cada seta indica a capacidade atual dos melhores sistemas de IA de atingir diversos objetivos, ilustrando que a inteligência artificial de hoje costuma ser *limitada*, com cada sistema capaz de realizar apenas objetivos muito específicos. Em contraste, a inteligência humana é notoriamente ampla: uma criança saudável pode aprender a se aperfeiçoar em quase qualquer coisa.

Embora a palavra "inteligência" tenda a conotações positivas, é importante notar que a estamos usando de uma maneira completamente neutra em termos de valor: como capacidade de atingir objetivos complexos, independentemente de esses objetivos serem considerados bons ou ruins. Assim, uma pessoa inteligente pode ser muito boa em ajudar outras pessoas ou muito boa em fazer mal a elas. Vamos explorar a questão dos objetivos no Capítulo 7. Em relação aos objetivos, também precisamos esclarecer a sutileza dos objetivos a que estamos nos referindo. Suponha que o seu futuro assistente pessoal

Que propriedade física fundamental todos eles têm em comum que os torna úteis como dispositivos de memória, ou seja, dispositivos para armazenar informações? A resposta é que todos podem estar em *estados duradouros diferentes* – duradouros o suficiente para codificar as informações até que sejam necessárias. Vamos ver um exemplo simples, suponha que você coloque uma bola em uma superfície montanhosa com 16 vales diferentes, como na Figura 2.3. Quando tiver rolado para baixo e parado, a bola estará em um dos 16 locais, então você poderá usar a posição dela como uma maneira de lembrar qualquer número entre 1 e 16.

Esse dispositivo de memória é bastante robusto, porque, mesmo que seja um pouco sacudido e perturbado por forças externas, é provável que a bola permaneça no mesmo vale em que você a colocou, então você ainda pode saber qual número está sendo armazenado. A razão pela qual essa memória é tão estável é que tirar a bola de seu vale requer mais energia do que distúrbios aleatórios provavelmente vão fornecer. Essa mesma ideia pode oferecer memórias estáveis de modo muito mais genérico do que para uma bola móvel: a energia de um sistema físico complicado pode depender de todos os tipos de propriedades mecânicas, químicas, elétricas e magnéticas, e, enquanto for necessária energia para afastar o sistema do estado de que você deseja que ele se lembre, esse estado será estável. É por isso que os sólidos têm muitos estados de vida longa, enquanto líquidos e gases, não: se você gravar o nome de alguém em um anel de ouro, as informações ainda estarão lá anos depois, porque a remodelagem do ouro requer uma energia significativa, mas se você as gravar na superfície de uma lagoa, ele vai se perder em um segundo, quando a superfície da água mudar sua forma sem esforço.

O dispositivo de memória mais simples possível tem apenas dois estados estáveis (Figura 2.3, ao lado). Podemos, portanto, pensar nisso como a codificação de um dígito binário (ou bit, abreviação de *binary digit*), ou seja, zero ou um. As informações armazenadas por qualquer dispositivo de memória mais complicado podem ser armazenadas de modo equivalente em múltiplos bits: por exemplo, juntos, os quatro bits mostrados na Figura 2.3 (à direita) podem estar em $2 \times 2 \times 2 \times 2 = 16$ estados diferentes 0000, 0001, 0010, 0011, ..., 1111, então, em conjunto, eles têm exatamente a mesma capacidade de memória do sistema mais complicado de 16 estados (à esquerda). Podemos, portanto, considerar os bits como átomos de informação – o menor pedaço indivisível de informação, que pode ser combinado para formar qualquer informação. Por exemplo, apenas digitei a palavra "*word*", e meu laptop a representou em sua memória como a sequência de quatro números "119 111 114 100", armazenando cada um desses números

E os dispositivos de memória que evoluíram, em vez de serem projetados por humanos? Os biólogos ainda não sabem qual foi a primeira forma de vida que copiou seus projetos entre gerações, mas pode ter sido bem pequena. Uma equipe liderada por Philipp Holliger, na Universidade de Cambridge, criou uma molécula de RNA em 2016 que codificava 412 bits de informação genética e era capaz de copiar as cadeias de RNA por mais tempo do que ela própria, reforçando a hipótese do "mundo do RNA" de que a vida na Terra primitiva envolveu pequenos fragmentos de RNA autorreplicantes. Até agora, o menor dispositivo de memória conhecido por evoluir e ser utilizado na natureza é o genoma da bactéria *Candidatus Carsonella ruddii*, que armazena cerca de 40 kilobytes, enquanto o nosso DNA humano armazena cerca de 1,6 gigabytes, comparável a um filme baixado. Como mencionado no capítulo anterior, nosso cérebro armazena muito mais informações que nossos genes: aproximadamente 10 gigabytes eletricamente (especificando quais de seus 100 bilhões de neurônios estão disparando a qualquer momento) e 100 terabytes química/biologicamente (especificando quão fortemente diferentes neurônios estão ligados por sinapses). A comparação desses números com a memória das máquinas mostra que os melhores computadores do mundo agora podem equivaler a qualquer sistema biológico – a um custo que está caindo rapidamente e chegou a alguns milhares de dólares em 2016.

A memória do seu cérebro funciona de maneira muito diferente da memória do computador, não apenas em termos de como é construída, mas também de como é usada. Considerando que você recupera memórias de um computador ou disco rígido especificando *onde* estão armazenadas, você recupera memórias do seu cérebro especificando algo sobre *o que* está armazenado. Cada grupo de bits na memória do computador tem um endereço numérico, e, para recuperar uma informação, o computador especifica em que endereço procurar, como se eu dissesse: "Vá até a minha estante, pegue o quinto livro da direita na prateleira superior e me diga o que está escrito na página 314". Por outro lado, você recupera informações do seu cérebro de maneira semelhante à de um mecanismo de pesquisa: você especifica uma parte da informação ou algo relacionado a ela, e ela é exibida. Se eu disser "Ser ou não" para você, ou se pesquisar no Google, é provável que o resultado alcançado seja "Ser ou não ser, eis a questão". De fato, é provável que funcione mesmo que eu use outra parte da citação ou embaralhe um pouco as coisas. Esses sistemas de memória são chamados *autoassociativo*, uma vez que se lembram por associação, e não por endereço.

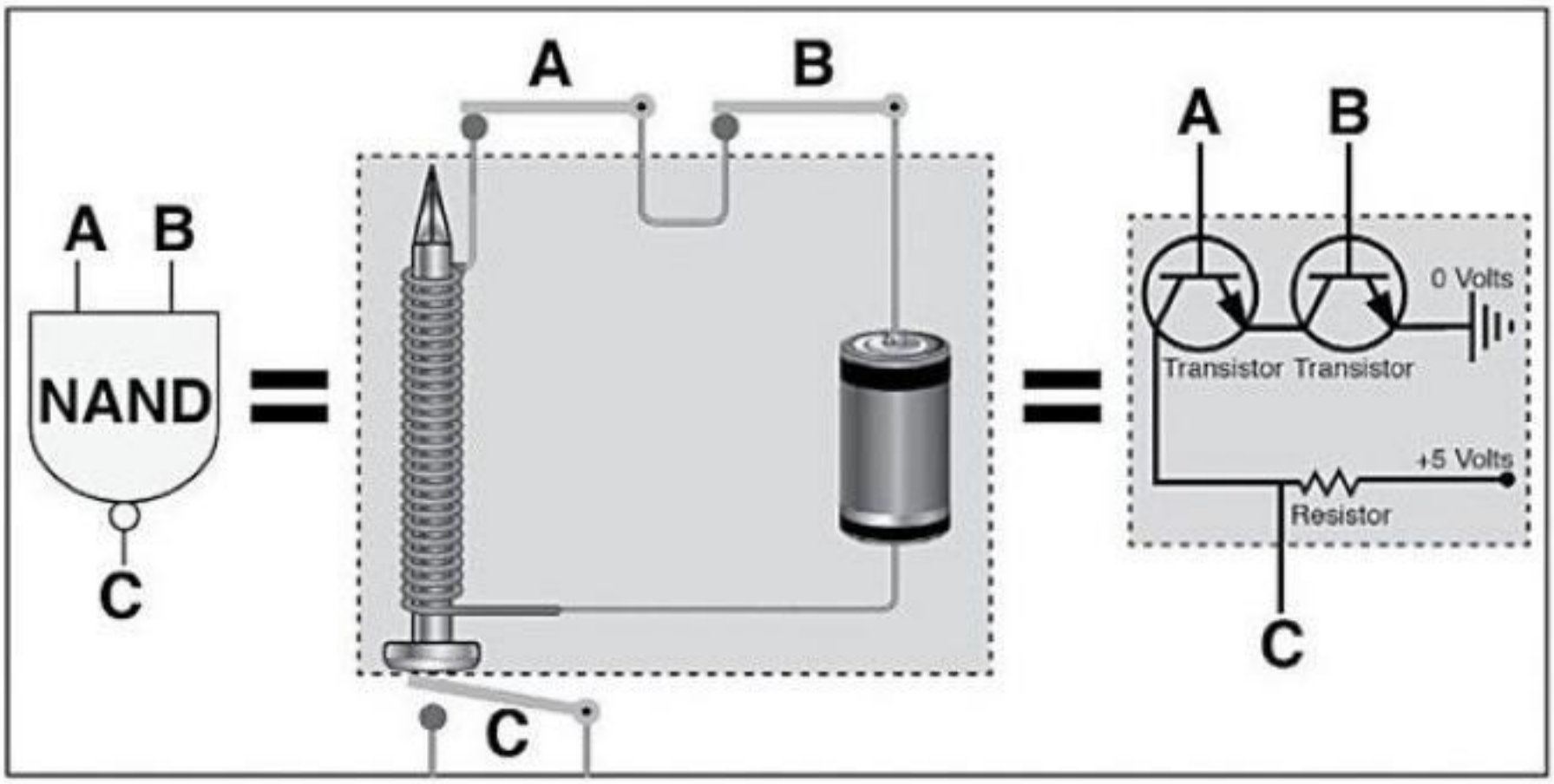


Figura 2.6: Uma porta chamada NAND recebe dois bits, A e B, como entradas e calcula um bit C como saída, de acordo com a regra de que C = 0 se A = B = 1 e, em qualquer outra combinação de A e B, C = 1. Muitos sistemas físicos podem ser usados como portas NAND. No exemplo do meio, os comutadores são interpretados como bits em que 0 = aberto, 1 = fechado e, quando os comutadores A e B são fechados, um eletroímã abre o comutador C. À direita da figura, tensões (potenciais elétricos) são interpretadas como bits em que 1 = cinco volts, 0 = zero volt e, quando os fios A e B estão ambos em cinco volts, os dois transistores conduzem eletricidade e o fio C cai para aproximadamente zero volt.

Existe um teorema conhecido na ciência da computação que diz que as portas NAND são *universais*, o que significa que você pode implementar *qualquer* função bem definida simplesmente conectando portas NAND.[13] Portanto, se você pode construir portas NAND suficientes, pode construir um dispositivo que computa qualquer coisa! Caso queira ver como isso funciona, a Figura 2.7 mostra como multiplicar números usando nada além de portas NAND.

Os pesquisadores do MIT Norman Margolus e Tommaso Toffoli cunharam o nome *computronium* para qualquer substância que consiga executar cálculos arbitrários. Acabamos de ver que a produção de *computronium* não precisa ser particularmente difícil: a substância precisa apenas ser capaz de implementar portas NAND conectadas da maneira desejada. De fato, existem muitos outros tipos de *computronium* também. Uma variante simples que também funciona envolve a substituição das portas NAND pelas portas NOR, que produzem 1 somente quando ambas as entradas são 0. Na próxima seção, vamos explorar as redes neurais, que também podem implementar cálculos arbitrários, ou seja, agir como *computronium*. O cientista e empresário Stephen Wolfram mostrou que o mesmo vale para dispositivos simples chamados autômatos celulares, que

permitiria que você comprasse todos os bens e serviços produzidos na Terra este ano. Essa queda drástica nos custos é, obviamente, uma das principais razões pelas quais a computação está em toda parte nos dias de hoje, tendo se espalhado das instalações de computação do tamanho de um prédio do passado para nossas casas, nossos carros e nossos bolsos – e aparecendo até em lugares inesperados, como nos tênis.

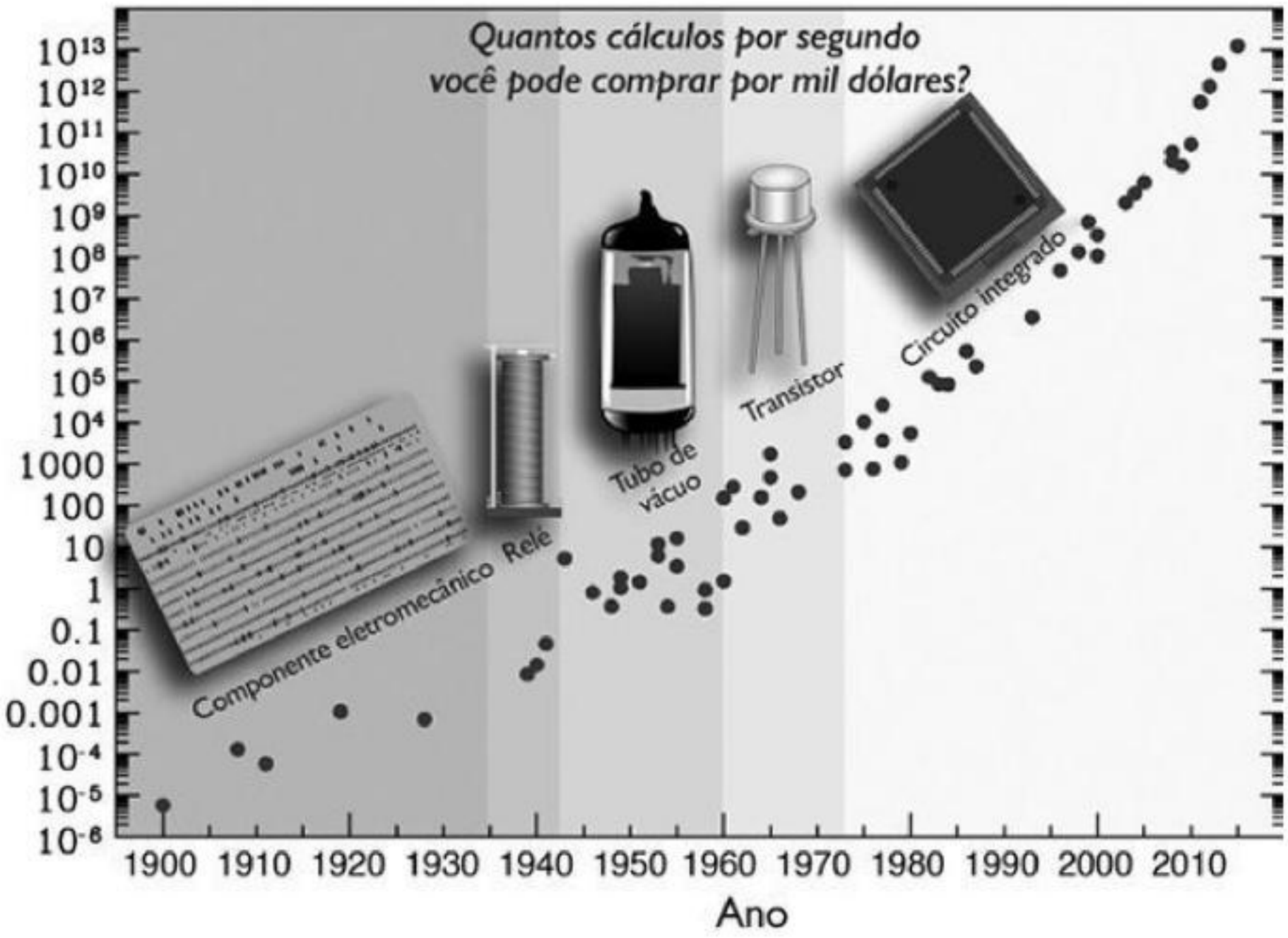


Figura 2.8: Desde 1900, a computação ficou duas vezes mais barata a cada dois anos. O gráfico mostra o poder de computação medido em operações de ponto flutuante por segundo (FLOPS) que podem ser compradas por mil dólares.[14] O cálculo específico que define uma operação de ponto flutuante corresponde a cerca de 10^5 operações lógicas elementares, como inversões de bits ou avaliações NAND.

Por que nossa tecnologia continua ganhando o dobro de poder em intervalos regulares, exibindo o que os matemáticos chamam de crescimento exponencial? De fato, por que isso está acontecendo não apenas em termos de miniaturização de transistor (uma tendência conhecida como Lei de Moore), mas também de forma mais ampla para a computação como um todo (Figura 2.8), para a memória (Figura 2.4) e para uma infinidade de outras tecnologias que variam do sequenciamento do genoma à imagem do

Anteriormente, vimos como uma superfície com muitos vales (veja a Figura 2.3) pode ser usada como um dispositivo de memória: por exemplo, se o fundo de um dos vales estiver na posição $x = \pi \approx 3{,}14159$, e não houver outros vales próximos, então é possível colocar uma bola em $x = 3$ e observar o sistema calcular as casas decimais faltantes, deixando a bola rolar para baixo. Agora, suponha que a superfície seja feita de argila macia e comece completamente plana, como uma lousa em branco. Se alguns entusiastas da matemática colocarem a bola repetidas vezes nos locais de cada um de seus números favoritos, a gravidade criará gradualmente vales nesses locais, e, depois, a superfície da argila poderá ser usada para recuperar essas memórias armazenadas. Em outras palavras, a superfície da argila *aprendeu* a calcular dígitos de números como π.

Outros sistemas físicos, como o cérebro, podem aprender com muito mais eficiência com base na mesma ideia. John Hopfield mostrou que sua rede de neurônios interconectados mencionada anteriormente pode aprender de maneira análoga: se você a colocar em certos estados de modo repetitivo, ela vai aprender de modo gradual esses estados e retornar a eles de qualquer estado próximo. Se você vê os membros de sua família muitas vezes, as memórias de como são podem ser acionadas por qualquer coisa relacionada a eles.

As redes neurais agora transformaram a inteligência biológica e artificial e recentemente começaram a dominar o subcampo da IA conhecido como aprendizado de máquina, ou *machine learning* (o estudo de algoritmos que melhoram com a experiência). Antes de nos aprofundarmos em como essas redes podem aprender, vamos primeiro entender como elas podem calcular. Uma rede neural é simplesmente um grupo de neurônios interconectados capazes de influenciar o comportamento um do outro. Seu cérebro contém quase tantos neurônios quanto o número de estrelas em nossa galáxia: na casa dos cem bilhões. Em média, cada um desses neurônios está conectado a cerca de mil outros através de junções chamadas *sinapses*, e são os pontos fortes dessas conexões de sinapse de cerca de cem trilhões que codificam a maioria das informações em seu cérebro.

Podemos desenhar esquematicamente uma rede neural como uma coleção de pontos representando neurônios conectados por linhas representando sinapses (veja a Figura 2.9). Os neurônios do mundo real são dispositivos eletroquímicos muito complicados que não se parecem em nada com esta ilustração esquemática: eles envolvem partes diferentes com nomes como axônios e dendritos; existem muitos tipos diferentes de neurônios que operam de várias maneiras, e os detalhes exatos de como e quando a atividade elétrica em um neurônio afeta os outros ainda é objeto de estudo ativo. No entanto, os pesquisadores

parte da resposta está na física. Descobrimos que a classe de funções que as leis da física jogam em nós e nos leva a ter interesse em computação também é uma classe bem pequena porque, por razões que ainda não entendemos por completo, as leis da física são notoriamente simples. Além disso, a pequena fração de funções que as redes neurais podem calcular é muito semelhante à pequena fração em que a física nos deixa interessados! Também ampliamos o trabalho anterior, mostrando que as redes neurais de aprendizado profundo (chamadas de *profundo* se elas contêm muitas camadas) são muito mais eficientes do que as superficiais para muitas dessas funções de interesse. Por exemplo, junto com outro aluno incrível do MIT, David Rolnick, mostramos que a tarefa simples de multiplicar n números requer uma quantidade imensa de 2^n neurônios para uma rede com apenas uma camada, mas precisa de apenas cerca de 4^n neurônios em uma rede profunda. Isso ajuda a explicar não apenas por que as redes neurais agora estão na moda entre os pesquisadores de IA, mas também por que desenvolvemos redes neurais em nossos cérebros: se desenvolvemos cérebros para prever o futuro, faz sentido evoluirmos uma arquitetura computacional boa naqueles problemas computacionais que importam no mundo físico.

Agora que exploramos como as redes neurais funcionam e calculam, vamos voltar à questão de como podem aprender. Especificamente, como uma rede neural pode melhorar a computação atualizando suas sinapses?

Em seu seminal livro de 1949, *The Organization of Behavior: A Neuropsychological Theory*, o psicólogo canadense Donald Hebb argumentou que, se dois neurônios próximos estivessem frequentemente ativos ("disparando") ao mesmo tempo, seu acoplamento sináptico se fortaleceria para que aprendessem a ajudar a desencadear um ao outro – uma ideia captada pelo slogan popular "Acende junto, se conecta junto". Embora os detalhes de como os cérebros reais aprendem ainda estejam longe de serem compreendidos, e a pesquisa tenha mostrado que as respostas são, em muitos casos, muito mais complicadas, também foi demonstrado que mesmo essa regra simples de aprendizado (conhecida como aprendizado hebbiano) permite que redes neurais aprendam coisas interessantes. John Hopfield mostrou que o aprendizado hebbiano permitiu que sua rede neural artificial simplificada armazenasse muitas memórias complexas simplesmente sendo expostas a elas repetidas vezes. Essa exposição à informação a ser aprendida costuma ser chamada de "treinamento" quando se refere a redes neurais artificiais (ou a animais ou pessoas aprendendo habilidades), embora "estudo", "educação" ou "experiência" possam ser igualmente adequados. As redes neurais artificiais que alimentam os sistemas atuais de

aprendizado profundo transformou muitos aspectos da visão computacional, da transcrição à mão à análise de vídeo em tempo real para carros autônomos. Da mesma forma, revolucionou a capacidade dos computadores de transformar a linguagem falada em texto e traduzi-la para outros idiomas, mesmo em tempo real – e é por isso que agora podemos conversar com assistentes digitais pessoais como Siri, Google Now e Cortana. Os quebra-cabeças irritantes do CAPTCHA, nos quais precisamos convencer um site de que somos humanos, estão se tornando cada vez mais difíceis para se manter à frente do que a tecnologia de aprendizado de máquina pode fazer. Em 2015, o Google DeepMind lançou um sistema de IA, usando aprendizado profundo, capaz de dominar dezenas de jogos de computador, como uma criança faria – sem nenhuma instrução –, exceto que logo aprendeu a jogar melhor do que qualquer humano. Em 2016, a mesma empresa construiu o AlphaGo, um sistema de computador que utilizou o aprendizado profundo para avaliar a força de diferentes posições do tabuleiro e derrotou o campeão mundial do jogo Go. Esse progresso está alimentando um círculo virtuoso, trazendo cada vez mais financiamento e talento para a pesquisa em IA – o que gera mais progresso.

Passamos este capítulo explorando a natureza da inteligência e seu desenvolvimento até agora. Quanto tempo vai demorar até que as máquinas possam nos superar em *todas* as tarefas cognitivas? Claramente, não sabemos, e precisamos estar abertos à possibilidade de que a resposta seja "nunca". No entanto, uma mensagem básica deste capítulo é que também precisamos considerar a possibilidade de que *vai* acontecer, talvez até durante a nossa vida. Afinal, a matéria pode ser organizada de modo que, quando obedece às leis da física, ela se lembre, calcule e aprenda – e não precise ser biológica. Os pesquisadores de IA têm sido acusados com frequência de prometer demais e apresentar resultados insuficientes, mas, para ser justo, alguns de seus críticos também não têm o melhor histórico. Alguns continuam mexendo nas metas, definindo efetivamente a inteligência como aquilo que os computadores ainda não podem fazer ou como aquilo que nos impressiona. Agora, as máquinas são boas ou excelentes em aritmética, jogar xadrez, provar teoremas matemáticos, escolher ações, legendar imagens, dirigir, jogar jogos eletrônicos, jogar Go, sintetizar a fala, transcrever a fala, traduzir e diagnosticar câncer, mas alguns críticos zombam com desdém "Claro, mas isso não é inteligência *real*!". Eles podem continuar alegando que a inteligência real envolve apenas os topos das montanhas da paisagem de Moravec (Figura 2.2) que ainda não foram submersos, assim como algumas pessoas no passado argumentavam que legendagem de imagens e o jogo Go deveriam contar – enquanto a água continuou subindo.

Pessoalmente, não me incomoda que as máquinas de hoje me superem em habilidades manuais, como cavar e tricotar, já que esses não são meus hobbies nem minhas fontes de renda ou de autovalorização. Aliás, quaisquer ilusões que eu tenha alimentado sobre minhas habilidades a esse respeito foram destruídas aos 8 anos de idade, quando minha escola me forçou a fazer uma aula de tricô na qual quase fui reprovado, e concluí meu projeto só porque um aluno solidário da quinta série ficou com pena de mim e me ajudou.

Contudo, à medida que a tecnologia continua melhorando, será que a ascensão da IA vai acabar encobrindo também essas habilidades que criam minha noção atual de autoestima e meu valor no mercado de trabalho? Stuart Russell me disse que ele e muitos de seus colegas pesquisadores de IA haviam vivenciado recentemente um momento de "Minha nossa!", quando testemunharam a IA fazendo algo que não esperavam ver por muitos anos. Por falar nisso, por favor, deixe-me contar sobre alguns dos meus momentos de surpresa, e como os vejo como precursores de habilidades humanas que logo serão superadas.

Descobertas

Agentes de aprendizado por reforço profundo

Fiquei extremamente boquiaberto em 2014 enquanto assistia a um vídeo do sistema de IA DeepMind aprendendo a jogar jogos de computador. Para ser mais específico, a IA estava jogando Breakout (veja a Figura 3.1), um clássico jogo de Atari da minha adolescência do qual me lembro com carinho. O objetivo é mover uma barra para bater várias vezes uma bola contra uma parede de tijolos; toda vez que você bate em um tijolo, ele desaparece, e sua pontuação aumenta.

Na época, eu já tinha criado alguns jogos de computador por conta própria e sabia que não era difícil escrever um programa que pudesse jogar Breakout – mas *não* foi isso que a equipe do DeepMind tinha feito. Na verdade, eles criaram uma IA nova que não sabia nada sobre esse jogo – nem sobre outros jogos, nem mesmo sobre os *conceitos* jogos, barras, tijolos ou bolas. Tudo o que a IA sabia era que uma longa lista de números era inserida em intervalos regulares: a pontuação atual e uma longa lista de números que nós

Intuição, criatividade e estratégia

Outro momento decisivo para mim foi quando o AlphaGo, sistema de IA do DeepMind, venceu uma disputa de cinco partidas de Go contra Lee Sedol, considerado o melhor do mundo nesse jogo no início do século XXI.

Era de se esperar que jogadores humanos de Go fossem destronados por máquinas em algum momento, já que isso acontecera com seus colegas enxadristas duas décadas antes. No entanto, a maioria dos especialistas em Go previu que isso levaria mais uma década, então o triunfo do AlphaGo foi um momento crucial para eles e para mim. Nick Bostrom e Ray Kurzweil enfatizaram como pode ser difícil o avanço da IA, o que fica evidente em entrevistas com o próprio Lee Sedol antes e depois de perder os três primeiros jogos:

- Outubro de 2015: "Com base no nível visto... acho que vencerei o jogo quase de lavada."
- Fevereiro de 2016: "Ouvi dizer que a IA do Google DeepMind é surpreendentemente forte e está ficando mais forte, mas estou confiante de que posso ganhar pelo menos desta vez."
- 9 de março de 2016: "Fiquei muito surpreso porque achava que não ia perder."
- 10 de março de 2016: "Fiquei sem palavras... estou em choque. Preciso admitir que... o terceiro jogo não vai ser fácil para mim."
- 12 de março de 2016: "Eu meio que me senti impotente."

Em menos de um ano depois de jogar com Lee Sedol, um AlphaGo melhorado jogou com todos os 20 melhores jogadores do mundo sem perder uma única partida.

Por que isso foi tão importante para mim pessoalmente? Bem, confessei anteriormente que vejo a intuição e a criatividade como dois dos meus principais traços humanos, e como vou explicar agora, sinto que o AlphaGo exibia ambos.

Os jogadores de Go se revezam colocando pedras pretas e brancas nas 19 linhas verticais e 19 horizontais do tabuleiro (veja a Figura 3.2). Existem muito mais posições possíveis no Go do que átomos em nosso Universo, o que significa que tentar analisar todas as sequências interessantes de movimentos futuros rapidamente se torna inútil. Os jogadores, portanto, dependem fortemente da intuição subconsciente para complementar seu raciocínio consciente, com especialistas desenvolvendo uma sensação quase estranha de quais posições são fortes e quais são fracas. Como vimos no capítulo

recentes progressos no aprendizado profundo da conversão de fala em texto e de texto em fala, esses usuários agora podem falar com seus smartphones em um idioma e ouvir o resultado traduzido.

O processamento de linguagem natural é agora um dos campos de IA de mais rápido avanço, e acho que o sucesso adicional terá um grande impacto, porque a linguagem é muito central para o ser humano. Quanto melhor uma IA se tornar em previsões linguísticas, melhor poderá criar respostas razoáveis por e-mail ou continuar uma conversa falada. Isso pode, pelo menos para alguém de fora, dar a impressão de que o pensamento humano está ocorrendo. Os sistemas de aprendizado profundo estão, portanto, dando pequenos passos para passar pelo famoso teste de Turing, no qual uma máquina precisa conversar suficientemente bem por escrito para induzir uma pessoa a pensar que ela também é humana.

A IA de processamento de línguas ainda tem um longo caminho a percorrer. Embora eu precise confessar que fico um pouco desanimado quando sou traduzido por uma IA, me sinto melhor quando lembro que, até agora, ela não *compreende* o que está dizendo de nenhum modo significativo. Ao ser treinada com grandes conjuntos de dados, ela descobre padrões e relações envolvendo palavras sem nunca as relacionar com nada no mundo real. Por exemplo, ela pode representar cada palavra com uma lista de mil números que especificam sua semelhança com outras palavras e, então, concluir que a diferença entre "rei" e "rainha" é semelhante à diferença entre "marido" e "esposa" – mas ainda não tem ideia do que significa ser homem ou mulher, nem mesmo que exista uma realidade física com espaço, tempo e matéria.

Como o teste de Turing trata fundamentalmente de decepção, foi criticado por testar a credulidade humana mais do que a verdadeira inteligência artificial. Por outro lado, um teste rival chamado Winograd Schema Challenge (ou Desafio do Esquema de Winograd, em tradução livre) vai direto na jugular, adotando o entendimento de senso comum que os atuais sistemas de aprendizado profundo tendem a não ter. Nós, humanos, ao analisar uma frase, com frequência usamos o conhecimento do mundo real para descobrir a que um determinado pronome se refere. Por exemplo, um desafio típico de Winograd pergunta a que "eles" se referem aqui:

1. "Os vereadores da cidade não deram permissão aos manifestantes porque eles temiam violência."
2. "Os vereadores da cidade não deram permissão aos manifestantes porque eles

verificação, validação, segurança e *controle*.[23] Para impedir que as coisas fiquem muito nerds e secas, vamos fazer isso explorando sucessos e falhas passados da tecnologia da informação em diferentes áreas, bem como lições valiosas com as quais podemos aprender e desafios de pesquisa que elas apresentam.

Embora a maioria dessas histórias seja antiga, envolvendo sistemas de computador de baixa tecnologia que quase ninguém chamaria de IA e que causaram poucas baixas, ou nenhuma, veremos que elas ainda nos ensinam lições valiosas para projetar futuros sistemas de IA seguros e poderosos, cujas falhas podem ser de fato catastróficas.

IA para exploração espacial

Vamos começar com algo de que gosto: exploração espacial. A tecnologia dos computadores nos permitiu levar as pessoas para a Lua e enviar naves espaciais não tripuladas para explorar todos os planetas do nosso Sistema Solar, chegando até a lua de Saturno, Titã, e a um cometa. Como vamos explorar no Capítulo 6, a futura IA pode nos ajudar a explorar outros sistemas solares e outras galáxias – se não houver bugs. Em 4 de junho de 1996, os cientistas que esperavam pesquisar a magnetosfera da Terra aplaudiram felizes quando um foguete Ariane 5, da Agência Espacial Europeia, rugiu através do céu com os instrumentos científicos que eles haviam construído. Trinta e sete segundos depois, seus sorrisos desapareceram quando o foguete explodiu em uma queima de fogos de artifício, custando centenas de milhões de dólares.[24] Verificou-se que a causa tinha sido um software com bugs manipulando um número muito grande para caber nos 16 bits alocados para ele.[25] Dois anos depois, o Mars Climate Orbiter da Nasa acidentalmente entrou na atmosfera do Planeta Vermelho e se desintegrou porque duas partes distintas do software usaram unidades diferentes de força, causando um erro de 445% no controle de impulso do motor do foguete.[26] Esse foi o segundo bug supercaro da Nasa: sua missão Mariner I para Vênus explodiu após o lançamento no Cabo Canaveral, em 22 de julho de 1962, depois que o software de controle de voo foi enganado por um sinal de pontuação incorreto.[27] Como que para mostrar que não apenas os ocidentais dominaram a arte de lançar bugs no espaço, a missão soviética Phobos 1 falhou em 2 de setembro de 1988. Essa foi a espaçonave interplanetária mais pesada já lançada, com o objetivo espetacular de implantar uma sonda na lua de Marte, Phobos – frustrada quando a ausência de um hífen

evitados com uma melhor validação: os robôs causaram danos não por causa de bugs ou maldade, mas porque fizeram suposições inválidas – de que a pessoa não estava presente ou que era uma peça de automóvel.

IA para transporte

Embora possa salvar muitas vidas na indústria, a IA pode conseguir economizar ainda mais no transporte. Só os acidentes de carro mataram mais de 1,2 milhão de pessoas em 2015, e os acidentes de aeronaves, trens e barcos juntos mataram milhares mais. Nos Estados Unidos, com seus altos padrões de segurança, os acidentes de automóvel mataram cerca de 35 mil pessoas no ano passado – sete vezes mais do que todos os acidentes industriais juntos.[38] Quando tivemos um painel de discussão sobre isso em Austin, Texas, na reunião anual de 2016 da Associação para o Avanço da Inteligência Artificial, o cientista de computação israelense Moshe Vardi ficou bastante emocionado e argumentou que a IA não apenas *poderia* reduzir as fatalidades nas estradas, mas *deve* fazê-lo: "É um imperativo moral!", exclamou. Como quase todos os acidentes de carro são causados por erro humano, acredita-se amplamente que carros autônomos movidos a IA podem eliminar pelo menos 90% das mortes nas estradas, e esse otimismo está incentivando um grande progresso no sentido de colocar carros autônomos nas estradas. Elon Musk prevê que futuros veículos autônomos não serão apenas mais seguros, mas também ganharão dinheiro para seus proprietários enquanto não forem necessários, competindo com a Uber e a Lyft.

Até agora, os carros autônomos têm, de fato, um registro de segurança melhor do que os motoristas humanos, e os acidentes ocorridos ressaltam a importância e a dificuldade da validação. A primeira colisão causada por um carro autônomo do Google ocorreu em 14 de fevereiro de 2016 porque o sistema fez uma suposição incorreta sobre um ônibus: que seu motorista cederia passagem quando o carro parasse na frente dele. O primeiro acidente fatal causado por um Tesla autônomo, que atingiu o baú de um caminhão que atravessava a estrada, em 7 de maio de 2016, foi causado por duas suposições incorretas:[39] que o lado branco brilhante do baú era apenas parte do céu claro e que o motorista (que supostamente estava assistindo a um filme de Harry Potter) estava prestando atenção e interviria se algo desse errado.[40]