MÁQUINAS ÉTICAS

SEU GUIA CONCISO PARA UMA **IA** TOTALMENTE IMPARCIAL, TRANSPARENTE E RESPEITOSA

REID BLACKMAN

Fundador e CEO da Virtue, consultoria digital de risco ético

ALTA BOOKS
GRUPO EDITORIAL
Rio de Janeiro, 2024

SUMÁRIO

1

Como Se Deve Pensar sobre Ética

Este livro dá conselhos sobre como construir, adquirir e implementar a IA de maneira eticamente (e, portanto, reputacional, regulamentar e legalmente) segura, e fazê-lo em escala. Não estamos aqui para abordar questões existenciais e metafísicas do tipo: "Como a IA afeta o que devemos pensar sobre o que é ser humano?" ou "O que a IA nos ensina sobre a natureza da consciência?". Dito isso, não podemos obter uma direção firme sem fundações conceituais firmes. Este capítulo estabelece essas bases.

O executivo sênior que descreveu a ética de IA como "gosmenta" não era um idiota; era uma pessoa com uma carreira longa e bem-sucedida em risco e em conformidade. E os outros que a descrevem como "confusa" e "subjetiva" também são pessoas inteligentes e realizadas.

Como ex-professor de filosofia, ouvi essas opiniões por quase vinte anos. Agora as vejo outra vez com meus clientes, e sempre que as pessoas falam sobre ética de IA em quase qualquer contexto. Quando as ouço, logo aponto que tal mentalidade impede o progresso.

Quando dizem que a ética é gosmenta — ou, como direi a partir de agora, "subjetiva" —, as pessoas estão de fato dizendo que não sabem ao certo o que pensar sobre ela, e geralmente desistem de fazê-lo.

É bastante difícil para os líderes seniores tentarem efetuar mudanças dentro da organização. Esses líderes estão tentando criar um programa abrangente de risco ético de IA, que exige adesão em todos os níveis da organização. Eles muitas vezes têm a experiência de entrar em uma sala de engenheiros, dizendo-lhes que a ética de IA é muito importante e, em seguida, enfrentar a pergunta inevitável: "Mas a ética não é subjetiva?"

É o beijo da morte. Engenheiros tendem a gostar de coisas que são concretas, quantificáveis, empiricamente verificáveis. Se não for, não merece atenção ou cuidado intelectual. O líder sênior falando sobre ética em IA? Isso é só RP. É o politicamente correto atrapalhando o progresso tecnológico. É um assunto delicado que não tem espaço em uma conversa entre pessoas sérias e, certamente, não tem nos negócios.

O líder que não sabe como responder a tais reclamações está em apuros. Se disser: "Bem, sim, é subjetiva, mas…", ele já perdeu. Portanto, os líderes seniores precisam acertar isso se quiserem obter a adesão organizacional de que precisarão para impulsionar — desculpe, aninhar — a ética de IA em todas as operações.

Mas isso é só o começo. Também é preciso que as pessoas pensem sobre ética de maneira que não seja recebida com indiferença sempre que precisarem fazer alguma reflexão ética, conforme exigido pelo programa de ética de IA que você implementará; na realização do processo de due diligence de risco ético durante o desenvolvimento do produto, por exemplo, ou nas deliberações de um comitê de ética.

Portanto, fazer com que os funcionários pensem a ética da IA como algo diferente de subjetivo é imperativo, tanto para a adesão organizacional quanto para a análise eficaz de riscos éticos durante o desenvolvimento, aquisição e implementação do produto. Por acaso,

há uma razão muito boa para você e sua equipe pararem de pensar a ética como subjetiva e começarem a pensá-la de maneira que se preste a discussões frutíferas e, por fim, identificação e mitigação de riscos. Falando de modo um pouco distinto, se você está propenso a falar sobre "IA responsável", precisará pensar sobre ética de maneira que se preste a uma investigação responsável sobre os riscos éticos de IA. E, para aqueles que estão no "fundão", falando de outro modo mais uma vez: a ética da IA é sobre duas coisas — IA e ética. No capítulo anterior, explicamos o que é IA e AM e como funcionam. Agora é hora de explicar a ética.

Não se preocupe: não vou escrever um tratado filosófico aqui. Tudo o que você precisa para pensar sobre ética de maneira eficaz é saber sobre *uma pergunta*, *uma confusão* e *três argumentos notoriamente ruins, mas onipresentes, que fazem você pensar que a ética é subjetiva*. Quando fizermos isso, vamos colocar a ética em prática.

A Pergunta

Uma pergunta que me fazem muito é "O que é ética?". Quem indaga isso normalmente está procurando uma "definição" de ética. Eles podem até perguntar: "Como você define 'ética'?" ou "Qual é a sua definição de 'ética'?".

Mas minha visão de como entender *o que é* ética — e isso é realmente o que a pessoa quer saber — é pensar em algumas das perguntas centrais que caracterizamos naturalmente como questões *éticas*:

- O que é uma boa vida?
- Temos alguma obrigação um com o outro? Quais são elas?
- A compaixão é uma virtude? E a coragem? E a generosidade?
- O aborto é eticamente permitido? Pena de morte? Eutanásia?

- O que é privacidade? As pessoas têm direito a ela?
- O que é discriminação? O que a torna ruim?
- As pessoas têm o mesmo valor moral?
- Os indivíduos têm obrigação de se autoaperfeiçoarem?
- É eticamente admissível mentir?
- As empresas têm obrigações com seus funcionários? E com a sociedade em geral?
- O Facebook está incentivando ou manipulando injustificadamente os usuários a clicar em anúncios?
- É eticamente permissível usar algoritmos caixa preta para diagnosticar doenças?

E assim por diante. O que é ética? Bem, não se preocupe com a definição do termo — se você quer uma, basta procurá-la no dicionário. Se quiser saber do que se trata a ética, pense sobre esses tipos de perguntas e aquelas que as permeiam. Se entende isso, não há razão para se preocupar com definições.

A Confusão

Uma fonte significativa de confusão para muitas pessoas que pensam a ética como subjetiva é não distinguir entre as *crenças* das pessoas sobre a ética — o que acreditam ser eticamente certo ou errado, bom ou ruim e assim por diante — e a própria ética. E ao juntar essas duas coisas, fazem afirmações equivocadas sobre a subjetividade da ética quando estão, na verdade, fazendo afirmações sobre a variação das crenças das pessoas. Para vermos isso, daremos um passo para trás.

Por um lado, há nossa crença sobre a Terra ser plana ou esférica e, por outro, há a forma real da Terra. Por um lado, há nossa crença sobre

a composição química da água ser H_2O ou H_3O e, por outro, há a composição química real da água. Por um lado, há nossa crença sobre se a eleição norte-americana de 2020 foi roubada ou legítima e, por outro, há a legitimidade real da eleição.

Em geral, distinguimos nossas crenças sobre X e como X é realmente, e às vezes nossas crenças são verdadeiras, às vezes são falsas. Se não fizéssemos essa distinção entre nossas crenças sobre X, por um lado, e como X é realmente, por outro, teríamos que pensar que acreditar em X o faz ser de um jeito, mas ninguém pensa que acreditar que a Terra é esférica ou plana, que a água é H_3O ou H_2O, ou que a eleição foi roubada ou legítima são coisas que tornam a Terra esférica, ou a água composta de H_2O, ou a eleição legítima.

É claro que as crenças das pessoas sobre isso podem mudar ou evoluir ao longo do tempo. Em um ponto, a maioria das pessoas acreditava que a Terra é plana, não acreditava que a água é H_2O (em sua defesa, não sabiam nada sobre química), e algumas pessoas deixaram de acreditar que a eleição foi roubada e agora acreditam que foi legítima. Então, nossas crenças mudam, mas as coisas sobre as quais se tinha (ou não) crenças eram o que eram o tempo todo. Não é como se a Terra mudasse de plana para esférica.

Vamos continuar com essa distinção: por um lado, há nossa crença sobre a permissibilidade ética ou a inadmissibilidade da escravidão e, por outro, se a escravidão é eticamente permissível. Se há alguma coisa eticamente *inadmissível*, é a escravidão.

Em um ponto, a maioria das pessoas — particularmente aqueles que se beneficiaram da escravidão — acreditava que a escravidão era eticamente permitida. Mas as crenças das pessoas mudaram ou evoluíram ao longo do tempo, e agora todos acreditam que a escravidão está errada. O erro da escravidão não mudou; sempre foi errado. (Nota rápida: há uma questão à parte sobre até que ponto aqueles que pensavam

que era eticamente permissível são merecedores de *culpa*, uma vez que todos ao seu redor também pensavam que era permissível, mas não discutiremos isso aqui.)

De certa forma, tudo isso é bastante óbvio. *É claro* que há diferença entre o que as pessoas acreditam sobre X e como X realmente é. Mas as coisas tendem a ficar muito estranhas quando as pessoas falam sobre ética; a distinção vai direto por água abaixo. As pessoas dirão coisas como: "Sua ética é diferente da minha" ou "A ética é subjetiva porque a ética ou a moralidade varia entre culturas e indivíduos", ou "A ética evoluiu ao longo do tempo; as pessoas já pensaram que a escravidão era eticamente permitida e agora acham que não é".

Mas agora somos capazes de ver que "sua ética é diferente da minha ética" pode significar "o que é eticamente certo para você é eticamente errado para mim" ou "o que você acredita ser eticamente certo é algo que acredito ser eticamente errado". E já vimos que, embora as crenças éticas claras mudem ou evoluam ao longo do tempo, não significa que o que é certo ou errado mude ao longo do tempo. O que é estranho é que, quando as pessoas dizem essas coisas, muitas vezes pensam em crenças éticas como *a mesma coisa* que o certo e errado ético, e isso é simplesmente uma confusão.

Explicando melhor, a questão sobre a ética ser subjetiva ou não está relacionada ao fato de as crenças éticas das pessoas variarem ao longo do tempo e entre indivíduos e culturas. Claro que é assim! A questão sobre a ética ser subjetiva se relaciona ao fato de que algo ser certo ou errado, ou bom ou ruim, varia ao longo do tempo, entre indivíduos e culturas. Agora que entendemos isso, podemos examinar as razões comuns para pensar que a ética é subjetiva.

Três Argumentos Muito Ruins que Fazem Você Pensar que a Ética É Subjetiva

Dizer que a ética é subjetiva é dizer que não há fatos sobre o que é eticamente certo, errado, bom, ruim, permissível, inadmissível e assim por diante. Se a ética é subjetiva, então não apenas *as crenças* éticas variam de acordo com o indivíduo e a cultura, mas a *ética em si* varia de acordo com o indivíduo e a cultura. Se a ética é subjetiva, então não há tal coisa como investigação ética *responsável*, porque ninguém pode estar incorreto em suas conclusões (o mesmo vale para a ética responsável da IA, ou "IA responsável"). Se a ética é subjetiva, então é sensível, gosmenta, difusa e não é um assunto para pessoas sérias e certamente não para pessoas sérias em um contexto de negócios.

Agora sabemos do que se trata a ética. E sabemos distinguir entre crenças éticas sobre o que é certo ou errado e *o que é* certo ou errado. No entanto, mesmo as pessoas que conhecem essas coisas ainda podem pensar que a ética é subjetiva. E nesses quase vinte anos de ensino de filosofia, notei três argumentos principais para a crença de que a ética é subjetiva, cada um dos quais é totalmente equivocado. Vou expô-los e depois explicar o que há de errado com eles. E para deixar mais explicado: não é apenas a minha opinião de que são argumentos ruins. Os filósofos não concordam muito, mas há um consenso de que, mesmo que a ética seja subjetiva, não é por nenhum destes argumentos.

Argumento Muito Ruim nº 1: A ética é subjetiva porque as pessoas discordam sobre o que é certo e errado. As pessoas se envolvem em controvérsias éticas; elas discordam sobre a permissibilidade moral do aborto e da pena capital, sobre mentir para a polícia para proteger o amigo e se a coleta de dados das pessoas sem o conhecimento delas em troca do uso livre de serviços é eticamente permissível. E como há

tanta discordância — tantas crenças morais e éticas diferentes —, a ética é subjetiva; *não há verdade* sobre esse tema.

Argumento Muito Ruim nº 2: A ciência nos entrega a verdade. Ética não é ciência, então não nos dá a verdade. A ciência, e mais especificamente, o método científico, é *a única* maneira de descobrirmos verdades sobre o mundo. Observações empíricas ("ver é crer") e investigações (experimentos científicos, por exemplo) fornecem fatos sobre o mundo. Todo o resto é interpretação, ou seja, subjetivo. Mais uma vez, a ética é subjetiva porque as observações empíricas têm o monopólio da verdade; ética e investigação ética, por não serem empíricas, dizem respeito ao reino da não verdade. Em suma: *apenas afirmações cientificamente verificáveis são verdadeiras*.

Argumento Muito Ruim nº 3: A ética requer uma figura de autoridade para dizer o que é certo e errado; caso contrário, é subjetiva. Você tem suas crenças, eu tenho as minhas e outra pessoa tem as dela. E não é como se tivéssemos evidências científicas de que uma visão está certa e outra está errada, então quem deve dizer o que é certo e o que é errado? É tudo subjetivo. Em suma: *se existem verdades éticas, então deve haver uma figura de autoridade que torna isso certo e aquilo errado.*

O que Há de Tão Ruim nos Argumentos Ruins

Por que o Argumento Muito Ruim nº 1 é muito ruim? O primeiro argumento para pensar que a ética é subjetiva é a discordância das pessoas sobre o que é certo ou errado, e se elas discordam sobre essas coisas, então não há verdade no assunto. Este é um bom argumento? Acertou: é muito ruim! E é possível perceber como é ruim quando se leva em conta o seguinte princípio:

Se as pessoas discordam sobre X, então
não há verdade na questão sobre X.

Veja, esse princípio é obviamente falso. As pessoas discordam de todos os tipos de coisas sobre as quais há uma verdade no assunto. As pessoas discordam sobre se os humanos são produto da evolução, se os carros autônomos substituirão os dirigidos por humanos em uma década, se há algo no centro de um buraco negro e até se a Terra é plana ou esférica. Mas ninguém pensa: "Bem, acho que não há verdade na questão sobre a forma da Terra!"

O fato de as pessoas discordarem sobre X não comprova que não há verdade sobre X.

E assim é, também, com a ética. O fato de as pessoas discordarem sobre mentir para proteger seu amigo da polícia, se as pessoas devem ser donas dos próprios dados, se o Facebook se envolve em manipulação eticamente inaceitável de seus usuários e assim por diante, não comprova que não há verdade sobre essas questões.

"Mas", as pessoas rebatem, "com a ética é diferente. É uma exceção ao princípio".

Mas por que deveríamos pensar que com a ética é diferente? Por que pensar que está isenta da lição que acabamos de aprender sobre discordância e verdade?

A resposta é a mesma 99% das vezes: "Porque os outros casos de discordância podem ser cientificamente comprovados. Não há como a ciência resolver a ética."

Essa é uma boa resposta, de certa forma. Mas é um total *abandono* do primeiro argumento e uma retirada para o Argumento Muito Ruim nº 2 para pensar que a ética é subjetiva. A resposta apenas diz: "Somente afirmações cientificamente verificáveis são verdadeiras." Então vamos investigar isso.

Por que o Argumento Muito Ruim nº 2 é muito ruim? Este é surpreendentemente fácil de refutar. Diz que apenas afirmações cientificamente verificáveis são verdadeiras. Na verdade, vamos destacar isso em letras garrafais:

Afirmação: Somente afirmações cientificamente verificáveis são verdadeiras.

Se você é particularmente astuto, acabou de se fazer uma pergunta: "Se esta é uma afirmação, e a afirmação diz que apenas afirmações cientificamente verificáveis são verdadeiras, o que dizer dessa afirmação?"

A pergunta revela o problema com esta posição: ela é autodestrutiva. Afinal, como você verificaria cientificamente essa afirmação? Apresente-a ao químico, ou ao biólogo, ou ao físico, ou ao geólogo, e diga: "Por favor, realize um experimento para verificar essa afirmação." O que eles poderiam fazer? Escrevê-la em um pedaço de papel e medir quanto peso o papel ganhou? Anexá-la a um leitor sísmico? Colocar algumas células nela? Não há nada que possam fazer, e isso é porque a afirmação não é cientificamente verificável. Então, qualquer um que acredite nela, se quiser ser coerente, terá que parar de acreditar. E para aqueles que desde o início nunca acreditaram, tudo bem. Então, faça o que fizer, não acredite na afirmação. É falsa.

Por que o Argumento Muito Ruim nº 3 é muito ruim? OK, estamos quase acabando. Você pode pensar que, para que a ética não seja subjetiva, tem que haver fatos éticos, e para que haja fatos éticos, terá que haver uma figura de autoridade para dizer o que é certo e errado.

Mas isso é ignorar algumas maneiras básicas pelas quais pensamos sobre os fatos. Ninguém diz: se há fatos sobre a forma da Terra, ou sobre a história evolutiva dos seres humanos, ou sobre a composição química da água, então deve haver uma figura de autoridade que faz dessas coisas fatos. Em vez disso, há fatos sobre essas coisas e há

pessoas (cientistas, é claro) que nos dão a *evidência* para as afirmações de que a Terra é esférica, os seres humanos são produto da evolução biológica e a água é composta de H_2O. É a evidência, os argumentos que eles nos oferecem para essas conclusões, que ganha o dia, e certamente não são os *descobridores* dessa evidência que tornam a Terra esférica ou a água composta de H_2O.

Se há fatos morais ou éticos, então devemos esperar que sejam da mesma maneira que os demais fatos. Não há necessidade de uma figura de autoridade para torná-los verdadeiros. Ao contrário, há pessoas (filósofos e teólogos, por exemplo) que fornecem provas ou argumentos para as afirmações éticas que fazem. Pense nos muitos argumentos, contra-argumentos e contracontra-argumentos nas discussões sobre a permissibilidade moral do aborto. Nenhuma dessas pessoas diz: "Eu acho que é errado, então é." Se o fizessem, não lhes daríamos atenção. Quando fazemos o nosso melhor, prestamos atenção aos seus argumentos e investigamos se são sólidos, assim como fazemos com argumentos científicos.

Por que Isso Importa?

É importante nos livrarmos dessas confusões e argumentos ruins. Na ética, inclusive em áreas relacionadas à inteligência artificial, enfrentamos problemas éticos muito reais. Esses problemas, se não forem devidamente resolvidos tanto ética quanto tecnicamente, podem levar a consequências desastrosas. Mas se abrirmos mão da ética como sendo algo objetivo, como algo sobre o qual podemos raciocinar e apresentar argumentos e mudar razoavelmente nossas mentes, então abrimos mão de que a investigação ética seja uma ferramenta à nossa disposição para resolver esses problemas reais.

Deixe-me ser um pouco mais específico. Eu testemunhei centenas, se não milhares, de discussões sobre questões de importância ética. E

todas seguem o mesmo caminho, desde que as pessoas se sintam confortáveis para expor suas opiniões. Temos algumas pessoas argumentando por um lado, outras argumentando por outro, e aqueles que não têm certeza de que posição tomar. Isso é difícil. E, então, alguém diz: "Assim, o que importa? Tudo isso é subjetivo, afinal." E então todos se olham, param e dão de ombros. Ponto-final. Todas. As. Vezes.

Até eu perguntar: "Por que você acha que a ética é subjetiva?" E aí sempre ouço os Argumentos Muito Ruins. Quando os desarticulamos, as pessoas retomam o assunto, desta vez invulneráveis a um comentário que as faria perder o rumo totalmente.

Você não precisa começar seu programa de ética de IA falando sobre a natureza da ética. Mas se não envolver isso em algum momento, eu lhe prometo — eu lhe *prometo*: as pessoas vão trazer à tona os Argumentos Muito Ruins. E então você terá um monte de gente que se ressente do politicamente correto ou das coisinhas delicadas que estão atrapalhando a grandeza tecnológica. Também terá conformidade reduzida com seu programa de ética de IA e maior risco. O que as pessoas pensam sobre a ética de IA terá impacto sobre a efetividade do seu programa de ética de IA.

Qual É... Ética? Objetiva?

Eu *não estou* tentando convencê-lo de que a ética não é subjetiva. Não estou tentando convencê-lo de que existem fatos éticos. Estou tentando convencê-lo de que os argumentos-padrão para pensar que a ética é subjetiva são muito ruins, e não enxergar isso pode levar a muitos problemas. Mas o fato de que existem três argumentos muito ruins para pensar que a ética é subjetiva não significa que não possa haver um quarto argumento, muito bom, para pensar que a ética é subjetiva.

Neste livro, não entraremos nesse mérito. O ponto de neutralizar os argumentos muito ruins é que continuamente interrompem discussões frutíferas e são um empecilho para a aceitação organizacional

genuína, de cima para baixo. Noventa e nove por cento das pessoas que veem por que são argumentos muito ruins estão prontas para aceitar que a ética não é subjetiva e, portanto, o objetivo prático de discutir isso é alcançado.

Para aqueles de vocês que não estão convencidos, no entanto, deve-se pelo menos achar que pensar que existem fatos éticos não é uma posição irracional. Certamente não é *loucura*. E assim, com *o propósito de criar e participar de um programa de risco ético de IA*, convido você a se juntar a essas pessoas na prática. Isso lhe permitirá pensar sobre como seria um sistema eficaz para identificar e mitigar esses riscos. Possibilitará que tenha conversas baseadas na razão ou em evidências com colegas sobre qual é a coisa certa a fazer. E isso garantirá que você não encerre conversas importantes que levam à proteção de sua organização contra riscos éticos, reputacionais, regulatórios e legais.

Por que Não Falar Apenas sobre a Percepção do Consumidor em Vez de Ética?

Você pode estar se perguntando por que precisamos falar sobre ética. Por que não falar apenas sobre crenças éticas do consumidor ou sobre a percepção do consumidor de forma mais geral? Então seria possível fazer a pesquisa de mercado padrão e simplesmente torná-los seus padrões éticos internamente. A ética de IA, na realidade, são apenas valores de marca incorporados ao desenvolvimento e à implementação de produtos de IA, então vamos deixar todo esse papo de ética de lado. Na verdade, vamos abandonar esse rótulo de "Ética de IA" e chamá-lo do que é: "Percepções do Consumidor de IA", OK?

Esta é uma pergunta totalmente razoável. Não acho que se baseia em confusão, mal-entendido, ingenuidade ou falha de caráter ético. Não é uma coisa *louca* de se fazer. Dito isso, é imprudente. Aqui estão três razões pelas quais eu aconselho contra isso.