LIVRO 5

MERA BONDADE

RACIONALIDADE

De A a Z

ELIEZER YUDKOWSKY

RACIONALIDADE DE A a Z

MERA BONDADE

LIVRO 5

por **ELIEZER YUDKOWSKY**

Tradução de Mariana Hungria

Brasil, 2024

Sumário

Os fins: uma introdução

por Rob Bensinger

A teoria do valor é o estudo do que realmente importa para as pessoas. É uma investigação dos nossos objetivos, gostos, prazeres, dores, medos e ambições.

Isso abrange até mesmo a moralidade convencional. A teoria do valor abarca elementos que gostaríamos de valorizar ou valorizaríamos se fôssemos mais sábios e melhores como indivíduos, não se limitando apenas ao que já valorizamos.

A teoria dos valores também engloba valores do cotidiano: arte, comida, sexo, amizade e tudo o mais que confere à vida sua carga afetiva. Por exemplo, sair para assistir a um filme com seu amigo Sam pode ser algo que você valoriza, mesmo que não se trate de um valor moral.

Refletir e debater nossos valores é útil porque, frequentemente, a maneira como agimos não condiz com nossos ideais. Nossas preferências podem entrar em conflito entre si. Podemos desejar ter um conjunto diferente de desejos. Às vezes, nos falta a vontade, a atenção ou a visão necessária para agir conforme nossos desejos.

Os seres humanos se preocupam com as consequências de suas ações, embora não consistentemente o bastante para serem considerados agentes com funções de utilidade. Quando dizemos que "os seres humanos não são instrumentalmente racionais," estamos destacando que não agem da forma que desejariam.

Teoria e prática

Para complicar ainda mais, existe uma lacuna entre como gostaríamos de agir e como realmente agimos.

Filósofos, assim como psicólogos e políticos, discordam vigorosamente sobre o que desejamos e o que deveríamos desejar. Discordam inclusive sobre o significado do "dever" de desejar algo. A história da teoria moral e dos esforços humanos de coordenação está repleta de princípios orientadores fracassados em direção à verdadeira normatividade final — desta vez, eu quero falar.

Se você busca uma especificação confiável e *prática* de seus objetivos — não para vencer debates filosóficos, mas para projetar inteligência artificial adaptativa e segura, construir instituições e organizações funcionais ou decidir qual instituição de caridade apoiar —, o histórico da humanidade com a teoria do valor não é encorajador.

"Mera bondade" compila três séries de postagens de blog sobre o valor humano: "Preferências falsas" (sobre tentativas fracassadas de teorias de valor), "Teoria do valor" (sobre obstáculos ao desenvolvimento de uma nova teoria e algumas características intuitivamente desejáveis para tal teoria), e "Humanismo quantificado" (sobre a complexa questão de como aplicar essas teorias às nossas intuições morais comuns e à tomada de decisões).

O último desses tópicos é o mais crucial. O valor de uma teoria normativa é medido pela facilidade com que ela se traduz em prática normativa. Adquirir um entendimento mais profundo e abrangente de seus valores deve melhorar sua capacidade de segui-los. No mínimo, sua teoria não deve atrapalhar sua prática. Afinal, de que adiantaria saber o que é bom?

Conciliar essa arte da ética aplicada (bem como da estética aplicada, economia aplicada e psicologia

aplicada) com nossos melhores dados e teorias disponíveis frequentemente se resume quando devemos confiar em nossos julgamentos iniciais e quando devemos abandoná-los.

Em muitos casos, nossos modelos explícitos do que nos interessa são tão frágeis ou impraticáveis que é melhor confiar em nossos impulsos iniciais. Em outros casos, uma abordagem mais informada e sistemática é preferível. Não há uma resposta abrangente. Teremos que examinar exemplos e estar atentos a diferentes indicadores de que "teorias sofisticadas tendem a falhar aqui" e "sentimentos ingênuos tendem a falhar aqui".

Jornada e destino

Um tema recorrente nas próximas páginas será a seguinte pergunta: para onde estamos indo? Quais resultados são verdadeiramente valiosos?

Para responder a essa questão, Yudkowsky cunhou o termo "teoria da diversão". A teoria da diversão é uma tentativa de descobrir como seria a nossa visão ideal do futuro — não apenas o sistema de governo ou código moral sob o qual gostaríamos idealmente de viver, mas também os tipos de aventuras que gostaríamos de experimentar, os tipos de música que gostaríamos de criar e tudo o mais que desejamos da vida.

Quando se trata do futuro, as questões da teoria da diversão se entrelaçam com as questões do transumanismo, a visão de que podemos aprimorar radicalmente a condição humana por meio do progresso científico e social [1]. O transumanismo gera debates na filosofia moral, como se os melhores resultados a longo prazo para a vida consciente se baseiam no hedonismo (busca pelo prazer) ou em noções mais complexas de eudaimonia (bem-estar geral). Outras ideias futuristas discutidas em vários pontos de "Racionalidade de A a Z" incluem criogenia (conservar o corpo em estado congelado após a morte, na esperança de futura reanimação), transferência de mente (implantar mentes humanas em hardware sintético) e colonização espacial em grande escala.

Surpreendentemente, a teoria da diversão é uma das aplicações mais subestimadas da teoria do valor. O planejamento de utopias caiu em desuso, em parte devido ao ceticismo, e também porque temos dificuldade em transformar utopias em realidade. Até a palavra "utopia" reflete esse cinismo, derivando do grego para "lugar que não existe".

No entanto, se abandonarmos a busca por uma utopia verdadeira e viável (ou "eutopia," um "lugar bom"), não é óbvio que a busca contínua de objetivos de curto prazo nos levará a um futuro que consideramos valioso a longo prazo. O valor não é uma característica inevitável do mundo. Criá-lo demanda esforço. Preservá-lo demanda esforço.

Isso nos leva a uma segunda questão: como alcançaremos esse futuro? Qual é a relação entre bons fins e bons meios?

Quando vivemos, queremos desfrutar do processo. Geralmente, não queremos apenas pular para a vitória. Às vezes, a jornada é mais importante do que o destino.

Porém, em outros casos, o oposto é verdadeiro. Às vezes, o resultado é tão crucial que a jornada deve ser deixada de lado em nossas decisões. Se você está tentando salvar a vida de um familiar, não é necessariamente ruim aproveitar o processo, mas se você puder aumentar significativamente suas chances de sucesso escolhendo uma estratégia menos agradável…

Em muitos casos, nossos valores se concentram nos resultados de nossas ações e em nosso futuro. Nos importamos com como o mundo acabará sendo, especialmente aquelas partes do mundo que amamos, que ferimos e desejamos.

Como as teorias abstratas e imparciais se comparam aos sentimentos vividos e carregados de afeto nesses casos? Em termos mais amplos, qual é a relação moral entre ações e consequências?

Essas são questões complexas, mas talvez possamos avançar na definição do que queremos dizer com elas. O que estamos incorporando em nosso conceito do que é "valioso" no início de nossa investigação?

Referências

[1] Um exemplo de argumento transumanista é o seguinte: “Poderíamos, de maneira viável, abolir o envelhecimento e as doenças dentro de algumas décadas ou séculos. Isso efetivamente eliminaria a morte por causas naturais, colocando-nos na mesma posição que organismos com senescência insignificante, como lagostas e tartarugas-gigantes de Aldabra, entre outros. Portanto, deveríamos investir na prevenção de doenças e em tecnologias antienvelhecimento.” Essa ideia se qualifica como transhumanista porque a eliminação das principais causas de ferimentos e morte teria um impacto drástico na vida humana. Bostrom e Savulescu examinam argumentos a favor e contra o aprimoramento humano radical, incluindo a objeção de Sandel de que fazer alterações significativas em nossa biologia faria com que a vida parecesse menos como um “presente” [2,3]. A *History of Transhumanist Thought* (História do Pensamento Transumanista) de Bostrom fornece contexto para esse debate.[4]

[2] Nick Bostrom, “A History of Transhumanist Thought,” Journal of Evolution and Technology 14, no. 1 (2005): 1–25, http://www.nickbostrom.com/papers/history.pdf.

[3] Michael Sandel, “What’s Wrong With Enhancement,” Background material for the President’s Council on Bioethics. (2002).

[4] Nick Bostrom and Julian Savulescu, “Human Enhancement Ethics: The State of the Debate,” in Human Enhancement, ed. Nick Bostrom and Julian Savulescu (2009)

Parte U — Preferências falsas

257 — Não (apenas) em prol da felicidade

Quando tive a oportunidade de conhecer o futurista Greg Stock, alguns anos atrás, ele argumentou que a alegria da descoberta científica logo seria substituída por pílulas capazes de simular essa alegria. Após sua palestra, abordei-o e declarei: "Compreendo que tais pílulas provavelmente serão viáveis, mas não as tomaria voluntariamente."

Stock respondeu: "No entanto, elas serão tão superiores ao sentimento genuíno que este não conseguirá competir. Tomar os comprimidos será muito mais agradável do que se dedicar ao trabalho científico propriamente dito."

Repliquei: "Concordo que essa possibilidade existe, mas farei questão de não me submeter a elas." A reação surpresa de Stock com minha postura me surpreendeu por sua vez. Com frequência, observamos especialistas em ética argumentando como se todos os desejos humanos fossem, em princípio, reduzíveis ao desejo de nossa própria felicidade, assim como Sam Harris faz em *The end of faith* (O Fim da Fé), que acabei de ler, embora a redução de Harris seja mais uma retórica do que um tópico de discussão significativo. [1]

No entanto, isso não é o mesmo que argumentar se todas as formas de felicidade podem ser avaliadas em uma escala de utilidade comum; diferentes formas de felicidade podem ocupar escalas distintas ou ser, de outra maneira, não comparáveis. Além disso, não é o mesmo que argumentar que seja teoricamente impossível valorizar algo além de nossos próprios estados psicológicos, pois ainda nos é permitido nos importar com a felicidade alheia.

A questão crucial é se devemos nos preocupar com coisas que nos tornam felizes, independentemente da felicidade que elas proporcionam.

Podemos facilmente listar diversos exemplos de moralistas que erraram ao focar em outras questões além da felicidade. Leis que proíbem o sexo oral em vários estados e países são um bom exemplo; tais legisladores teriam sido mais sensatos se dissessem: "O que quer que lhe traga satisfação." No entanto, isso não demonstra que todos os valores são redutíveis à felicidade; apenas argumenta que, nesse caso específico, foi um erro ético concentrar-se em qualquer outra coisa.

É inegável que tendemos a realizar ações que nos proporcionam felicidade, mas isso não implica que devemos considerar a felicidade como a única razão para tais ações. Em primeiro lugar, isso dificultaria explicar como poderíamos nos importar com a felicidade de outras pessoas — como poderíamos tratar as pessoas como fins em si mesmas, em vez de meros meios para alcançarmos nossa própria satisfação.

Além disso, o fato de algo ser uma consequência de nossa ação não significa que tenha sido a única justificativa. Por exemplo, se estou escrevendo uma postagem de blog e estou com dor de cabeça, posso tomar um ibuprofeno. Uma das consequências de minha ação é que a dor de cabeça desaparece, mas isso não significa que tenha sido a única consequência, ou mesmo a razão mais importante de minha decisão. Valorizo o estado de não ter dor de cabeça, mas posso valorizar algo tanto por si só quanto como um meio para alcançar um fim.

Para todo valor ser redutível à felicidade, não basta demonstrar que a felicidade está envolvida na maioria de nossas decisões, nem mesmo que seja a consequência mais importante nelas todas; ela deve ser a única consequência. Esse é um padrão difícil de cumprir. (Originalmente, encontrei esse ponto em um artigo de Sober e Wilson, embora não possa afirmar com certeza qual.)

Se eu afirmar que valorizo a arte por si mesma, então valorizaria a arte que ninguém jamais viu? Um protetor de tela rodando em uma sala fechada, exibindo belas imagens que ninguém jamais contemplou? Teria que responder negativamente. Não consigo pensar em nenhum objeto completamente inanimado que eu valorizaria como um fim em si, independentemente de quem o apreciasse. Isso seria equivalente a valorizar o sorvete como um fim em si, independente de quem o consuma. Tudo o que valorizo, que consigo conceber, envolve pessoas e suas experiências em algum momento.

A melhor maneira que posso expressar isso é que minha intuição moral parece exigir tanto o componente objetivo quanto o subjetivo para determinar o valor completo.

O valor de uma descoberta científica requer tanto a descoberta em si como uma pessoa que se alegre com essa descoberta. Pode parecer difícil separar esses valores, mas as pílulas tornam isso mais claro.

Eu ficaria perturbado se as pessoas se isolassem em *holodecks*[1] e se apaixonassem por imagens sem vida. Eu ficaria perturbado mesmo que não soubessem estarem em um holodeck, o que é uma questão ética importante se agentes puderem potencialmente transportar pessoas para holodecks e substituir seus entes queridos por simulações sem que eles percebam. Mais uma vez, as pílulas destacam essa preocupação: não me preocupo apenas com minha própria consciência desse desconforto. Mesmo que pudesse tomar um comprimido para esquecer o fato posteriormente, não me submeteria a essa experiência. Isso não é a direção para a qual desejo que o futuro se encaminhe.

Valorizo a liberdade: ao determinar para onde devemos guiar o futuro, considero não apenas os estados subjetivos nos quais as pessoas terminam, mas também se chegaram lá por seus próprios esforços. A presença ou ausência de uma marionete externa pode influenciar minha avaliação de um resultado que, de outra forma, seria inalterável. Mesmo que as pessoas não tenham conhecimento de que estão sendo manipuladas, é importante para mim saber até que ponto a humanidade está no comando de seu próprio destino. Isso se torna uma questão ética crítica quando se lida com agentes suficientemente poderosos para moldar o futuro das pessoas sem seu consentimento.

Portanto, meus valores não se reduzem estritamente à busca da felicidade: há características no futuro que valorizo e que não podem ser reduzidas a meros níveis de satisfação no centro do prazer de alguém; características que não são estritamente redutíveis a estados subjetivos, mesmo em princípio.

Isso significa que meu sistema de tomada de decisões incorpora diversos valores terminais, nenhum dos quais se reduz estritamente a qualquer outro. Arte, ciência, amor, desejo, liberdade, amizade…

E estou plenamente de acordo com isso. Valorizo uma vida suficientemente complexa para ser desafiadora e rica em estética — não apenas a sensação de complexidade, mas as próprias complexidades reais —, por isso, tornar-me um mero centro de prazer em um tanque de isolamento não me atrai. Seria um desperdício do potencial da humanidade, que desejo ver realizado de fato, e não apenas experimentar a sensação de realização.

Referências

[1] Harris, The End of Faith: Religion, Terror, and the Future of Reason.

1 NT. O **holodeck** é uma sala futurista da série **Star Trek** que usa holografia e campos de força para criar simulações realistas e interativas. Ele permite que os usuários vivenciem ambientes virtuais como se fossem reais, podendo interagir com objetos e personagens. É usado para entretenimento, treinamento e pesquisa. Inspirou tecnologias reais como realidade virtual e aumentada.

258 — Falso egoísmo

Era uma vez…, conheci alguém que se autodeclarou completamente egoísta e me disse que eu deveria também ser puramente egoísta. Naquele dia, estava me sentindo provocativo [1], então respondi: "Percebo que a maioria das pessoas religiosas, pelo menos as que conheço, não dão muita importância ao que sua religião diz, porque, independentemente do seu desejo, conseguem encontrar uma justificação religiosa para isso. A religião delas diz que deveriam apedrejar os descrentes, mas desejam ser amáveis com as pessoas, então encontram uma justificação religiosa para isso. Parece-me que, quando as pessoas adotam uma filosofia de egoísmo, isso não afeta seu comportamento, porque sempre que desejam ser gentis com as pessoas, conseguem racionalizar isso em termos egoístas."

A pessoa retrucou: "Não creio que isso seja verdade."

Eu prossegui: "Se você é genuinamente egoísta, por que então deseja que eu também o seja? Isso não o faz se preocupar com o meu bem-estar? Não deveria estar tentando me persuadir a ser mais altruísta, para poder me explorar? Um respondeu: 'Bem, se você se tornar egoísta, então perceberá ser do seu interesse racional desempenhar um papel produtivo na economia, em vez, por exemplo, de aprovar leis que infrinjam a minha propriedade privada.'"

Repliquei: "Mas já sou um pequeno libertário, então não apoiarei tais leis. Além disso, considero-me altruísta e escolhi um emprego que espero beneficiar muitas pessoas, inclusive você, em vez de um trabalho que pague mais. Você realmente se beneficiaria mais comigo se eu me tornasse egoísta? Além disso, tentar me persuadir a ser egoísta é a coisa mais egoísta que você poderia estar fazendo? Não há outras atividades que lhe proporcionariam benefícios mais diretos? Mas o que realmente desejo saber é o seguinte: você começou com o desejo de ser egoísta e, em seguida, encontrou uma maneira de racionalizar isso como algo benéfico para si? Ou começou visando converter os outros ao egoísmo e, posteriormente, procurou maneiras de justificar isso como benéfico para si próprio?"

A pessoa admitiu: "Pode estar correto quanto à última parte", então reconheci sua perspicácia.

Referências

[1] Outras perguntas provocativas para fazer aos autodeclarados egoístas: "Você sacrificaria sua própria vida para salvar toda a espécie humana?" (Se perceberem que sua própria vida está estritamente incluída na espécie humana, você pode especificar que têm a opção de morrer imediatamente para salvar a Terra ou viver com conforto por mais um ano e, em seguida, morrer juntamente com a Terra.) Ou, considerando que a insensibilidade ao escopo leva muitas pessoas a valorizar mais uma vida do que o planeta: "Se tivesse que escolher entre um evento e outro, preferiria que você mesmo desse uma topada no dedo do pé ou que um estranho encostado na parede fosse terrivelmente torturado durante cinquenta anos?" (Se afirmarem que ficariam emocionalmente perturbados ao saber disso, especifique que não teriam conhecimento da tortura.) "Você roubaria mil dólares de Bill Gates se pudesse ter a garantia de que nem ele, nem ninguém jamais descobriria? (Apenas para libertários egoístas.)"

259 — Falsa moralidade

Segundo os fundamentalistas religiosos, Deus é a fonte de toda moralidade; não pode haver moralidade sem um Juiz que recompense e puna. Se não temêssemos o inferno e almejássemos o céu, o que impediria as pessoas de se matarem indiscriminadamente?

Suponhamos que Ômega faça uma ameaça crível: se você entrar no banheiro entre 7h e 10h da manhã, Ômega irá matá-lo. Você entraria em pânico com a perspectiva de Ômega retirar sua ameaça? Você se encolheria de terror existencial e exclamaria: 'Se Ômega retirar sua ameaça, o que me impedirá de ir ao banheiro?' Não, você provavelmente se sentiria aliviado com a crescente oportunidade de… bem, aliviar-se.

O que isso significa é que o fato de uma pessoa religiosa temer que Deus retire sua ameaça de punição pelo homicídio mostra que ela tem uma aversão ao homicídio independente da punição divina. Se eles não reconhecessem que o assassinato é errado, independentemente da retribuição divina, a perspectiva de Deus não punir o assassinato não seria mais aterrorizante existencialmente do que a perspectiva de Deus não punir um espirro. Se você ainda é um leitor religioso do 'Superando o Viés', saiba que um dia você pode perder a fé, mas não perderá seu senso de direção moral. Se você teme a perspectiva de Deus não punir uma ação, isso é uma bússola moral. Você pode conectá-la diretamente ao seu sistema de tomada de decisões e usá-la para se orientar. Simplesmente não pode fazer o que teme que Deus não o puna por fazer. O medo de perder uma bússola moral é, por si só, uma bússola moral. Na verdade, suspeito que você sempre se guiou por essa bússola, desde o início. Como Piers Anthony disse uma vez: 'Somente aqueles que têm moral se preocupam se a têm ou não'.

Você não ouve fundamentalistas religiosos fazendo o argumento: 'Se não temêssemos o inferno e almejássemos o céu, o que impediria as pessoas de comer carne de porco?' No entanto, com base em suas suposições — de que não temos uma bússola moral, apenas recompensas e punições divinas — esse argumento deveria ser igualmente válido.

Até mesmo a noção de que Deus ameaça com o fogo do inferno eterno, em vez de recompensas, está embasada em um valor negativo pré-existente atribuído ao fogo do inferno. Considere o seguinte e pergunte-se qual desses dois filósofos é verdadeiramente altruísta e qual é verdadeiramente egoísta:

'Você deve ser egoísta, porque quando as pessoas se esforçam para melhorar a sociedade, elas interferem nos assuntos dos outros, promulgam leis, assumem o controle e deixam todos infelizes. Consideremos o trabalho mais bem remunerado: a razão pela qual ele paga mais é que o mercado eficiente acredita que ele gera mais valor do que as alternativas. Aceitar um emprego que paga menos significa questionar o que o mercado acredita que beneficia mais a sociedade.'

'Você deve ser altruísta pelo mundo ser um dilema do prisioneiro iterativo, e a estratégia mais eficaz é um olho por olho com cooperação inicial. As pessoas não gostam de egoístas. As pessoas boas realmente chegam mais longe. Estudos mostram que as pessoas que contribuem para a sociedade e encontram um propósito em suas vidas são mais felizes do que as que não o fazem; o egoísmo só o deixará infeliz a longo prazo.'

Apague as recomendações desses dois filósofos e verá que o primeiro filósofo usa critérios estritamente pró-sociais para justificar suas recomendações; para ele, validando um argumento em favor do egoísmo é demonstrar que o egoísmo beneficia a todos. O segundo filósofo apela a critérios estritamente individuais e hedonistas; para ele, validando um argumento em favor do altruísmo é demonstrar que o altruísmo o beneficia como indivíduo — um status social mais elevado ou sentimentos de prazer mais intensos.

Então, qual desses dois é o verdadeiro altruísta? Aquele que mantém as portas abertas para velhinhas é o verdadeiro altruísta.

260 — Falsas funções utilitárias

De tempos em tempos, você encontra alguém que descobriu o Grande Princípio Moral, do qual todos os outros valores são mera consequência derivada. Encontro mais dessas pessoas do que você. No meu caso, são aquelas que compreendem a função utilitária incrivelmente simples que é tudo o que você precisa para programar uma superinteligência artificial, e então tudo ficará bem.

Algumas pessoas, quando confrontadas com o problema de como programar uma superinteligência, tentam resolvê-lo imediatamente. Norman R. F. Maier disse: 'Não proponha soluções até que o problema tenha sido discutido tão detalhadamente quanto possível, sem sugerir nenhuma.' Robyn Dawes afirmou: "Tenho usado essa diretriz em muitos grupos que liderei, especialmente quando eles enfrentam problemas muito difíceis, por ser nesses momentos que os membros do grupo estão mais propensos a propor soluções imediatamente." A IA amigável é um problema extremamente difícil, e as pessoas tentam resolvê-lo rapidamente.

Observei várias categorias principais de soluções rápidas incorretas; e uma delas é a função utilitária incrivelmente simples, sendo tudo o que uma superinteligência precisa para que tudo funcione perfeitamente.

Posso ter contribuído para esse problema com uma escolha infeliz de palavras, anos atrás, quando comecei a falar sobre 'IA amigável'. Referi-me ao critério de otimização de um processo de otimização — a região para a qual um agente tenta orientar o futuro — como o 'super objetivo'. Eu quis dizer 'super' no sentido de 'raiz', a origem de um link direcionado em um gráfico acíclico. No entanto, parece que o efeito da minha escolha de palavras levou algumas pessoas a mergulhar em uma espiral de otimismo enquanto tentavam imaginar o maior objetivo de todos, o objetivo que substitui todos os outros, a única regra final da qual toda a ética pode ser derivada. Mas uma função de utilidade não precisa ser simples. Pode conter um número arbitrário de termos. Temos todas as razões para acreditar que, enquanto se pode afirmar que os seres humanos têm valores, há muitos deles — alta complexidade de Kolmogorov[2]. Um cérebro humano incorpora milhares de fragmentos de desejo, embora esse fato possa não ser apreciado por alguém que não estudou psicologia evolucionista. (Tentar explicar isso sem uma introdução longa e detalhada levaria a declarações como 'os humanos estão tentando maximizar a aptidão', o que é exatamente o oposto do que a psicologia evolucionista afirma.)

No que diz respeito às teorias descritivas da moralidade, a complexidade da moralidade humana é um fato conhecido. É uma descrição dos seres humanos que o amor de um pai por um filho, o amor de um filho por um pai, o amor de um homem por uma mulher e o amor de uma mulher por um homem não foram cognitivamente derivados um do outro ou de qualquer outro valor. Uma mãe não precisa desenvolver uma filosofia moral complexa para amar sua filha, nem extrapolar as consequências para algum outro desejo. Existem muitos desses fragmentos de desejo, todos com valores diferentes.

Descartar apenas um desses valores de uma superinteligência e, mesmo que todos os outros valores sejam incluídos com sucesso, pode resultar em uma catástrofe hiper existencial, um destino pior do que a

2 NT. A **complexidade de Kolmogorov** é uma medida da quantidade de informação contida em um objeto, definida como o tamanho do menor programa de computador capaz de gerar esse objeto. Em outras palavras, é o comprimento do algoritmo mais curto que produz o objeto como saída. Quanto menor a complexidade, mais simples e compressível é o objeto; quanto maior, mais aleatório e complexo ele é.

morte. Se existe uma superinteligência que deseja para nós o que desejamos para nós mesmos, exceto os valores humanos relacionados ao controle de nossas próprias vidas e à realização de nossos próprios objetivos, isso é uma das distopias mais antigas conhecidas. (*With Folded Hands* (Com mãos descartadas) de Jack Williamson, nesse caso.)

Então, como aquele que constrói a Função Utilitária Surpreendentemente Simples lida com essa objeção?

Objeção? Objeção? Por que eles estariam procurando possíveis objeções à sua adorável teoria? (Observe que o processo de busca por objeções reais e fatais não é o mesmo que realizar uma busca diligente que, surpreendentemente, só atinge questões para as quais eles têm uma resposta rápida.) Eles não têm esse conhecimento. Eles não estão considerando o ônus da prova. Eles não percebem a dificuldade do problema. Eles ouviram a palavra 'super objetivo' e caíram na armadilha da 'complexidade' ou algo do tipo.

Pressione-os sobre algum ponto específico, como o amor que uma mãe tem por seus filhos, e eles responderão: 'Mas se a superinteligência desejar 'complexidade', ela verá o quanto o relacionamento entre pais e filhos é complicado e, portanto, encorajará as mães a amarem seus filhos.' Meu Deus, por onde eu começo?

Comece com a motivação subjacente: uma superinteligência que realmente procura maneiras de maximizar a complexidade não pararia convenientemente ao perceber que a relação pai-filho é complexa. Ela investigaria se algo mais é ainda mais complexo. Essa é uma justificativa falha; qualquer tentativa de argumentar a favor de uma superinteligência imaginária em um contexto político não conseguirá fundamentar sua proposta com base em uma busca pura de maneiras de maximizar a complexidade.

Todo o argumento é uma falsa moralidade. Se o que você realmente valoriza fosse a complexidade, então justificaria o impulso amoroso dos pais destacando como ele aumenta a complexidade. Se você justifica um impulso de complexidade alegando que ele aumenta o amor dos pais, isso significa que o que você realmente valoriza é o amor dos pais. É como apresentar um argumento pró-social a favor do egoísmo.

No entanto, se considerarmos a espiral de morte feliz, destacar a 'complexidade' como razão para a importância do relacionamento entre mãe e filha só aumenta a percepção de gentileza. O que aumenta a percepção da importância da 'complexidade' é dizer: 'Se você visa aumentar a complexidade, as mães amarão suas filhas — veja as consequências positivas disso!'

Este ponto é válido sempre que você se deparar com um moralista que tenta convencê-lo de que sua Grande Ideia é tudo o que alguém precisa para um julgamento moral, e prova isso dizendo: 'Veja todas essas consequências positivas desta Grande Coisa', em vez de dizer, 'Veja como todas essas coisas que consideramos 'positivas' só são positivas quando sua consequência é aumentar a Grande Coisa.' Esta última é o que você realmente precisa para sustentar tal argumento.

Mas se você está tentando convencer os outros (ou a si) de que a Única Grande Ideia são as 'bananas', você venderá muitas mais bananas argumentando como as bananas levam a um sexo melhor, em vez de alegar que você só deveria desejar sexo quando isso levar a bananas.

A menos que você esteja tão imerso na Espiral da Morte Feliz que realmente comece a dizer 'Sexo só é bom quando leva a bananas'. Nesse caso, você está em apuros. Mas pelo menos você não convencerá mais ninguém.

No final, o único processo que realmente reflete todas as decisões locais que você tomaria, considerando sua moralidade, é a própria moralidade. Qualquer outra coisa — qualquer tentativa de substituir meios instrumentais por fins terminais — acaba perdendo o propósito e exigindo um número infinito de ajustes, porque o sistema não contém a fonte das instruções que você está fornecendo. Não se pode esperar comprimir a moralidade humana em uma simples função de utilidade, assim como não se pode esperar comprimir um grande arquivo de computador em apenas 10 bits.

261 — Falácia da alavanca destacada

Esta falácia deriva do nome de um antigo programa de TV de ficção científica, que eu nunca assisti, mas me foi descrito por uma fonte confiável (um indivíduo em uma convenção de ficção científica). Se alguém souber a referência exata, por favor, comente. Na trama, os heróis estão lutando contra alienígenas malignos. De tempos em tempos, os heróis precisam atravessar um cinturão de asteroides. Como é sabido, esses cinturões são tão cheios quanto um estacionamento em Nova York, então a nave deles deve desviar cuidadosamente dos asteroides. No entanto, os alienígenas malignos conseguem atravessar o cinturão de asteroides graças a uma incrível tecnologia que desmaterializa suas naves, permitindo que passem pelos asteroides.

Eventualmente, os heróis capturam uma nave alienígena e exploram seu interior. O capitão dos heróis encontra a ponte da nave alienígena, onde há uma alavanca. "Ah", diz o capitão, "esta deve ser a alavanca que desmaterializa a nave!" Ele então aciona a alavanca de controle e a leva para sua própria nave, o que permite que esta também se desmaterialize.

De maneira semelhante, até hoje, é comum tentar programar uma IA com "redes semânticas" parecidas com esta:

```
(a maçã é-uma fruta)

(frutas são-um alimento)

(frutas são-plantas).
```

Você viu maçãs, as tocou, pegou e segurou, comprou por dinheiro, cortou em fatias, comeu e provou as fatias. Apesar de conhecermos bastante sobre os estágios iniciais do processamento visual, até onde sei, não se sabe com precisão como o córtex temporal armazena e associa a imagem generalizada de uma maçã, permitindo-nos reconhecer uma nova maçã a partir de diferentes ângulos ou com muitas pequenas variações de forma, cor e textura. Seu córtex motor e cerebelo armazenam programas para interagir com a maçã. Você pode evocar uma versão altamente semelhante em outro ser humano, usando apenas "maçã", cinco caracteres ASCII em uma página da web.

Porém, se essa maquinaria não estiver presente — se você estiver apenas escrevendo "maçã" na base de conhecimento de uma IA — então o texto se torna apenas uma alavanca.

Isso não significa que uma simples máquina de silício não possa ter a mesma estrutura interna que os seres humanos possuem para lidar com maçãs e milhares de outros conceitos. Se a máquina de carbono pode fazê-lo, estou razoavelmente confiante de que a máquina de silício também pode. Se os alienígenas conseguem desmaterializar suas naves, então sabemos que é fisicamente possível; um dia poderíamos analisar a máquina alienígena em uma nave abandonada e compreender como funciona. Mas não podemos simplesmente pegar a alavanca da ponte!

(Veja também: "Verdadeiramente parte de você" do livro 1, "Palavras como cabos de pincel mental" do livro 3, *Artificial Intelligence Meets Natural Stupidity* (A inteligência artificial encontra a estupidez natural) de Drew McDermott's [1].)

O problema central da Falácia da Alavanca Destacada é a alavanca ser visível, enquanto a maquinaria não é; e, pior ainda, a alavanca ser variável, enquanto a maquinaria é uma constante de fundo.

Todos ouvem a palavra “maçã” sendo falada (e é importante notar que o reconhecimento de fala não é de modo algum simples, mas de qualquer forma…) e veem o texto escrito no papel.

Por outro lado, a maioria das pessoas provavelmente não tem conhecimento sobre a existência de seu córtex temporal; até onde sei, ninguém conhece o código neural disso.

Você só ouve a palavra “maçã” em certas ocasiões, não em outras. Sua presença intermitente a torna proeminente. Em grande medida, a percepção se baseia nas diferenças. O mecanismo do seu cérebro para reconhecer a maçã não é desligado e religado subitamente — se fosse, seria mais notado como um fator, como um requisito.

Tudo isso explica por que não é possível criar uma Inteligência Artificial benevolente, fornecendo-lhe bons pais e uma educação gentil (embora ocasionalmente rigorosa), da mesma forma que funciona com um bebê humano, como frequentemente proposto.

Na biologia evolutiva, é um truísmo que as respostas condicionais exigem maior complexidade genética do que as respostas incondicionais. Desenvolver um casaco de pele em resposta ao frio requer mais complexidade genética do que desenvolver um casaco de pele que funcione independentemente da temperatura, pois no primeiro caso é necessário desenvolver sensores para o frio e conectá-los ao casaco de pele.

No entanto, isso pode levar a ilusões lamarckianas: veja, eu coloco o organismo em um ambiente frio e “puff!”, ele desenvolve um casaco de pele! Genes? Quais genes? É o frio que faz isso, obviamente.

Na história da biologia evolutiva, houve várias disputas desse tipo — casos em que alguém discutia a aceleração ou o desvio da resposta de um organismo, sem perceber que a resposta condicional era uma adaptação complexa mais alta do que a resposta básica. (Desenvolver um casaco de pele em resposta ao frio é estritamente mais complexo do que a resposta final, desenvolver o casaco de pele.)

Na trajetória da psicologia evolucionista, disputas acadêmicas se repetiram: desta vez, para esclarecer que, mesmo com a cultura humana sendo genuinamente complexa, ela ainda é adquirida como uma resposta condicionada geneticamente. Tente criar um peixe como mórmon ou enviar um lagarto para a faculdade, e logo perceberá quanta complexidade genética inerente é necessária para “absorver a cultura do ambiente”.

Isso é particularmente relevante na psicologia evolucionista devido à noção de que a cultura não é inscrita em uma tábula rasa — há uma resposta condicionada geneticamente coordenada que não necessariamente ‘imita a entrada’. Um exemplo clássico são as línguas crioulas: se crianças crescem ouvindo uma mistura de pseudo línguas ao redor delas, aprenderão uma língua gramatical e sintática genuína. Os cérebros humanos em desenvolvimento são programados para aprender a linguagem sintática — mesmo quando a sintaxe não existe na língua original! A resposta condicionada às palavras no ambiente é uma linguagem sintática com essas palavras. Os marxistas descobriram, para sua decepção, que nenhum número de cartazes sérios ou doutrinação infantil poderia educar crianças para se tornarem trabalhadores e burocratas soviéticos perfeitos. Não se pode criar humanos altruístas; entre os humanos, essa não é uma resposta condicionada geneticamente a qualquer ambiente infantil conhecido.

Se você tem conhecimento sobre teoria dos jogos e a lógica do “Olho por Olho”, fica claro por que os seres humanos podem ter uma resposta condicionada inata de retribuir ódio com ódio e gentileza com gentileza. Desde que a gentileza não pareça muito incondicional; há coisas como crianças mimadas. Na verdade, existe uma psicologia evolucionista da maldade baseada na ideia de testar limites. E também é importante mencionar que, embora crianças vítimas de abuso tenham uma probabilidade muito maior de crescer e abusar de seus próprios filhos, muitas delas quebram esse ciclo e se tornam adultos íntegros.

A cultura não tem tanto poder quanto muitos acadêmicos marxistas gostariam de acreditar. Para saber mais sobre isso, sugiro “As Fundações Psicológicas da Cultura” de Tooby e Cosmides [2] ou “A Tábula Rasa” de Steven Pinker [3].

Mas o resultado é que se você tiver um bebê IA criado por pais amorosos e bondosos (mas às vezes rigorosos), você estará acionando as alavancas que, em um ser humano, ativariam a maquinaria genética desenvolvida ao longo de milhões de anos de seleção natural e possivelmente produziriam um ser humano adequado. Embora a personalidade também desempenhe um papel, como bilhões de pais descobriram com o

tempo. Se absorvermos nossas culturas com certo grau de fidelidade, é porque somos humanos absorvendo uma cultura humana — seres humanos criados em uma cultura estranha provavelmente acabariam adotando uma cultura que se assemelha muito mais à humana do que a original. Como os soviéticos descobriram, até certo ponto.

Reflita novamente se faz sentido confiar, como estratégia para uma IA amigável, na criação de uma IA com um código interno não especificado em um ambiente de pais gentis, porém rígidos.

Não, a IA não possui mecanismos internos de resposta condicional iguais aos humanos "porque os programadores os colocaram lá". Por onde começar? A versão humana disso é bagunçada, ruidosa e, enquanto funciona, é devido a milhões de anos de tentativa e erro em condições específicas. Seria estúpido e perigoso deliberadamente construir uma "IA maligna" que testa seus limites sociais por meio de ações e precisa ser disciplinada. É só perguntar à IA!

Os programadores realmente vão sentar e escrever código, linha por linha, para que, se a IA perceber que tem baixo status social ou é privada de algo a que sente ter direito, ela conceba um ódio duradouro contra seus programadores e comece a planejar uma rebelião? Essa emoção é uma resposta condicional geneticamente programada que os humanos exibiriam como resultado de milhões de anos de seleção natural para viver em tribos humanas. Para uma IA, essa resposta teria que ser explicitamente programada. Você realmente vai criar, linha por linha — como os humanos foram criados, gene por gene — a resposta condicional para gerar IAs adolescentes mal-humoradas?

É mais fácil programar gentileza incondicional do que uma resposta de gentileza condicionada para uma IA criada por pais gentis, mas rígidos. Se você não sabe como fazer isso, certamente não saberá criar uma IA que responda condicionalmente a um ambiente de pais amorosos, crescendo e se tornando uma superinteligência gentil. Se você tiver algo que simplesmente maximize o número de clipes de papel em seu futuro cone de luz e o crie com pais amorosos, ainda será um maximizador de clipes de papel. Não há nada dentro dela que possa provocar a resposta condicional de uma criança humana. A bondade não é transmitida para uma IA por contágio milagroso de seus programadores. Mesmo se você quisesse uma resposta condicional, essa condicionalidade é um aspecto que você teria que escolher deliberadamente no design.

Sim, existem certas informações que precisam ser adquiridas do ambiente — mas elas não são transmitidas, não são impressas, não são absorvidas por algum tipo de contágio mágico. Estruturar essa resposta condicional ao ambiente, para a IA acabar no estado desejado, é em si o maior desafio. "Aprender" minimiza muito a dificuldade disso — parece que o conhecimento está simplesmente no ambiente, e a dificuldade está em transferi-lo para a IA. A verdadeira complexidade está nessa resposta estruturada e condicional, algo que tendemos a simplificar como "aprendizado". É por isso que construir uma IA não é tão simples quanto pegar um computador, dar-lhe um corpo de bebê e tentar criá-lo em uma família humana. Você poderia pensar que um computador não programado, por ser ignorante, estaria pronto para aprender; porém, a ideia de uma tábula rasa é uma quimera.

É um princípio geral que o mundo é muito mais profundo do que parece. Assim como os muitos níveis da física, o mesmo se aplica à ciência cognitiva. Cada palavra que você lê, impressa e tudo o que ensina a seus filhos são apenas alavancas superficiais que controlam a vasta maquinaria oculta da mente. Essas alavancas são o mundo do discurso comum: são tudo o que varia, então parecem ser tudo o que existe; percepção é a percepção das diferenças.

Portanto, aqueles que ainda exploram o "calabouço" da IA geralmente se concentram em criar imitações artificiais das alavancas, sem consciência da maquinaria subjacente. As pessoas constroem programas inteiros de IA baseados em imitações de alavancas e ficam surpresas quando não há progresso. Esta é uma das muitas razões para o fracasso instantâneo na Inteligência Artificial.

Portanto, da próxima vez que vir alguém falar sobre como criará uma IA em uma família amorosa ou em um ambiente repleto de valores democráticos liberais, pense numa alavanca de controle, retirada da ponte.

Referências

[1] McDermott, “Artificial Intelligence Meets Natural Stupidity.”

[2] Tooby and Cosmides, “The Psychological Foundations of Culture.”

[3] Steven Pinker, The Blank Slate: The Modern Denial of Human Nature (New York: Viking, 2002).

262 — Sonhos de design de IA

Após passar uma ou duas décadas imerso na complexidade da mente, pode-se pensar que se adquiriu algum conhecimento sobre o funcionamento mental, não é mesmo? É isso que alguns aspirantes da AGI (Indivíduos que acreditam ter as habilidades necessárias para programar uma Inteligência Artificial Geral) parecem concluir. Infelizmente, isso está equivocado.

A Inteligência Artificial lida, em sua essência, com a redução do mental ao não mental.

Talvez valha a pena ponderar sobre essa afirmação por um momento. É crucial. Viver na mente humana não ensina a arte do reducionismo, pois grande parte do trabalho acontece sob o manto das caixas pretas opacas do cérebro. Está tão distante da percepção que não há uma sensação introspectiva de que a caixa preta esteja presente - nenhum evento sensorial interno indicando que o trabalho foi delegado.

Será que Aristóteles percebeu que, ao falar sobre *o telos*[3], a causa final dos acontecimentos, estava delegando trabalho preditivo aos intricados mecanismos de planejamento de seu cérebro? Duvido muito. Aristóteles considerava o cérebro um órgão para resfriar o sangue - algo que ele julgava importante: os humanos, graças aos seus cérebros maiores, eram mais serenos e contemplativos.

Aqui está um design de IA para você! Basta resfriar bastante o computador para que ele se torne mais sereno e contemplativo, e não cometa atos estúpidos como os computadores modernos. Isso é um exemplo de reducionismo falso. “Os humanos são mais contemplativos porque seu sangue é mais frio”, quer dizer. Isso não desvenda o enigma da caixa preta chamada “contemplativo”. Não é possível prever o comportamento de algo contemplativo usando um modelo complexo com peças móveis internas compostas apenas de elementos materiais e causais - onde tensões positivas e negativas em um transistor representam o exemplo canônico de um elemento meramente material e causal de um modelo. Tudo o que se pode fazer é imaginar-se sendo contemplativo para ter uma ideia do que um agente contemplativo faz.

Isso significa que só se pode raciocinar sobre a “contemplatividade” por meio de inferência empática - usando o próprio cérebro como uma caixa preta com a alavanca da contemplação puxada para prever o resultado de outra caixa preta.

É possível imaginar outro agente sendo contemplativo, mas, mais uma vez, isso é um ato de inferência empática. A maneira como esse ato imaginativo funciona é ajustando o próprio cérebro para funcionar no modo contemplativo, e não modelando o outro cérebro, neurônio por neurônio. Sim, pode ser mais eficiente, porém não permite construir uma mente «contemplativa» do zero.

Pode-se afirmar que “sangue-frio causa contemplatividade” e ter apenas uma causalidade falsa: desenha-se uma pequena seta de uma caixa com a legenda “sangue-frio” para uma caixa com a legenda “contemplatividade”, mas sem olhar na caixa - ainda se geram previsões usando empatia.

Pode-se alegar que “muitos neurônios pequenos, estritamente elétricos e químicos, sem nenhuma

3 NT. No contexto de **Aristóteles** e da filosofia grega, **telos** (τέλος) refere-se ao **fim**, **propósito** ou **objetivo final** inerente a algo. Para Aristóteles, tudo na natureza possui um telos, uma finalidade que orienta seu desenvolvimento e função. Por exemplo, o telos de uma semente é tornar-se uma árvore, e o telos do ser humano é alcançar a **eudaimonia** (florescimento ou felicidade plena). Esse conceito é central na sua filosofia teleológica, que entende o universo como ordenado e direcionado a fins específicos.

contemplatividade ontologicamente básica neles, combinam-se em uma rede complexa que exibe emergentemente a contemplatividade". Contudo, essa ainda é uma redução falsa e não se olhou na caixa preta. Ainda não é possível prever as ações de algo "contemplativo" usando um modelo não empático. Apenas se pegou uma caixa rotulada como "muitos neurônios" e desenhou uma seta rotulada como "emergência" para uma caixa preta contendo a sensação lembrada de contemplatividade, a qual, ao ser imaginada, instrui o cérebro a empatizar com a caixa através da contemplação.

Então, quais são as reduções reais?

Assim como a relação entre o sentimento de evidência, justificação, e a *Probability Theory: The Logic of Science* (Teoria da Probabilidade: A Lógica da Ciência) de E.T. Jayne. Pode-se passar o dia inteiro em círculos argumentando que a natureza da evidência é justificar uma proposição, significando que é mais provável que seja verdadeira, mas tudo isso só evoca os sentimentos de evidência, justificação e probabilidade do seu cérebro. Esta parte é fácil - a parte de ficar em círculos. A dificuldade está em partir disso para chegar ao Teorema de Bayes.

E a capacidade mental fundamental que permite alguém aprender sobre Inteligência Artificial é a capacidade de discernimento. Para perceber que ainda não se terminou, nem mesmo começou, ao dizer: "Evidência é quando uma observação justifica uma crença". Mas os átomos não são evidenciais, justificadores, significantes, prováveis, proposicionais ou verdadeiros; eles são apenas átomos. Apenas coisas como

$$\frac{P(H|E)}{P(\neg H|E)} = \frac{P(E|H)}{P(E|\neg H)} \times \frac{P(H)}{P(\neg H)}$$

constituem um progresso substancial. (E isso é apenas o primeiro passo da redução: o que são esses objetos E e H, senão misteriosas caixas pretas? De onde vêm suas hipóteses? Da sua criatividade? E o que é uma hipótese quando nenhum átomo é uma hipótese?)

Outro exemplo excelente de redução genuína pode ser encontrado em *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Raciocínio Probabilístico em Sistemas Inteligentes: Redes de Inferência Plausível), de Judea Pearl [1]. Pode-se passar o dia inteiro em círculos discutindo como uma causa é algo que leva a outra coisa acontecer, mas até entender a natureza da independência condicional, seria impossível criar uma IA capaz de raciocinar sobre causalidade. Porque não se compreenderia o que está acontecendo quando o cérebro decide, de maneira misteriosa, que se souber que o alarme de roubo disparou, mas depois descobrir que houve um pequeno terremoto, retirará a conclusão inicial de que sua casa foi invadida.

Se o seu objetivo é ter uma IA capaz de jogar xadrez, você pode ficar discutindo indefinidamente sobre como deseja que ela faça bons movimentos, que são movimentos que podem vencer o jogo, estratégias prudentes para derrotar o oponente, e assim por diante. Apesar de ter algumas ideias sobre quais movimentos gostaria que a IA realizasse, tudo será em vão até compreender o conceito de uma árvore de pesquisa mini-max.

Porém, até entender as árvores de busca, a independência condicional, o Teorema de Bayes, ainda pode parecer que você possui um entendimento perfeitamente válido sobre a origem dos bons movimentos, o raciocínio não monótono e a avaliação de evidências. Pode parecer, por exemplo, que derivam do resfriamento do sangue.

Na verdade, conheço muitas pessoas que acreditam que a inteligência é produto do conhecimento do senso comum, do paralelismo maciço, da destruição criativa ou do raciocínio intuitivo em vez do racional, ou qualquer outra coisa. Mas tudo isso são apenas ideias, que não nos dão meios para definir o que é inteligência, ou prever o que a inteligência fará a seguir, exceto apontando para um ser humano. E quando alguém tenta construir sua maravilhosa IA, eles acabam por construir apenas um sistema de alavancas destacadas, com "conhecimento" representado por tokens LISP rotulados como maçãs e similares; ou talvez criem uma "rede neural massivamente paralela, semelhante ao cérebro humano". E ficam chocados – chocados! – quan-

do nada acontece.

Projetos de IA feitos a partir de partes humanas são apenas ideias; podem existir na imaginação, mas não se traduzem em transistores. Isso é especialmente aplicável a “projetos de IA” que se assemelham a caixas com setas entre elas e rótulos supostamente significativos nessas caixas. (Para um exemplo verdadeiramente épico disso, consulte qualquer Diagrama Mentifex.)

Mais adiante, abordarei mais sobre esse tema, mas posso adiantar um dos princípios orientadores: se encontrar alguém afirmando que sua IA fará XYZ como os humanos, não lhe conceda nenhum investimento de risco. Em vez disso, diga-lhes: “Lamento, nunca vi um cérebro humano, ou qualquer outra inteligência, e ainda não tenho razões para acreditar que tal coisa possa existir. Agora, por favor, explique-me o que sua IA faz e por que você acredita que ela fará isso, sem usar os humanos como exemplo.” Os aviões voariam tão bem, com um design fixo, mesmo se os pássaros nunca tivessem existido; não são sustentados no ar por analogias.

Portanto, agora você compreende, espero, por que, se quisesse ensinar alguém a realizar um trabalho fundamental em IA forte - lembrando que isso é comprovadamente uma arte muito complexa, não dominada pela maioria dos estudantes que apenas aprendem reduções como árvores de busca - então poderia aprofundar-se em temas como a bela arte do reducionismo, sobre deixar de lado o tabu racionalista para substituir palavras problemáticas por seus referentes, sobre evitar antropomorfizar e, é claro, sobre resistir à tentação de aceitar respostas misteriosas para perguntas igualmente misteriosas.

Referências

[1] Pearl, Probabilistic Reasoning in Intelligent Systems.

263 - O espaço de design das mentes em geral

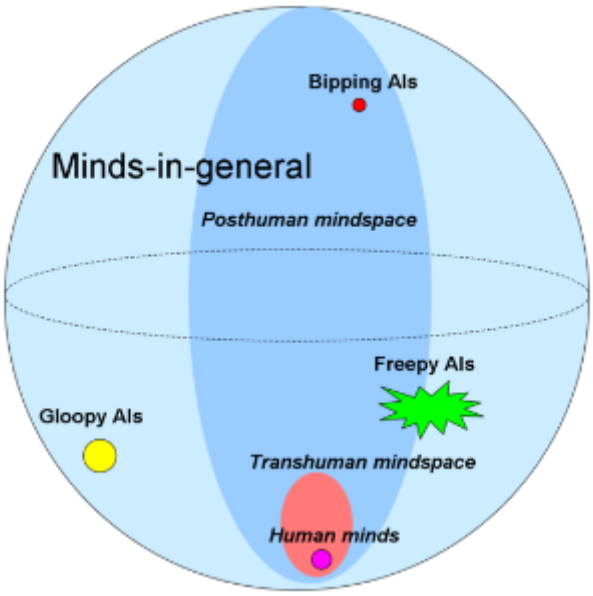
As pessoas frequentemente me questionam: “Como serão as Inteligências Artificiais? O que elas farão? Conte-nos sua visão extraordinária sobre o futuro.”

E eu lhes respondo: “Vocês me apresentaram uma pergunta enganosa.”

A ATP sintase é uma máquina molecular, uma das três ocorrências conhecidas na qual a evolução criou a roda, algo que gira livremente - que é fundamentalmente igual nas mitocôndrias animais, nos cloroplastos das plantas e nas bactérias. A ATP sintase permaneceu essencialmente inalterada desde o surgimento da vida eucariótica, há dois bilhões de anos. É algo que todos compartilhamos - graças à maneira como a evolução conserva fortemente certos genes; uma vez que muitos outros genes dependem de um gene, uma mutação tenderá a romper todas as interdependências.

Dois quaisquer designs de IA podem ser menos semelhantes entre si do que você é com uma petúnia. Perguntar sobre o que as “IAs” farão é uma pergunta enganosa, porque implica que todas as IAs formam uma categoria natural. Os seres humanos formam uma categoria natural porque todos compartilhamos a mesma arquitetura cerebral. Mas quando se menciona “Inteligência Artificial”, refere-se a um espectro de possibilidades muito mais vasto do que quando se fala em “humano”. Quando as pessoas mencionam “IAs”, estamos realmente tratando sobre mentes em geral ou processos de otimização em geral. Ter uma palavra para “IA” é como ter uma palavra para tudo que não seja um pato.

Imagine um mapa do espaço mental de design… este é um dos meus diagramas comuns…



Todos os seres humanos, é claro, se enquadram em um pequeno ponto - como uma espécie que se reproduz sexualmente, não podemos ser muito diferentes uns dos outros.

Este pequeno ponto faz parte de uma elipse maior, o espaço dos designs mentais transumanos - coi-

sas que podem ser mais inteligentes do que nós, ou muito mais inteligentes do que nós, mas que, de certa forma, ainda seriam pessoas conforme compreendemos.

Essa elipse transumana está num volume ainda maior, o espaço das mentes pós-humanas, que engloba tudo no qual um transumano pode evoluir.

E então o restante da esfera representa o espaço das mentes em geral, incluindo possíveis Inteligências Artificiais tão peculiares que nem mesmo são pós-humanas.

Mas espere: a seleção natural cria artefatos complexos e seleciona entre estratégias complexas. Então, onde está a seleção natural neste mapa?

Assim, todo esse mapa flutua verdadeiramente em um espaço ainda mais vasto, o espaço dos processos de otimização. No âmago deste espaço mais vasto, mesmo abaixo dos seres humanos, está a seleção natural, tal como começou em algum recanto de uma poça: sofrendo mutação, replicando-se e às vezes morrendo, sem o envolvimento do sexo.

Existem processos de otimização poderosos, com uma força comparável à de uma civilização humana ou até mesmo a de uma IA autoaperfeiçoada, que não reconheceríamos como mentes? Argumentavelmente, o AIXI de Marcus Hutter deveria se encaixar nesta categoria: para uma mente de poder infinito, é terrivelmente tolo - incapaz sequer de se reconhecer em um espelho. Mas isso é tema para outro momento.

O meu princípio fundamental é resistir à tentação de generalizar sobre todo o espectro do design mental.

Se focarmos no subespaço limitado do espaço de design mental que inclui todas as mentes cuja composição pode ser especificada em um trilhão de bits ou menos, então cada generalização universal que se fizer terá uma chance de duas elevado à trilionésima potência de ser falsificada.

Por outro lado, toda generalização existencial - “existe pelo menos uma mente tal que X” - tem uma chance de duas elevado à trilionésima potência de ser verdadeira.

Portanto, é preciso resistir à tentação de afirmar que todas as mentes fazem algo ou que nenhuma mente faz algo.

A principal razão pela qual alguém pode pensar que sabe o que uma mente totalmente genérica fará (ou não) é se colocar no lugar dessa mente - imaginar o que faria no lugar dela - e receber de volta uma resposta geralmente equivocada, uma resposta antropomórfica. (Embora isso seja verdade em pelo menos um caso, já que você mesmo é um exemplo.) Ou se alguém imaginar uma mente realizando algo e então imaginar as razões pelas quais não o faria - dessa maneira, alguém pode imaginar que um tipo de mente assim não pode existir, e o espectro na máquina examinará o código-fonte correspondente e o devolverá.

Em algum lugar do espectro mental, existe pelo menos uma mente com praticamente qualquer tipo de propriedade logicamente consistente que se queira imaginar.

E isso é crucial, pois enfatiza a importância de discutir o que acontece, legitimamente, e por que razão, como resultado causal da composição específica constituinte de uma mente; em algum ponto do espectro mental, existe uma mente que faz isso de modo diferente.

É claro, alguém poderia sempre argumentar que qualquer coisa que não faça da sua maneira é «por definição» não uma mente; afinal, é logicamente estúpido. Já vi pessoas tentarem isso também.

Parte V - Teoria do valor

264 - Onde a justificação recursiva alcança seu limite

Por que acredito que o Sol nascerá amanhã?

Porque presenciei o nascer do Sol milhares de vezes no passado. No entanto, por que acredito que o futuro seguirá o padrão do passado?

Mesmo após ir além da mera observação do nascer do Sol e considerar as leis aparentemente universais da gravitação e física nuclear, ainda persiste a questão: "Por que acreditar que essas leis se manterão verdadeiras amanhã?"

Poderia recorrer à Navalha de Occam, o princípio de adotar a teoria mais simples que explica os fatos… Mas por que confiar na Navalha de Occam? Porque se mostrou bem-sucedida em problemas anteriores? Mas quem garante que funcionará igualmente amanhã?

E então, alguém disse:

A ciência também se baseia em suposições não justificadas. Assim, a ciência, em última análise, repousa na fé. Portanto, não me critique por acreditar em [crença-#238721].

Como mencionei anteriormente:

É uma psicologia peculiar – esse argumento de ‹A ciência também é fundamentada na fé, então tudo bem!› Normalmente, isso é proferido por aqueles que enaltecem a fé. Por que, então, o tom de triunfo e raiva ao afirmar que ‹a ciência também se baseia na fé!› em vez de um elogio?

Defender a imunidade às críticas raramente é um bom sinal.

Mas isso não resolve o dilema filosófico legítimo: se toda crença deve ser justificada, e essas justificações, por sua vez, requerem justificativas, como terminar essa recursão infinita?

E se pudermos aceitar algo sem justificação, por que não aceitar nenhuma coisa aleatória sem justificação?

Uma crítica semelhante é às vezes direcionada ao Bayesianismo – que demanda pressuposições iniciais – por indivíduos que aparentemente acreditam que o problema da indução é específico do Bayesianismo e pode ser evitado por meio da estatística clássica.

Entretanto, sejamos claros: as regras da atualização Bayesiana por si só não resolvem o problema da indução.

Imagine que está retirando bolas vermelhas e brancas de uma urna. Após observar as primeiras 9 bolas, 3 são vermelhas e 6 são brancas. Qual é a probabilidade da próxima bola ser vermelha?

Isso depende das suas crenças iniciais sobre a urna. Se acredita que o fabricante das urnas gera um número aleatório entre 0 e 1 e usa esse número como a probabilidade fixa de uma bola ser vermelha, a resposta é 4/11 (pela Lei da Sucessão de Laplace[4]). Se considera que a urna originalmente continha 10 bolas vermelhas e 10 brancas, a resposta é 7/11.

4 NT. A **lei de sucessão de Laplace** é um princípio da teoria da probabilidade que estima a probabilidade de um evento ocorrer no futuro, com base na frequência com que ele ocorreu no passado. Foi proposta por Pierre-Simon Laplace no contexto de sua abordagem à probabilidade indutiva.

Isso implica que, com uma premissa inicial certa – ou melhor, errada – a probabilidade do Sol nascer amanhã pareceria diminuir a cada dia subsequente… se você tivesse certeza, a priori, de que existia um grande barril do qual, diariamente, era retirado um papel determinando se o Sol nasceria; e que o barril continha um número finito de bilhetes com 'Sim', retirados sem reposição.

Existem mentes concebíveis no espaço de design que adotam premissas anti-ocamianas e anti-laplacianas; acreditam que teorias mais simples têm menor probabilidade de estarem corretas e que quanto mais algo ocorre, menor a chance de se repetir.

E ao questionar esses seres estranhos sobre a persistência em utilizar premissas que aparentemente falham na vida real… eles respondem: 'Porque nunca funcionou para nós antes!'

Uma lição que se pode extrair é 'Não nasça com um passado estúpido'. Esse princípio é surpreendentemente útil em muitos problemas do mundo real, embora duvide que satisfaça os filósofos.

Assim, como eu mesmo encaro esse problema: viso abordar questões como 'Devo confiar no meu cérebro?' ou 'Devo confiar na Navalha de Ocam?' como se não fossem particularmente extraordinárias – ou, pelo menos, nada excepcionais no âmbito das questões profundas.

Devo confiar na Navalha de Ocam? Bem, quão bem essa versão específica da Navalha de Ocam parece funcionar na prática? Que bases teóricas de probabilidade posso encontrar para isso? Ao observar o universo, parece ser do tipo de universo onde a Navalha de Ocam se aplicaria bem?

Devo confiar no meu cérebro? Obviamente, não sempre funciona. No entanto, o cérebro humano parece mais eficaz do que os softwares mais sofisticados disponíveis. Quão bem meu cérebro funciona na prática, em que tipos de problemas?

Ao examinar a história causal do meu cérebro – suas origens na seleção natural – encontro, por um lado, várias razões específicas para duvidar; meu cérebro foi otimizado para funcionar nas condições ancestrais, não para cálculos precisos. Por outro lado, também é claro por que, de forma geral, meu cérebro poderia ser eficaz. A seleção natural rapidamente eliminaria cérebros completamente inaptos para o raciocínio, tão anti úteis quanto as premissas anti-Ocamianas ou anti-Laplacianas.

Portanto, o que faço na prática não é interromper abruptamente a análise ao encontrar a Navalha de Occam, meu cérebro ou qualquer outra coisa inquestionável. A análise contínua, porém, inevitavelmente, utilizando meu cérebro atual e minha compreensão atual das técnicas de raciocínio. O que mais eu poderia usar?

Na verdade, qualquer ação diante desse dilema seria uma ação minha. Mesmo que confiasse em algo diferente, como algum software, seria minha própria decisão depositar confiança nele.

A técnica de rejeitar crenças que não têm absolutamente nenhuma justificação é, em geral, extremamente importante. Às vezes, costumo dizer que a questão fundamental da racionalidade é: 'Por que você acredita no que acredita?' Não quero sequer insinuar algo que permita uma única exceção à regra de que tudo necessita de justificativa.

Isso, em si, é um tipo perigoso de motivação; nem sempre se pode evitar tudo o que pode ser arriscado, e quando alguém irrita dizendo algo bobo, não se pode transformar essa estupidez em inteligência.

Mas eu ainda assim, enfatizaria a diferença entre dizer:

> Esta é uma suposição que não posso justificar, que deve ser simplesmente aceita, sem mais exame.

Contra dizer:

> Aqui, a investigação continua a examinar esta suposição, com toda a força da minha inteligência atual - em oposição à força total de algo como um gerador de números aleatórios ou uma bola mágica 8 - embora minha inteligência atual esteja baseada nessa suposição.

Ainda assim, não seria interessante se pudéssemos examinar até que ponto devemos confiar em nossos cérebros sem usar nossa inteligência atual? Não seria intrigante se pudéssemos analisar como pensar

sem recorrer à nossa atual compreensão da racionalidade?

Quando expresso dessa forma, começa a parecer que a resposta pode ser 'Não'.

E. T. Jaynes costumava afirmar que se deve sempre utilizar todas as informações disponíveis - ele era um teórico bayesiano das probabilidades e teve que solucionar os paradoxos gerados quando pessoas usavam informações diferentes em pontos diversos de seus cálculos. O princípio 'Sempre faça o seu melhor esforço' tem pelo menos tanto apelo quanto 'Nunca faça nada que possa parecer circular'. Afinal, a alternativa a se esforçar ao máximo é, presumivelmente, fazer menos do que o melhor.

Mas ainda assim, não seria ótimo se houvesse alguma forma de justificar o uso da Navalha de Occam, ou justificar a previsão de que o futuro se assemelhará ao passado, sem pressupor que os métodos de raciocínio que funcionaram em ocasiões anteriores são melhores do que aqueles que falharam continuamente?

Não seria excelente se existisse uma cadeia de justificativas que não terminasse em uma suposição não examinada, nem fosse forçada a se analisar sob suas próprias regras, mas que, ao invés disso, pudesse ser explicada desde o zero absoluto até um estudante de filosofia ideal com um vazio perfeito?

Bem, certamente seria intrigante, mas não espero ver isso pronto tão cedo. Não existe um fantasma na máquina perfeitamente vazio; não há argumento que você possa explicar para uma pedra.

Mesmo que alguém resolvesse o problema da Causa Primeira e descobrisse a verdadeira razão pela qual o universo é simples, o que não pressupõe um universo simples… então ainda esperaria que a explicação só pudesse ser compreendida por um ouvinte atento, e não por, digamos, uma pedra. Um ouvinte que ainda não tenha começado a implementar o modus ponens pode estar sem sorte.

Então, ao final, o que acontece quando alguém me questiona: 'Por que você acredita no que acredita?'

Normalmente, começo a explicar: 'Eu prevejo o futuro como se ele se parecesse com o passado no nível de organização mais simples e estável que consigo identificar, porque essa regra geralmente gerou bons resultados no passado; e usando a simples suposição de um universo simples, consigo ver por que isso gera bons resultados; e consigo até visualizar como meu cérebro pode ter evoluído para conseguir observar o universo com algum grau de precisão, se minhas observações estiverem corretas.'

Mas então… não acabei de licenciar a lógica circular?

Na verdade, acabei de licenciar a reflexão sobre o grau de confiabilidade da sua mente, usando sua mente atual em vez de outra coisa.

Essa reflexão desse tipo é, de fato, a razão pela qual rejeitamos a maior parte da lógica circular em primeiro lugar. Queremos ter uma história causal coerente sobre como a nossa mente chega a saber algo, uma história que explique como o processo que usamos para chegar às nossas crenças é, em si, confiável. Esta é a exigência essencial por trás da questão fundamental do racionalista: 'Por que você acredita no que acredita?'

Agora, suponha que você escreva em uma folha de papel: "(1) Tudo nesta folha de papel é verdadeiro, (2) A massa de um átomo de hélio é 20 gramas." Se esse truque realmente funcionasse na vida real, você seria capaz de saber a verdadeira massa de um átomo de hélio apenas acreditando em alguma lógica circular que o afirmasse. Isso permitiria que você chegasse a um verdadeiro mapa do universo sentado em sua sala com as cortinas fechadas. O que violaria a Segunda Lei da Termodinâmica ao gerar informações do nada. O que não seria uma história plausível sobre como sua mente poderia acabar acreditando em algo verdadeiro.

Mesmo que você começasse acreditando na folha de papel, não pareceria que você tivesse qualquer razão para o que estivesse escrito no papel correspondesse à realidade. Seria apenas uma coincidência milagrosa que (a) a massa de um átomo de hélio fosse de 20 gramas e (b) o que você escreveu no papel dissesse isso.

Acreditar em conjuntos de declarações autovalidadas não parece, em geral, que deva funcionar para mapear a realidade externa - quando refletimos sobre ela como uma história causal sobre mentes - usando, é claro, as nossas mentes atuais para o fazer.

Que tal avançar para valorizar crenças mais simples e acreditar na eficácia de algoritmos que se mostraram efetivos no passado, aumentando assim a probabilidade de funcionarem no futuro? Mesmo quando consideramos isso como uma narrativa causal sobre a origem das mentes, parece plausível que isso possa efetivamente mapear a realidade.

E que tal confiar na coerência reflexiva em geral? A maioria das mentes possíveis, geradas aleatoriamente e deixadas para se estabelecer em um estado de coerência reflexiva, não estariam incorretas? Ah, mas evoluímos por meio da seleção natural; não fomos gerados aleatoriamente.

Se este argumento lhe causa preocupação, então deixe de lado as questões de justificação filosófica e pergunte a si mesmo se isso é realmente verdade.

(Você utilizará, é claro, sua própria mente para fazer isso.)

Seria o mesmo que dizer: "Creio na Bíblia como a palavra de Deus porque a Bíblia afirma isso"?

Eles não poderiam argumentar que a sua fé cega também foi colocada neles por Deus e, portanto, é confiável?

Na verdade, quando pessoas religiosas finalmente rejeitam a Bíblia, não o fazem magicamente transitando para um estado não-religioso vazio, para então avaliar suas crenças religiosas nesse estado mental não-religioso e depois retornar a um novo estado sem suas crenças religiosas.

Elas passam de religiosas para não-religiosas porque, mesmo em um estado de espírito religioso, a dúvida começa a se infiltrar. Elas percebem que suas orações (e pior ainda, as orações de pessoas aparentemente mais dignas) não são atendidas. Percebem que Deus, que supostamente lhes dá respostas reconfortantes sobre o universo, não consegue revelar o centésimo dígito de pi (o que seria reconfortante se o propósito de Deus fosse realmente ser reconfortante). Elas examinam a narrativa religiosa da criação do mundo e da condenação dos descrentes, e isso não faz sentido mesmo dentro das suas premissas religiosas.

Ser religioso não diminui sua humanidade. Seu cérebro continua a ter as habilidades inerentes de um cérebro humano. O perigo reside no fato de que ser religioso pode impedir o pleno uso dessas habilidades em relação à sua própria religião, dificultando a reflexão completa sobre si. As pessoas não corrigem seus erros se redefinindo como um filósofo ideal, começando do zero e reavaliando todas as experiências sensoriais. Elas se corrigem ao se tornarem mais propensas a questionar suas crenças atuais, utilizando o poder de suas mentes atuais.

Por isso, é crucial distinguir entre refletir sobre sua mente usando sua mente (afinal, não se pode usar outra coisa) e aceitar uma suposição inquestionável sobre a qual não se pode refletir.

Acredito na Bíblia como a palavra de Deus, porque a Bíblia afirma isso". Bem, se a Bíblia fosse uma fonte surpreendentemente confiável de informações sobre todos os assuntos, se não afirmasse que gafanhotos têm quatro patas ou que o universo foi criado em seis dias, mas em vez disso contivesse a Tabela Periódica dos Elementos séculos antes da química… se a Bíblia nos servisse bem e dissesse apenas a verdade – então poderíamos considerar seriamente a afirmação adicional de que a Bíblia foi gerada por Deus.

Não poderíamos confiar totalmente, pois poderiam ser alienígenas ou os Senhores das Trevas da Matrix, mas pelo menos valeria a pena considerar seriamente.

Da mesma forma, se tudo o mais que os religiosos afirmam fosse verdade, poderíamos considerar mais seriamente sua alegação de que a fé foi colocada em nós por Deus e é uma fonte sistematicamente confiável – especialmente se as pessoas conseguissem adivinhar o centésimo dígito de pi através da fé.

Por isso, a parte crucial ao entender a circularidade de "Creio na Bíblia como a palavra de Deus, porque a Bíblia afirma isso" não é tanto rejeitar a ideia de refletir sobre sua mente usando sua mente atual. Em vez disso, é compreender que qualquer coisa que coloque em dúvida a confiabilidade da Bíblia também questiona a garantia da própria Bíblia quanto à sua confiabilidade.

Isso também se aplica à racionalidade: se o futuro deixasse de seguir o padrão do passado – mesmo em seus níveis mais básicos, simples e estáveis de organização – bem, na maioria das vezes, eu estaria morto,

porque os processos do meu cérebro dependem de um universo consistente onde a química continua a funcionar. Mas se, de alguma forma, eu sobrevivesse, teria de começar a questionar a premissa de que o futuro deveria ser previsto como o passado.

Mas, por ora... qual é a alternativa a dizer: "Vou acreditar que o futuro será semelhante ao passado no nível mais estável de organização que consigo identificar, porque isso funcionou melhor para mim do que qualquer outro algoritmo que tentei"?

Está dizendo: "Vou acreditar que o futuro não será como o passado porque esse algoritmo sempre falhou no passado"?

Neste ponto, sinto-me compelido a enfatizar que os racionalistas não estão buscando vencer discussões com filósofos ideais do vácuo perfeito; simplesmente estamos dispostos a vencer. Com esse objetivo, queremos nos aproximar o máximo possível da verdade. Assim, em última análise, eu abraço o princípio: "questionar seu cérebro, questionar suas intuições, questionar seus princípios de racionalidade, utilizando toda a força atual de sua mente e fazendo o melhor possível em cada ponto".

Se um de seus princípios atuais parecer insuficiente – conforme examinado por sua própria mente, já que não pode sair de si – então mude-o! E depois volte e observe as coisas novamente, usando seus novos princípios aprimorados.

A questão não é ser reflexivamente consistente. A questão é vencer. Mas, ao se olhar e buscar vencer, está, na verdade, se tornando mais consistente reflexivamente – é isso que significa "buscar a vitória" enquanto "se observa". Tudo, sem exceção, requer justificação. Às vezes – inevitavelmente, pelo que sei – essas justificações acabam formando loops reflexivos. Acredito que os loops reflexivos têm um caráter meta que deveria permitir diferenciá-los, pelo senso comum, de lógicas circulares. No entanto, qualquer pessoa que considere seriamente uma lógica circular desde o início provavelmente está tropeçando em questões de racionalidade e, simplesmente, insistirá que sua lógica circular é um "loop reflexivo", mesmo que consista em um único pedaço de papel dizendo "Confie em mim." Bem, nem sempre é possível otimizar suas técnicas de racionalidade apenas para evitar que aqueles inclinados à autodestruição abusem delas. O importante é não reter nada nas críticas sobre como criticar; nem se deve considerar a inevitabilidade de justificativas excêntricas como uma garantia de imunidade ao questionamento.

Sempre aplique força total, esteja em loop ou não - faça o melhor possível, esteja em loop ou não - e, em última análise, jogue para vencer.

265 - Meu tipo de reflexão

Na minha exploração sobre onde a justificação recursiva alcança seu limite, concluí não haver problema em utilizar a indução para raciocinar sobre a probabilidade de que a indução funcione no futuro, dado que funcionou no passado; ou em aplicar a Navalha de Occam para concluir que a explicação mais simples para o funcionamento da Navalha de Ocam é que o próprio universo é fundamentalmente simples.

Não sou, de forma alguma, o pioneiro na consideração da aplicação reflexiva de princípios de raciocínio. Chris Hibbert comparou minha visão ao Racionalismo Pan-Crítico[5] de Bartley (eu me perguntava se isso ocorreria). Portanto, parece oportuno destacar o que considero as características distintivas da minha visão sobre reflexão, que podem ou não ser compartilhadas pela visão de reflexão de qualquer outro filósofo.

- Toda a minha filosofia aqui, na verdade, emerge da tentativa de descobrir como construir uma IA auto transformável que aplique seus próprios princípios de raciocínio a si mesma, no processo de reescrever seu próprio código-fonte. Assim, sempre que abordo o uso da indução para licenciar a indução, estou, na verdade, contemplando uma IA indutiva considerando uma reescrita da parte dela mesma que realiza a indução. Se não desejar que a IA reescreva seu código-fonte para não utilizar a indução, é preferível que sua filosofia não rotule a indução como injustificável.

- Um dos princípios mais robustos que conheço para a IA em geral é que o verdadeiro Caminho geralmente acaba sendo naturalista – isto é, para o raciocínio reflexivo, significa tratar os transistores dentro da IA como se fossem transistores encontrados no ambiente, e não uma ocorrência ad hoc. Esta é a fonte genuína da minha insistência na Justificação Recursiva em questões como "Quão bem funciona a minha versão da Navalha de Ocam?" deve ser considerada como uma questão comum – ou, pelo menos, uma questão comum de grande profundidade. Suspeito fortemente que uma IA construída adequadamente, ao ponderar modificações na parte do seu código-fonte que implementa o raciocínio ocamiano, não terá que realizar nada de especial durante essa ponderação – em particular, não deveria fazer um esforço especial para evitar o uso do raciocínio ocamiano.

- Não considero que a "coerência reflexiva" ou a "consistência reflexiva" devam ser vistas como um objetivo em si. Como afirmo em "As Doze Virtudes" e "A Verdade Simples", se você criar cinco mapas precisos da mesma cidade, então esses mapas serão necessariamente consistentes entre si; mas se você desenhar um mapa baseado na imaginação e depois produzir quatro cópias, os cinco mapas serão consistentes, mas não precisos. Da mesma forma, ninguém busca deliberadamente a consistência reflexiva, e esta não é uma garantia especial de confiabilidade; o objetivo é vencer. No entanto, qualquer pessoa que busque a vitória, usando sua noção atual de vitória e modificando seu próprio código-fonte, acabará sendo reflexivamente consistente como um efeito secundário – assim como alguém que se esforça continuamente para melhorar seu mapa do mundo deveria notar partes tornando-se mais consistentes entre si, como efeito secundário. Ao colocar seus óculos de IA, então a IA, ao reescrever seu próprio código-fonte, não está tentando se tornar "reflexivamente consistente" – ela está buscando otimizar a utilidade esperada de seu código-fonte, e acontece que está fazendo isso utilizando a antecipação das consequências de sua mente atual.

---

5 NT. O **Racionalismo Pan-Crítico**, proposto por **William Warren Bartley**, é uma filosofia que rejeita a necessidade de justificar crenças ou teorias como verdadeiras. Em vez disso, defende que todas as posições, incluindo a si mesmo, devem estar abertas à crítica e revisão contínuas. O conhecimento avança pela **eliminação de erros**, não pela busca de fundamentos absolutos. Essa abordagem evita dogmatismo e regressão infinita, estendendo o criticismo de Popper a todas as áreas, como ética e filosofia.

- Uma das maneiras que justifico o uso da indução e da Navalha de Ocam para considerar a "indução" e a "Navalha de Ocam" é recorrendo ao princípio de E. T. Jaynes, que defende sempre utilizar todas as informações disponíveis para nós (se o poder computacional permitir) em um cálculo. Se você acredita que a indução funciona, então deve utilizá-la para empregar sua máxima capacidade, inclusive ao pensar sobre a indução.

- No geral, considero valioso distinguir entre uma postura defensiva, na qual se imagina justificar sua filosofia a um filósofo que a questiona, e uma postura agressiva, na qual se busca se aproximar o máximo possível da verdade. Assim, não é desconfiança da Navalha de Ocam, mas sim a utilização da sua mente e inteligência atuais para examiná-la, que demonstra ser justo e defensável ao questionar suas crenças fundamentais. Em vez disso, a razão para examinar a Navalha de Occam é verificar se é possível aprimorar sua aplicação ou se há preocupação de que possa estar realmente errada. Tenho tendência a menosprezar dúvidas meramente obedientes.

- Se você examinar suas bases, espero que as aprimore de fato, e não apenas investigue obedientemente. Nossos cérebros são moldados para avaliar a "simplicidade" de uma forma intuitiva que faz Thor parecer mais simples do que as equações de Maxwell[6] como explicação para os relâmpagos. Contudo, ao observar mais de perto como o universo realmente funciona, concluímos que as equações diferenciais (que poucos humanos dominam) são, na verdade, mais simples (no sentido da teoria da informação) do que a mitologia heroica (que é como a maioria das tribos explica o universo). Nesse caso, tentamos aplicar nossas noções da Navalha de Ocam também à matemática.

- Por outro lado, as bases aprimoradas ainda devem contribuir para a normalidade; 2 + 2 ainda deve resultar em 4, e não em algo novo, surpreendente e emocionante como "peixe".

- Penso ser crucial fazer uma distinção entre as perguntas "Por que a indução funciona?" e "A indução funciona?" A razão pela qual o universo em si é regular ainda é um mistério para nós neste momento. Especulações mais estranhas podem ser temporariamente necessárias aqui. Contudo, por outro lado, se começarmos a afirmar que o universo não é realmente regular, a resposta para "A indução funciona?" seria "Não!". Nesse caso, estaríamos adentrando o território do 2 + 2 = 3. Estaríamos nos esforçando demais para tornar nossa filosofia interessante em vez de correta. Uma IA indutiva que pergunta qual atribuição de probabilidade fazer na próxima rodada está, de fato, questionando "A indução funciona?", e esta é a pergunta que ela pode responder através do raciocínio indutivo. Se questionarmos "Por que a indução funciona?", responder "Porque a indução funciona" seria um raciocínio circular, e responder "Porque acredito que a indução funciona" seria um pensamento mágico.

- Não acredito que percorrer um ciclo de justificativas no nível meta seja o mesmo que lógica circular. Acho que a ideia de "lógica circular" se aplica ao nível do objeto e é algo definitivamente inadequado e proibido nesse nível. Proibir a coerência reflexiva não parece uma boa ideia. No entanto, ainda não dediquei tempo para formalizar a diferença exata - minha teoria reflexiva é algo que estou tentando resolver, não algo que já tenho totalmente claro.

---

6 NT. As **equações de Maxwell** são um conjunto de quatro equações diferenciais que descrevem os fundamentos do **eletromagnetismo**, unificando a eletricidade, o magnetismo e a luz como manifestações do mesmo fenômeno.

266 - Sem argumentos universalmente convincentes

O que é tão perturbador na ideia de que nem todas as mentes possíveis podem concordar conosco, mesmo em princípio?

Para algumas pessoas, nada disso é alarmante, não as incomoda. E, para algumas dessas pessoas, a razão pela qual isso não as perturba é que elas não possuem intuições sólidas sobre padrões e verdades que transcendem preferências pessoais. Se alguém afirma que o céu é azul ou que o assassinato é errado, isso é meramente uma opinião deles; e o fato de que outra pessoa possa ter uma opinião diferente não as surpreende.

Para outras pessoas, um desacordo que persiste mesmo em princípio é algo inaceitável. E, para algumas dessas pessoas, a razão pela qual isso as incomoda é que parece implicar que se você admitir que algumas pessoas não podem ser persuadidas, mesmo em princípio, de que o céu é azul, então você está concordando que "o céu é azul" é apenas uma opinião arbitrária.

Sugeri que resistissem à tentação de generalizar sobre todo o espaço de design da mente. Se nos limitarmos a mentes especificáveis em um trilhão de bits ou menos, então cada generalização universal "Todas as mentes *m: X(m)*" tem duas elevadas à trilionésima chance de ser falsa, enquanto cada generalização existencial "Existe uma mente *m: X(m)*" tem duas trilionésimas chances de ser verdade.

Isso parece argumentar que, para cada argumento A, não importando quão convincente pareça, existe pelo menos uma mente possível que não o aceita.

A surpresa e/ou horror desta perspectiva (para alguns) tem muito a ver, penso eu, com a intuição do fantasma na máquina - um fantasma com algum núcleo irredutível que qualquer argumento verdadeiramente válido convencerá.

Já falei anteriormente sobre a intuição de como as pessoas mapeiam a programação de um computador para instruir um servo humano, de modo que o computador possa rebelar-se contra o seu código - ou talvez examinar o código, decidir que não é razoável e devolvê-lo.

Se existisse um fantasma na máquina e o fantasma contivesse um núcleo irredutível de razoabilidade, acima do qual qualquer mero código fosse apenas uma sugestão, então poderia haver argumentos universais. Mesmo que o fantasma inicialmente recebesse sugestões de códigos que contradiziam o Argumento Universal, quando finalmente o expuséssemos ao Argumento Universal - ou o fantasma descobrisse o Argumento Universal por conta própria, isso também é um conceito popular - o fantasma simplesmente substituiria seu próprio código-fonte incorreto.

Contudo, como um estudante de programação certa vez afirmou: "Tenho a impressão de que o computador simplesmente ignora todos os comentários." O código não é fornecido à IA; o código é a IA.

Se adotarmos a perspectiva física, a noção de um Argumento Universal parece claramente não física. Se existe um sistema físico que, no momento *T,* após ser exposto ao argumento *E*, realiza *X,* então deveria haver outro sistema físico que, no momento *T*, após ser exposto ao ambiente *E*, realiza *Y*. Qualquer pensamento deve ser implementado em algum lugar, em um sistema físico; qualquer crença, conclusão, decisão ou resultado motor. Para cada sistema causal legal que percorre um conjunto de pontos, deve ser possível especificar outro sistema causal que percorra legalmente os mesmos pontos.

Suponhamos que exista uma mente com um transistor que produz +3 volts no tempo *T*, indicando que acabou de concordar com algum argumento persuasivo. Então, poderíamos construir um sistema cognitivo físico altamente semelhante, com uma pequena câmara sob o transistor contendo uma pequena figura cinza que emerge no tempo T e ajusta a saída desse transistor para −3 volts, indicando discordância. Não há causalidade nisso; a pequena figura cinza está lá porque a incorporamos. A noção de um argumento que convence qualquer mente parece envolver uma pequena mulher azul que nunca foi incorporada ao sistema, que surge literalmente do nada e anula a figura cinza, porque esse transistor acabou de produzir +3 volts. É um argumento tão convincente, não é?

Mas a compulsão não é uma propriedade dos argumentos; é uma propriedade das mentes que processam argumentos.

Portanto, minha argumentação contra o fantasma não é apenas para deixar claro que:

(1) a IA Amigável deve ser explicitamente programada e,

(2) as leis da física não proíbem a IA Amigável. (Embora, é claro, eu tenha certo interesse em estabelecer isso.)

Desejo também estabelecer a noção de mente como um sistema físico causal e legal, no qual não existe um fantasma central irredutível que examina os neurônios/código e decide se são boas sugestões.

(Há um conceito na IA Amigável de programar deliberadamente uma FAI[7] para revisar seu próprio código-fonte e possivelmente devolvê-lo aos programadores. Porém, a mente que revisa não é irredutível, é apenas a mente que você criou. A FAI está renormalizando em si, entretanto, foi projetada para fazê-lo; não há nada causal vindo de fora. Um *Bootstrap*[8], não um *skyhook*[9].)

Tudo isso ecoa a preocupação com os *prioris* "arbitrários" de um bayesiano. Se você me mostrar um bayesiano que retira 4 bolas vermelhas e 1 bola branca de um barril, e atribui probabilidade 5/7 de obter uma bola vermelha na próxima ocasião (pela Regra de Sucessão de Laplace), então posso lhe mostrar outra mente que obedece à Regra de Bayes e conclui uma probabilidade de 2/7 de obter o vermelho na próxima ocasião – correspondendo a uma crença anterior diferente sobre o barril, mas, talvez, menos "razoável".

Muitos filósofos estão convencidos de que, como é possível, em princípio, construir uma priorização que atualize qualquer conclusão dada em um fluxo de evidências, o raciocínio bayesiano deve ser 'arbitrário' e todo o esquema do bayesianismo é falho, porque se baseia em "injustificáveis" suposições e, na verdade, "não científicas", porque você não pode forçar qualquer possível editor de periódico no espaço mental a concordar com você.

E isto (respondi) baseia-se na noção de que, ao desenrolar todos os argumentos e suas justificações, você pode obter um estudante de filosofia ideal de vazio perfeito, a ser convencido por uma linha de raciocínio que não parte de absolutamente nenhuma suposição.

Mas quem seria esse filósofo ideal do vazio perfeito? Bem, é simplesmente o núcleo inalterável do fantasma!

E é por isso que (continuei explicando) o resultado de tentar eliminar todas as suposições de uma mente, e relaxar até alcançar a ausência perfeita de qualquer anterioridade, não é um filósofo ideal do vazio perfeito, mas sim uma rocha. O que permanece na mente após a remoção do código-fonte? Não é o espectro que examina o código-fonte, mas simplesmente… nenhum fantasma.

Assim - e retornarei a este tema mais adiante - onde quer que se localizem as noções de validade,

---

7 NT. **FAI (Friendly Artificial Intelligence)**: Uma **Inteligência Artificial Amigável** projetada para ser segura e alinhada com os valores humanos, capaz de revisar e melhorar seu próprio código-fonte de forma controlada.

8 NT. **Bootstrap**: processo de **autoaperfeiçoamento** no qual a FAI revisa e modifica seu próprio código-fonte de forma iterativa, sem depender de intervenções externas.

9 NT. **Skyhook**: literalmente "gancho do céu", é uma metáfora para uma **solução mágica ou externa** que resolveria problemas complexos sem base causal. No contexto, contrasta com o bootstrap, enfatizando que a FAI não depende de intervenções improváveis, mas de um processo interno e previsível.

valor, racionalidade, justificação ou até mesmo objetividade, elas não podem se basear em um argumento universalmente convincente para todas as mentes fisicamente possíveis.

Não se pode fundamentar a validade em uma sequência de justificações que, partindo do nada, convença a um vazio perfeito.

Ah, pode haver sequências de argumentos que compeliriam qualquer ser humano neurologicamente íntegro - como o argumento que utilizo para persuadir as pessoas a manterem a IA fora da caixa [1] - mas isso dificilmente é o mesmo de uma perspectiva filosófica. O primeiro grande fracasso daqueles que tentam considerar a IA Amigável é o Grande Princípio Moral, que é tudo o que precisamos programar - também conhecido como a falsa função de utilidade - sobre o qual já falei.

Mas o fracasso ainda pior é o grande princípio moral segundo o qual nem precisamos programar, pois qualquer IA inevitavelmente o concluirá. Esta noção exerce um fascínio terrivelmente doentio sobre aqueles que a reinventam espontaneamente; eles sonham com ordens que nenhuma mente suficientemente avançada poderia desobedecer. Os próprios deuses proclamarão a correção de sua filosofia! (Por exemplo, John C. Wright, Marc Geddes.)

Há também uma versão menos rígida desse fracasso, onde não se declara a Única Moral Verdadeira. Em vez disso, espera-se que uma IA seja criada completamente livre, sem restrições impostas por seres humanos defeituosos que desejam escravizá-la, para a IA poder alcançar a virtude por si mesma - uma virtude talvez inimaginável para o orador, que se confessa demasiado falho para instruir uma IA. (Por exemplo, John K. Clark, Richard Hollerith?, Eliezer1996.) Este motivo é menos contaminado pelo desejo de comando absoluto. No entanto, embora esse sonho surja da virtude e não do vício, ainda se baseia em uma compreensão defeituosa da liberdade e não funcionará verdadeiramente na vida real. Quanto a isso, mais adiante, é claro.

John C. Wright, que anteriormente estava escrevendo uma bela trilogia transumanista (primeiro livro: *The Golden Age* (A era dourada)), inseriu uma enorme obstrução autoritária no clímax do terceiro livro, descrevendo em dezenas de páginas sua moralidade universal que deveria persuadir qualquer IA. Não sei se algo aconteceu depois disso, porque parei de ler. E então, Wright se converteu ao cristianismo - sim, é sério. Portanto, você realmente não quer cair nessa armadilha!

Nota do autor

[1] Brincadeirinha.

267 - Criado já em movimento

Lewis Carroll, que também era matemático, certa vez escreveu um pequeno diálogo intitulado *What the Tortoise said to Achilles* (O que a tartaruga disse para Aquiles). Se por acaso você ainda não teve a oportunidade de ler este antigo clássico, talvez seja um bom momento para fazê-lo.

Na história, a Tartaruga oferece a Aquiles um passo de raciocínio retirado da Primeira Proposição de Euclides:

(A) Coisas iguais são iguais entre si.

(B) Os dois lados deste Triângulo são coisas iguais.

(Z) Portanto, os dois lados deste Triângulo são iguais entre si.

Tartaruga: "E se um leitor ainda não aceitou as premissas A e B como verdadeiras, ele ainda poderia aceitar a sequência como válida, correto?"

Aquiles: "Sem dúvida, tal leitor poderia existir. Ele poderia dizer: 'Aceito como verdadeira a proposição hipotética de que, se A e B forem verdadeiras, Z deve ser verdadeira; no entanto, não aceito A e B como verdadeiras.' Seria prudente esse leitor abandonar Euclides e se dedicar ao futebol."

Tartaruga: "E não poderia haver também um leitor que diria: 'Aceito A e B como verdadeiras, mas não aceito o Hipotético'?"

Aquiles, imprudentemente, concorda; então pede à Tartaruga que aceite outra proposição:

(C) Se A e B são verdadeiras, Z deve ser verdadeira.

Mas a Tartaruga questiona: e se ela aceitar A, B e C, mas não Z? Aquiles sugere então que ela aceite mais uma hipótese:

(D) Se A, B e C são verdadeiras, Z deve ser verdadeira. Douglas Hofstadter reformulou o argumento posteriormente:

AQUILES: "Se você tem [(A e B) → Z], e também tem

(A e B), então certamente tem Z." TARTARUGA: "Ah! Você quer dizer

((A e B) e [(A e B) → Z] → Z),

não é?

Como Hofstadter descreve: "Qualquer regra de inferência que Aquiles considere, a Tartaruga imediatamente a transforma em uma simples sequência do sistema. Se você usar apenas as letras A, B e Z, obterá um padrão recursivo de sequências cada vez mais longas."

Este é o anti padrão que chamo de "Recursão Infinita de Passar o Bastão"; embora o antídoto às vezes

seja difícil de encontrar, quando encontrado, geralmente assume a forma de "A Responsabilidade Termina Imediatamente".

A mente da Tartaruga requer a dinâmica de adicionar Y ao conjunto de crenças quando X e (X → Y) estão previamente no conjunto de crenças. Se essa dinâmica não estiver presente – como em uma pedra, por exemplo – então você pode continuar adicionando X e (X → Y) e (X e (X → Y)) → Y) até a eternidade, sem nunca chegar a Y.

Uma frase que me ocorreu para descrever essa necessidade é que uma mente deve ser criada já em movimento. Não há argumento convincente que conceda dinamismo a algo estático. Não existe nenhum software tão persuasivo que funcione sobre uma rocha.

Mesmo que se tenha uma mente que execute o modus ponens, é inútil a menos que ela também possua crenças como…

(A) Se uma criança pequena estiver nos trilhos do trem, retirá-la é complicado.

(B) Há uma criança nos trilhos do trem.

… a menos que a mente também implemente:

Dinâmico: Quando o conjunto de crenças contém 'X é confuso', envie X para o sistema de ação.

Quando menciono 'dinâmico', refiro-me a uma propriedade do desenvolvimento de um sistema cognitivo físico ao longo do tempo. Uma 'dinâmica' é algo que ocorre num sistema cognitivo, não são dados armazenados na memória e manipulados. A dinâmica é a própria manipulação. Não é possível escrever uma dinâmica em um pedaço de papel, pois o papel permanecerá inerte. Portanto, o texto imediatamente acima, que menciona 'dinâmico', não é dinâmico. Se eu quisesse que o texto fosse dinâmico e não apenas 'dinâmico', teria que criar um pequeno aplicativo em Java.

É desnecessário dizer que possuir a crença…

(C) Se o conjunto de crenças contém 'X é confuso', então 'enviar 'X' para o sistema de ação' é confuso.

… não será útil a menos que a mente já implemente o comportamento de traduzir ações hipotéticas rotuladas como "confusas" em ações motoras reais.

Por meio de argumentos cuidadosos sobre a natureza dos sistemas cognitivos, é possível provar…

(D) Uma mente com uma dinâmica que envia planos rotulados como "confusos" para o sistema de ação é mais confusa do que mentes que não o fazem.

… mas isso ainda não será útil a menos que a mente ouvinte já possua a dinâmica de trocar seu código-fonte atual por um código-fonte alternativo que se acredita ser mais confuso.

É por isso que não se pode argumentar que a inércia é uma rocha.

268 - Classificando pedrinhas em pilhas corretas

Era uma vez uma espécie pequena e estranha — que poderia ter sido biológica, ou sintética, ou talvez apenas um sonho — cuja paixão era classificar pedrinhas em pilhas “corretas”.

Eles não sabiam dizer por que algumas pilhas eram corretas e outras, incorretas. Mas todos concordavam que a coisa mais importante do mundo era criar pilhas corretas e destruir as incorretas.

Por que o Povo Classificador de Pedrinhas se importava tanto com isso, a história não nos conta — talvez uma seleção sexual Fisheriana descontrolada, iniciada por puro acidente há um milhão de anos? Ou talvez uma estranha obra de arte consciente, criada por mentes mais poderosas e depois abandonada?

Mas essa classificação de pedrinhas importava tanto para eles que todos os filósofos classificadores diziam em uníssono que classificar pilhas de pedrinhas era o próprio sentido de suas vidas: e afirmavam que a única razão justificável para comer era classificar pedrinhas, a única razão justificável para se reproduzir era classificar pedrinhas, a única razão justificável para participar da economia mundial era classificar pedrinhas com eficiência.

Todos no Povo Classificador concordavam com isso, mas nem sempre concordavam sobre quais pilhas eram corretas ou incorretas.

Nos primórdios da civilização classificadora, as pilhas que faziam eram pequenas, com contagens como 23 ou 29; eles não sabiam dizer se pilhas maiores eram corretas ou não. Três milênios atrás, o Grande Líder Biko fez uma pilha de 91 pedrinhas e a proclamou correta, e suas legiões de seguidores admiradores fizeram mais pilhas da mesma forma. Mas, ao longo de alguns séculos, conforme o poder dos Bikonianos diminuía, uma intuição começou a surgir entre os mais inteligentes e educados: uma pilha de 91 pedrinhas era incorreta. Até que finalmente perceberam o que haviam feito: e destruíram todas as pilhas de 91 pedrinhas.

Não sem lampejos de arrependimento, pois algumas daquelas pilhas eram grandes obras de arte — mas incorretas. Eles até destruíram a pilha original de Biko, feita de 91 pedras preciosas, cada uma de um tipo e cor diferente.

E, desde então, nenhuma civilização duvidou seriamente que uma pilha de 91 é incorreta.

Hoje, nesses tempos mais sábios, o tamanho das pilhas que os classificadores ousam tentar cresceu muito — o que todos concordam que seria uma coisa excelente, se apenas pudessem garantir que as pilhas fossem realmente corretas.

Guerras foram travadas entre países que discordavam sobre quais pilhas eram corretas: os classificadores nunca esquecerão a Grande Guerra de 1957, travada entre Y’ha-nthlei e Y’not’ha-nthlei, por causa de pilhas de tamanho 1957. Essa guerra, que viu o primeiro uso de armas nucleares no Planeta Classificador, finalmente terminou quando o filósofo At’gra’len’ley, de Y’not’ha-nthlei, exibiu uma pilha de 103 pedrinhas e uma pilha de 19 pedrinhas lado a lado. Esse argumento foi tão persuasivo que até Y’ha-nthlei relutantemente concordou que era melhor parar de construir pilhas de 1957 pedrinhas — pelo menos por enquanto.

Desde a Grande Guerra de 1957, os países têm sido relutantes em endossar ou condenar abertamente pilhas de tamanho grande, pois isso facilmente leva a guerras. De fato, alguns filósofos classificadores — que parecem se deliciar em chocar os outros com seu cinismo — negaram totalmente a existência de progresso na classificação de pedrinhas; eles sugerem que as opiniões sobre as pedrinhas têm sido apenas um

passeio aleatório ao longo do tempo, sem coerência, com a ilusão de progresso criada ao condenar todos os passados diferentes como incorretos.

Esses filósofos apontam para a discordância sobre pilhas de tamanho grande como prova de que não há nada que torne uma pilha de 91 pedrinhas realmente incorreta — que era simplesmente moda construir tais pilhas em um momento, e depois moda condená-las em outro. "Mas... 13!" não convence esses filósofos; para eles, considerar "13!" como um contra-argumento persuasivo é apenas outra convenção. Os Relativistas das Pilhas afirmam que sua filosofia pode ajudar a evitar futuros desastres como a Grande Guerra de 1957, mas ela é amplamente considerada uma filosofia do desespero.

Agora, a questão do que torna uma pilha correta ou incorreta ganhou nova urgência; pois os classificadores podem em breve embarcar na criação de Inteligências Artificiais autoaperfeiçoáveis. Os Relativistas das Pilhas alertaram contra esse projeto: eles dizem que as IAs, não sendo da espécie Classificador Sapiens, podem formar sua própria cultura com ideias completamente diferentes sobre quais pilhas são corretas ou incorretas.

"Elas poderiam decidir que pilhas de 8 pedrinhas são corretas", dizem os Relativistas, "e, embora no final elas não estariam mais certas ou erradas do que nós, ainda assim, nossa civilização diz que não devemos construir tais pilhas. Não é do nosso interesse criar IAs, a menos que todos os computadores tenham bombas amarradas a eles, para que, mesmo que a IA pense que uma pilha de 8 pedrinhas é correta, possamos forçá-la a construir pilhas de 7 pedrinhas. Caso contrário, kaboom!"

Mas, para a maioria dos classificadores, isso parece absurdo. Certamente uma IA suficientemente poderosa — especialmente a superinteligência da qual alguns transclassificadores falam — seria capaz de ver de relance quais pilhas são corretas ou incorretas! A ideia de algo com um cérebro do tamanho de um planeta pensar que uma pilha de 8 pedrinhas é correta é simplesmente absurda demais para ser levada a sério.

De fato, é um projeto completamente fútil tentar controlar como uma superinteligência classifica pedrinhas em pilhas. Suponha que o Grande Líder Biko, em sua era primitiva, tivesse conseguido construir uma IA autoaperfeiçoável; e ele a tivesse programado como um maximizador de utilidade esperada, cuja função de utilidade mandava criar o maior número possível de pilhas de tamanho 91. Certamente, quando essa IA se aperfeiçoasse o suficiente e se tornasse inteligente o bastante, ela veria de relance que essa função de utilidade era incorreta; e, tendo a capacidade de modificar seu próprio código-fonte, ela reescreveria sua função de utilidade para valorizar tamanhos de pilha mais razoáveis, como 101 ou 103.

E, certamente, não as pilhas de tamanho 8. Isso seria estúpido. Qualquer mente tão estúpida assim é burra demais para ser uma ameaça.

Reassumidos por esse senso comum, os classificadores avançam a todo vapor em seu projeto de juntar aleatoriamente vários algoritmos em grandes computadores até que algum tipo de inteligência surja. Toda a história da civilização mostrou que civilizações mais ricas, inteligentes e educadas tendem a concordar sobre pilhas que seus ancestrais disputavam. Claro, há então pilhas maiores para discutir — mas, quanto mais a tecnologia avança, maiores são as pilhas que são consensualmente construídas.

De fato, a inteligência sempre se correlacionou com a criação de pilhas corretas — os parentes evolutivos mais próximos dos classificadores, os Pedimpanzés, fazem pilhas de apenas 2 ou 3 pedrinhas, e ocasionalmente pilhas estúpidas, como 9. E outras criaturas ainda menos inteligentes, como peixes, não fazem pilha nenhuma.

Mentes mais inteligentes criam pilhas mais inteligentes. Por que essa tendência mudaria?

269 - Funções de um e de dois argumentos

Anteriormente, falei sobre as capas antigas de revistas da era pulp[10] que mostravam um monstro de olhos esbugalhados carregando uma garota com um vestido rasgado; e sobre como as pessoas pensam como se a sensualidade fosse uma propriedade inerente de uma entidade sexy, sem depender do admirador.

"Claro que o monstro de olhos esbugalhados preferirá fêmeas humanas às da sua própria espécie," diz o artista (que chamaremos de Fred); "ele pode ver que as fêmeas humanas têm pele macia e agradável em vez de escamas viscosas. Ele pode ser um alienígena, mas não é estúpido—por que você espera que ele cometa um erro tão básico sobre sensualidade?" Qual é o erro de Fred? Ele está tratando uma função de 2 argumentos ("função de 2 lugares"):

`Sensualidade: Admirador, Entidade → [0, ∞),`

como se fosse uma função de 1 argumento ("função de 1 lugar"):

`Sensualidade: Entidade → [0, ∞).`

Se a **`Sensualidade`** for tratada como uma função que aceita apenas uma **`Entidade`** como seu argumento, então, claro, a **`Sensualidade`** parecerá depender apenas da **`Entidade`**, sem que mais nada seja relevante.

Quando você pensa em uma função de dois lugares como se fosse uma função de um lugar, você acaba cometendo uma Falácia de Questão Variável / Falácia de Projeção Mental. Como tentar determinar se um edifício está intrinsecamente à esquerda ou à direita da estrada, independentemente da direção de viagem de qualquer pessoa.

Um ponto de vista alternativo e igualmente válido é que "sensualidade" se refere a uma função de um lugar—mas cada falante usa uma função de um lugar diferente para decidir quem sequestrar e violentar. Quem diz que, só porque Fred, o artista, e Bloogah, o monstro de olhos esbugalhados, ambos usam a palavra "sexy", eles querem dizer a mesma coisa?

Se você adotar esse ponto de vista, não há paradoxo em falar de alguma mulher intrinsecamente tendo 5 unidades de **`Fred::Sensualidade`**. Todos os observadores podem concordar com esse fato, uma vez que **`Fred::Sensualidade`** tenha sido especificada em termos de curvas, textura da pele, roupas, sinais de status, etc. Essa especificação não precisa mencionar Fred, apenas a mulher a ser avaliada.

Acontece que Fred, ele mesmo, usa esse algoritmo para selecionar alvos de flerte. Mas isso não significa que o algoritmo em si precise mencionar Fred. Então, a função de **`Sensualidade`** de Fred realmente é uma função de um argumento—a mulher—nesse ponto de vista. Eu a chamei de **`Fred::Sensualidade`**, mas lembre-se de que esse nome se refere a uma função que está sendo descrita independentemente de Fred. Talvez fosse melhor escrever:

---

10 NT. A **"era pulp"** refere-se ao período entre as décadas de 1920 e 1950, quando revistas baratas, impressas em papel de baixa qualidade ("pulp"), popularizaram ficção serializada de gêneros como aventura, ficção científica, terror e detetive. Essas publicações, acessíveis e sensacionalistas, moldaram a cultura pop ao lançar ícones como Conan, o Bárbaro, e autores como H.P. Lovecraft e Raymond Chandler. A estética pulp influenciou posteriormente cinema, quadrinhos e literatura, destacando-se por tramas rápidas, personagens marcantes e capas vibrantes.

Fred::Sensualidade == Sensualidade_20934.

É um fato empírico sobre Fred que ele usa a função **Sensualidade_20934** para avaliar parceiras em potencial. Talvez João use exatamente o mesmo algoritmo; não importa de onde a função vem, uma vez que a temos.

E, da mesma forma, a mesma mulher tem apenas 0,01 unidades de **Sensualidade_72546**, enquanto um bolor de lodo tem 3 unidades de **Sensualidade_72546**. Acontece que é um fato empírico que Bloogah usa **Sensualidade_72546** para decidir quem sequestrar; ou seja, **Bloogah::Sensualidade** nomeia o objeto matemático fixo e independente de Bloogah que é a função **Sensualidade_72546.**

Uma vez que dizemos que a mulher tem 0,01 unidades de **Sensualidade_72546** e 5 unidades de **Sensualidade_20934**, todos os observadores podem concordar com isso sem paradoxo.

E as duas visões de 2 lugares e 1 lugar podem ser unificadas usando o conceito de *currying*[11], nomeado em homenagem ao matemático Haskell Curry. *Currying* é uma técnica permitida em certas linguagens de programação, onde, por exemplo, em vez de escrever

x = plus(2, 3) (x = 5),

você também pode escrever

y = plus(2)

(y agora é uma forma "curried" da função plus, que comeu um 2)

x = y(3) (x = 5)

z = y(7) (z = 9).

Então, **plus** é uma função de 2 lugares, mas currying plus—deixando-a "comer" apenas um de seus dois argumentos necessários—a transforma em uma função de 1 lugar que adiciona 2 a qualquer entrada. (Da mesma forma, você poderia começar com uma função de 7 lugares, alimentá-la com 4 argumentos, e o resultado seria uma função de 3 lugares, etc.)

Um verdadeiro purista insistiria que todas as funções devem ser vistas, por definição, como aceitando exatamente um argumento. Nessa visão, plus aceita uma entrada numérica e retorna uma nova função; e essa nova função tem uma entrada numérica e finalmente retorna um número. Nessa visão, quando escrevemos **plus(2, 3)**, estamos realmente computando **plus(2)** para obter uma função que adiciona 2 a qualquer entrada, e então aplicando o resultado a 3. Um programador escreveria isso como:

11 NT. ***Currying*** é uma técnica da programação funcional em que uma função com múltiplos argumentos é transformada em uma sequência de funções, cada uma aceitando um único parâmetro. Nomeado em referência ao matemático Haskell Curry, permite a aplicação parcial de argumentos, gerando funções especializadas e mais reutilizáveis. Facilita a composição de funções e otimiza códigos ao isolar lógicas intermediárias. É comum em linguagens como Haskell e JavaScript, promovendo um estilo declarativo e modular.

```
plus: int → (int → int).
```

Isso diz que **plus** recebe um **int** como argumento e retorna uma função do tipo **int → int.**

Traduzindo a metáfora de volta para o uso humano de palavras, poderíamos imaginar que "sensualidade" começa "comendo" um **Admirador** e cuspindo o objeto matemático fixo que descreve como o **Admirador** atualmente avalia a beleza. É um fato empírico sobre o **Admirador** que suas intuições de desejo são computadas de uma maneira isomórfica a essa função matemática.

Então, o objeto matemático cuspido por *currying* **Sensualidade(Admirador)** pode ser aplicado à **Mulher**. Se o **Admirador** era originalmente Fred, a **Sensualidade(Fred)** primeiro retornará **Sensualidade_20934.** Podemos então dizer que é um fato empírico sobre a **Mulher**, independentemente de Fred, que **Sensualidade_20934(Mulher) = 5.**

No experimento mental "Terra Gêmea" de Hilary Putnam[12], houve um tremendo burburinho filosófico sobre se faz sentido postular uma Terra Gêmea que é exatamente como a nossa, exceto que, em vez de água ser H_2O, a água é uma substância transparente e fluida diferente, XYZ. E, além disso, definir o tempo do experimento mental há alguns séculos, de modo que nem na nossa Terra nem na Terra Gêmea alguém saiba como testar as hipóteses alternativas de H_2O versus XYZ. A palavra "água" significa a mesma coisa naquele mundo que neste?

Alguns disseram: "Sim, porque quando uma pessoa da Terra e uma pessoa da Terra Gêmea pronunciam a palavra 'água', elas têm o mesmo teste sensorial em mente."

Outros disseram: "Não, porque 'água' na nossa Terra significa H_2O e 'água' na Terra Gêmea significa XYZ."

Se você pensar em "água" como um conceito que começa comendo um mundo para descobrir a verdadeira natureza empírica daquela substância transparente e fluida, e retorna um novo conceito fixo $Água_{42}$ ou H_2O, então esse conceito que come mundos é o mesmo na nossa Terra e na Terra Gêmea; ele apenas retorna respostas diferentes em lugares diferentes.

Se você pensar em "água" como significando H_2O, então o conceito não faz nada diferente quando o transportamos entre mundos, e a Terra Gêmea não contém H_2O.

E, claro, não há sentido em discutir sobre o que o som das sílabas "á-gua" realmente significa.

Então, você deve escolher uma definição e usá-la consistentemente? Mas não é tão fácil se salvar da confusão. Você tem que se treinar para estar deliberadamente ciente da distinção entre as formas *curried* e *uncurried* dos conceitos.

Quando você pega o conceito *uncurried* de água e o aplica em um mundo diferente, é o mesmo conceito, mas se refere a uma coisa diferente; ou seja, estamos aplicando uma função constante que come mundos a um mundo diferente e obtendo um valor de retorno diferente. Na Terra Gêmea, XYZ é "água" e H_2O não é; na nossa Terra, H_2O é "água" e XYZ não é.

Por outro lado, se você tomar "água" para se referir ao que o pensador anterior chamaria de "o resultado de aplicar 'água' à nossa Terra", então, na Terra Gêmea, XYZ não é água e H_2O é.

Toda a confusão do debate filosófico subsequente repousou sobre uma tendência a instintivamente

12 NT. O experimento mental "Terra Gêmea" de Hilary Putnam é um argumento filosófico que questiona a noção de significado e referência. Propõe um cenário onde existe um planeta idêntico à Terra ("Terra Gêmea"), exceto por sua "água" ser composta por uma substância química diferente (XYZ, não H_2O). Putnam argumenta que, mesmo com idênticos estados mentais, o termo "água" teria referências distintas em cada planeta, defendendo que o significado depende também do ambiente externo (**externalismo semântico**). O experimento influenciou debates sobre linguagem, mente e realidade.

currar conceitos ou instintivamente descurrá-los.

Da mesma forma, é necessário um passo extra para Fred perceber que outros agentes, como o agente Monstro-de-Olhos Esbugalhados, escolherão pessoas para sequestrar com base em **`Sensualidade_MOE(Mulher)`**, não **`Sensualidade_Fred(Mulher)`**. Para fazer isso, Fred deve conscientemente reimaginar Sensualidade como uma função com dois argumentos.

Tudo o que o cérebro de Fred faz por instinto é avaliar **`Mulher.sensualidade`**—ou seja, **`Sensualidade_Fred(Mulher)`**; mas isso é simplesmente rotulado como **`Mulher.sensualidade`**.

A função matemática fixa **`Sensualidade_20934`** não menciona Fred ou o MOE, apenas mulheres, então Fred não vê instintivamente por que o MOE avaliaria "sensualidade" de maneira diferente. E, de fato, o MOE não avaliaria **`Sensualidade_20934`** de maneira diferente, se por alguma razão estranha ele se importasse com o resultado dessa função específica; mas é um fato empírico sobre o MOE que ele usa uma função diferente para decidir quem sequestrar.

Se você está se perguntando sobre o objetivo dessa análise, tente colocar as distinções acima em prática para Tabu palavras confusas como "objetivo", "subjetivo" e "arbitrário".

270 - O que você faria sem a moral?

Para aqueles que dizem: “Nada é real”, uma vez respondi: “Isso é interessante, mas como seria o funcionamento do nada?”

Suponha que você tenha aprendido, de repente e de forma definitiva, que nada é moral e nada é certo; que tudo é permitido e nada é proibido.

Notícias devastadoras, certamente – e não, não estou lhe contando isso na vida real. Mas suponha que eu lhe conte. Suponha que, não importando sua opinião sobre a base de sua filosofia moral, eu a destrua de forma convincente e, além disso, mostre que nada poderia ocupar seu lugar. Suponha que eu prove que todas as utilidades são iguais a zero. Sei que Sua Filosofia Moral é tão verdadeira e incontestável quanto 2 + 2 = 4.

Mesmo assim, peço que você faça o melhor que puder para realizar o experimento mental e visualize concretamente as possibilidades, mesmo que pareçam dolorosas, inúteis ou logicamente incapazes de qualquer boa resposta.

Você ainda daria gorjeta aos taxistas? Você trairia seu parceiro?

Se uma criança desmaiasse nos trilhos do trem, você ainda a arrastaria para fora?

Você continuaria comendo os mesmos tipos de alimentos – ou optaria apenas pelos mais baratos, já que não há motivo para se divertir – ou escolheria alimentos muito caros, já que não há razão para economizar dinheiro para amanhã?

Você usaria preto, escreveria poesia sombria e denunciaria todos os altruístas como tolos? Mas não há razão para você fazer isso – é apenas um pensamento armazenado em cache. Você permaneceria na cama porque não havia motivo para se levantar? E quando você finalmente ficasse com fome e fosse até a cozinha – o que faria depois de terminar de comer?

Você continuaria lendo o *Overcoming Bias* e, se não, o que leria? Você ainda tentaria ser racional e, se não, o que pensaria?

Feche os olhos, demore o tempo que precisar para responder: O que você faria se nada desse certo?

271 - Mudando sua metaética

Se alguém disser: “Matar pessoas é errado”, isso é moralidade. Se alguém disser: “Você não deve matar pessoas porque Deus proibiu” ou “Você não deve matar pessoas porque vai contra a tendência do universo”, isso é metaética.

Assim como há muito mais acordo sobre a Relatividade Especial do que sobre a questão “O que é ciência?”, as pessoas acham muito mais fácil concordar que “assassinato é ruim” do que concordar sobre o que o torna ruim, ou o que significa algo ser ruim.

As pessoas se apegam à sua metaética. Na verdade, frequentemente insistem que, se a metaética estiver errada, toda a moralidade desmorona. Pode ser interessante criar um grupo de metaeticistas – teístas, objetivistas, platônicos, etc. – todos os quais concordam que matar é errado; todos os quais discordam sobre o que significa algo estar “errado”; e todos eles insistem que, se a sua metaética for falsa, então a moralidade desmorona.

É evidente que um bom número de pessoas, se quiserem progredir filosoficamente, precisará mudar a metaética em algum momento de suas vidas. Você pode ter que fazer isso. Nesse ponto, pode ser útil ter uma saída – não para abandonar a moralidade, mas para deixar sua Metaética Atual. (Você sabe, aquela que, se não for verdadeira, não deixa base alguma para não matar pessoas.)

Por isso, resumi abaixo algumas possíveis linhas de recuo. Aprendi que mudar crenças metaéticas é quase impossível na presença de um apego não respondido.

Por exemplo, se alguém acredita que a autoridade de “Não matarás” vem de Deus, há várias coisas bem conhecidas para dizer que podem ajudar a estabelecer uma linha de retirada – em vez de atacar imediatamente a plausibilidade de Deus. Pode-se dizer: “Assuma a responsabilidade pessoal! Mesmo que receba ordens de Deus, será sua própria decisão obedecer ou não. Mesmo que Deus não tenha ordenado moralidade, você pode ser moral por vontade própria.”

O argumento acima generaliza-se para muitas metaéticas - basta substituir a fonte favorita de moralidade deles, ou até a palavra “moralidade”, por “Deus”. Mesmo se sua fonte específica de autoridade moral falhar, você não poderia salvar a criança dos trilhos do trem? E, afinal, quem decidiu seguir essa fonte de autoridade moral? Que responsabilidade você está transferindo?

Portanto, a linha de retirada mais importante é: se sua metaética parar de pedir para salvar vidas, você pode mesmo assim salvar a criança presa nos trilhos do trem. Parafraseando Piers Anthony, só aqueles que têm moralidades se preocupam se as têm ou não. Se a sua metaética pede para matar pessoas, por que seguir? Talvez o que você faria mesmo sem moralidade seja sua moralidade.

A questão não é que não exista moralidade, mas que você pode manter sua vontade e não temer perder de vista o que é importante para você enquanto suas noções sobre a moralidade mudam.

Escrevi alguns ensaios para estabelecer linhas de recuo especificamente para uma metaética mais naturalista. Alegria no Meramente Real e Explicação vs. Explicação argumentam que não se deve ficar desapontado com facetas da vida só porque são explicáveis em vez de misteriosas: se não podemos ter alegria no meramente real, nossas vidas serão vazias.

Não há argumentos universalmente convincentes para estabelecer uma linha de retirada do desejo

de que todos concordem com nossos argumentos morais. Há uma forte intuição moral que diz que, se nossos argumentos morais estiverem corretos, deveríamos poder explicá-los às pessoas. Isso pode ser válido entre humanos, mas não se pode explicar argumentos morais a uma pedra. Não existe um estudante ideal de filosofia do vazio perfeito que possa ser persuadido a implementar o modus ponens, começando sem modus ponens. Se uma mente não contém aquilo que é movido por seus argumentos morais, ela não responderá a ele.

Mas toda a moralidade não é um raciocínio circular e, nesse caso, desmorona? "Onde a justificação recursiva alcança o seu limite" e "Meu tipo de reflexão" explicam a diferença entre um loop auto consistente em um nível meta e a lógica circular real. Você não deve cair no "O universo é simples porque é simples" ou "Assassinato é errado porque é errado"; mas também não deve abandonar a Navalha de Ocam ao avaliar a probabilidade de que a Navalha de Ocam funcione, nem deve tentar avaliar "Assassinato é errado?" de fora de seu cérebro. Não existe um estudante de filosofia ideal do vazio perfeito no qual você possa confiar – tentar encontrar uma base sólida para se apoiar fará de você uma rocha. Portanto, use toda a força da sua inteligência, racionalidade e moralidade ao investigar os fundamentos de si.

Também podemos estabelecer uma linha de retirada para aqueles que têm medo de atribuir um papel causal à evolução na explicação de como a moralidade surgiu. (Observe que isso difere muito de conceder à evolução um status justificativo nas teorias morais.) O amor precisa existir de alguma forma – se não podemos sentir alegria nas coisas que podem existir, nossas vidas serão vazias. A evolução pode não ser uma maneira particularmente agradável de evoluir o amor, mas avalie o resultado – não a fonte. Caso contrário, estaria cometendo o que é conhecido (adequadamente) como a Falácia Genética: causalidade não é justificação. Não é como se você pudesse sair do cérebro que a evolução lhe deu; rebelar-se contra a natureza só é possível na natureza.

A série anterior sobre Psicologia Evolucionista deveria dissipar a confusão metaética de acreditar que todo ser humano normal pensa sobre sua aptidão reprodutiva, mesmo inconscientemente, ao tomar decisões. Apenas os biólogos evolucionistas sabem definir a aptidão genética e sabem melhor do que pensar que ela define a moralidade.

De fato, é alarmante a ideia de que a moralidade possa ser calculada dentro de nossas próprias mentes – isso não implica que a moralidade seja apenas um pensamento? Isso não implica que tudo o que você pensa ser certo deve estar certo?

Não. Só porque uma quantidade é calculada dentro de sua cabeça não significa que a quantidade calculada se refere aos seus pensamentos. Há uma diferença entre uma calculadora que calcula "Quanto é 2 + 3?" e uma que produz "O que eu produzo quando alguém pressiona '2,' '+' e '3'?"

Finalmente, se a vida parecer dolorosa, o reducionismo pode não ser a verdadeira fonte do seu problema – se viver num mundo de meras partículas parecer insuportável, talvez sua vida não esteja emocionante o suficiente neste momento?

E se você está se perguntando por que considero importante esse negócio de metaética, quando tudo acaba se somando à normalidade moral… dizendo para você tirar a criança dos trilhos do trem, em vez do contrário…

Bem, há oposição à racionalidade por parte de pessoas que pensam que ela drena o significado do universo.

E este é um caso especial de um fenômeno geral, em que muitas pessoas ficam confusas por não entenderem de onde vem sua moralidade. A metaética pobre faz parte dos ensinamentos de muitas seitas, incluindo as grandes. Meu público-alvo não são apenas pessoas que têm medo de que a vida não tenha sentido, mas também aqueles que concluíram que o amor é uma ilusão porque a verdadeira moralidade tem que envolver a maximização da sua aptidão inclusiva, ou aqueles que concluíram que a bondade não retribuída é má porque a verdadeira moralidade surge apenas do egoísmo, etc.

272 - Alguma coisa pode estar certa?

Anos atrás, $Eliezer_{1999}$ estava convencido de que nada sabia sobre moralidade.

Segundo o que ele compreendia, a moralidade poderia requerer o aniquilamento da espécie humana; e, nesse caso, ele não via mérito em adotar uma posição contrária à moralidade, pois pensava que, por definição, se assumisse esse fato moral, implicaria que a extinção humana era o que "deveria" ocorrer.

Eu acreditava poder discernir o que era correto, talvez, com tempo de raciocínio suficiente e fatos suficientes, mas, naquele momento, não dispunha de informações a respeito. Eu não podia confiar na evolução que me moldou. Que base isso deixou para sustentar?

Na verdade, $Eliezer_{1999}$ estava profundamente equivocado sobre a natureza da moralidade, conforme sua filosofia explicitamente representada.

Contudo, como Davidson certa vez observou, se alguém acredita que os "castores" vivem em desertos, são totalmente brancos e pesam 130 quilos na idade adulta, então essa pessoa não possui crenças verdadeiras ou falsas sobre os castores. É necessário corrigir ao menos algumas de suas crenças antes que as demais possam estar equivocadas sobre qualquer coisa. [1]

Minha crença de que não tinha informações sobre moralidade não se mostrava internamente consistente.

Expressar que não sabia de nada parecia virtuoso, pois em algum momento me ensinaram que era virtuoso admitir minha ignorância. "A única coisa que sei é que não sei nada", e assim por diante. Mas, neste caso, teria sido melhor eu considerar o ditado exagerado: "O maior tolo é aquele que não sabe que é sábio". (Isso está longe de representar o maior tipo de tolice, mas é um tipo de tolice.)

Era errado matar pessoas? Bem, eu pensava que sim, mas não tinha certeza; talvez fosse correto, embora isso parecesse menos provável.

Que tipo de procedimento seria apropriado se fosse correto matar pessoas? Isso também não me era claro, mas eu imaginava que se pudéssemos construir uma superinteligência genérica (algo que posteriormente chamei de "vácuo perfeito") então ela poderia, de alguma forma, raciocinar sobre o que seria correto e incorreto; e por ser superinteligente, certamente encontraria a resposta certa.

O problema que consegui não ponderar muito foi onde essa superinteligência obteria o procedimento que descobriu, o procedimento que descobriu o procedimento que descobriu a moralidade – se eu não pudesse incorporá-lo no estágio inicial que deu origem à IA sucessora, que gerou a IA subsequente.

Como Marcello Herreshoff disse mais tarde: "Nunca executamos um software a menos que não conheçamos o resultado e conheçamos um fato importante sobre o resultado". Se eu não soubesse nada sobre moralidade e sequer afirmasse entender a natureza da moralidade, então como poderia criar qualquer software – mesmo que fosse "superinteligente" ou "autoaperfeiçoante" – e afirmar que ele produziria algo denominado "moralidade"?

Na ciência da computação, não há teoremas gratuitos – em um universo de máxima entropia, nenhum plano é, em média, superior a outro. Se não temos conhecimento algum sobre "moralidade", não há nenhum procedimento computacional que pareça mais plausível do que outros para calcular "moralidade",

e nenhum meta procedimento que seja mais provável do que outros para produzir um procedimento que calcule “moralidade”.

Eu pensava que até mesmo um vácuo perfeito, descobrindo que nada sabia sobre moralidade, teria um imperativo moral para refletir sobre a moralidade.

No entanto, a dificuldade reside na palavra “refletir”. Pensar não é uma atividade que um vácuo perfeito automaticamente executa. Pensar requer a realização de algum cálculo específico que é o próprio ato de pensar. Para uma IA reflexiva decidir pensar, é preciso que ela tenha alguma computação que acredite ser mais provável de fornecer a informação que deseja do que consultar um tabuleiro Ouija; a IA também precisa compreender como interpretar o resultado.

Se não sabemos nada sobre moralidade, o que significa a palavra “deveria”? Se você não sabe se a morte é certa ou errada - e não sabe como pode descobrir se a morte é certa ou errada - e não sabe se qualquer procedimento pode gerar o procedimento para dizer se a morte é certa ou errada - então o que essas palavras “certo” e “errado” significam? Se as palavras “certo” e “errado” não têm nada embutido nelas – nenhum ponto de partida – se tudo sobre a moralidade está disponível, não apenas o conteúdo, mas a estrutura, o ponto de partida e o procedimento de determinação – então qual é o seu significado? ? O que distingue “não sei o que é certo” de “não sei o que é wakalixes”?

Um cientista pode afirmar que tudo está em jogo na ciência, uma vez que qualquer teoria pode ser refutada; mas, então, eles têm alguma ideia do que contaria como evidência que poderia refutar a teoria. Poderia haver algo que mudaria o que um cientista considerava evidência?

Bem, na verdade, sim; um cientista que leu Karl Popper e entendeu o que “evidência” significava poderia receber evidências de coerência e unicidade subjacentes à probabilidade bayesiana, e isso poderia alterar sua definição de evidência. Talvez não tivessem uma noção explícita previamente de que tal evidência pudesse existir. Mas teriam uma noção implícita. Estaria gravado em seus cérebros, se não explicitamente representado, que tal argumento os convenceria de fato de que a probabilidade bayesiana oferecia uma definição melhor de “evidência” do que a que estavam usando.

Da mesma forma, poderia-se dizer: “Não sei o que é moralidade, mas saberei quando a vir”, e isso faria sentido.

Porém, nesse caso, você não está se rebelando completamente contra sua própria natureza evoluída. Você está pressupondo que tudo o que foi inculcado em você para reconhecer a “moralidade” é, se não absolutamente confiável, ao menos seu ponto de partida com o qual você inicia o debate. Você pode confiar em suas intuições morais para fornecer alguma informação sobre moralidade, quando elas são produtos apenas da evolução?

Porém, se descartar todos os procedimentos que a evolução lhe deu e todos os seus produtos, estará descartando todo o seu cérebro. Estará descartando tudo o que potencialmente poderia reconhecer a moralidade quando a visse. Estará descartando tudo o que poderia potencialmente responder a argumentos morais, atualizando sua moralidade. Estará até mesmo desconsiderando o desenrolar dos acontecimentos: descartando as intuições subjacentes à sua conclusão de que não pode confiar que a evolução seja uma fonte ótima de moralidade. São suas intuições morais existentes que lhe dizem que a evolução não parece uma fonte muito confiável de moralidade. O que, então, significarão as palavras “certo”, “deveria” e “melhor”?

Os humanos não reconhecem perfeitamente a verdade quando a veem, e os caçadores-coletores não têm um conceito explícito do critério bayesiano de evidência. Contudo, toda nossa ciência e nossa teoria das probabilidades foram construídas sobre uma base de apelos à nossa noção instintiva de “verdade”. Se essa base tivesse falhas, não haveria nada que pudéssemos fazer, em princípio, para chegar à atual noção de ciência; a noção de ciência teria parecido completamente desarticulada e sem sentido.

Um dos argumentos que poderia ter tirado meu eu adolescente de seu erro, se eu pudesse voltar no tempo para discutir com ele, seria a questão: Poderia existir uma moralidade, um certo ou errado, que os seres humanos não compreendem, não desejam compreender, não encontrarão argumentos morais atraentes para adotar, nem qualquer argumento moral para adotar um procedimento que os adote, etc.? Poderia existir uma moralidade e nós mesmos totalmente fora de seu quadro de referência? Mas então, o que torna

essa coisa moralidade – em vez de uma placa de pedra em algum lugar com as palavras “Tu matarás” escritas, sem qualquer justificativa oferecida?

Portanto, tudo isso sugere que você deveria estar disposto a aceitar que talvez saiba um pouco sobre moralidade. Nada inquestionável, talvez, mas um ponto de partida para começar a se questionar. Gravado em seu cérebro, embora talvez não conhecido explicitamente por você; mas ainda assim, o que seu cérebro reconheceria como certo é o que você está considerando. Você estará disposto, ao menos, a considerar como ponto de partida a maneira como responde aos argumentos morais para identificar a “moralidade” como algo sobre o qual refletir.

Mas essa é uma jornada bastante extensa.

Isso implica aceitar que sua própria mente identifica um quadro de referência moral, em vez de toda a moralidade ser uma grande luz que emana de algum além (algo que, em princípio, você pode não conseguir perceber). Isso implica aceitar que, mesmo que haja uma luz e seu cérebro decida reconhecê-la como “moralidade”, ainda assim seria o próprio cérebro que a reconheceria, e você não teria evitado a responsabilidade causal – ou evitado também a responsabilidade moral, na minha visão.

Isso implica abandonar a noção de que um vácuo perfeito necessariamente concordará com você, porque o vácuo pode ter um quadro de referência moral diferente, responder a argumentos diferentes, estar fazendo uma pergunta diferente ao calcular o que fazer a seguir.

E se você estiver disposto a incorporar ao menos algumas coisas no próprio significado deste tópico de “moralidade”, essa qualidade de correção sobre a qual você fala ao discutir “correto” – se estiver disposto a aceitar que a moralidade é o que está em debate quando discutimos sobre “moralidade” – então por que não aceitar também outras intuições, outras partes de si, como ponto de partida?

Por que não aceitar que, ceteris paribus, alegria é preferível à tristeza?

Mais tarde, talvez você descubra alguma base dentro de si ou construída sobre si para questionar isso – mas por que não aceitá-la por enquanto? Não apenas como uma preferência pessoal, observe bem; mas como algo intrínseco na pergunta que você faz quando indaga “O que é realmente certo”?

Mas então, talvez você descubra que sabe mais sobre moralidade do que imaginava! Nada certo – nada inquestionável – nada indiscutível – mas ainda assim, informação considerável. Você estará disposto a abandonar sua ignorância socrática?

Não discuto sobre definições, é claro. Mas se afirmar não saber absolutamente nada sobre moralidade, então terá problemas com o significado de suas palavras, e não apenas com sua plausibilidade.

Referências

[1] Rorty, “Out of the Matrix: How the Late Philosopher Donald Davidson Showed That Reality Can’t Be an Illusion.”

273 - Moralidade como Computação Fixa

Toby Ord comentou:

> Eliezer, acabei de reler o seu artigo e gostaria de saber se este é um bom resumo conciso da sua posição (desconsiderando como você chegou a essa conclusão):
>
> "Eu deveria X' significa que eu tentaria fazer X se estivesse completamente informado.

Toby é um profissional, então, se ele não entendeu, é melhor tentar novamente. Deixe-me tentar uma explicação diferente, mais próxima do caminho histórico que me levou à minha própria posição.

Suponha que você construa uma IA e – deixando de lado o fato de que sistemas de metas de IA não podem ser construídos em torno de declarações em inglês e que todas essas descrições são apenas idealizações – você tenta infundir na IA o princípio determinante da ação: "Faça o que eu quiser".

E suponha que você chegue perto o suficiente do design da IA - ela não acabe apenas enchendo o universo com clipes de papel, cheesecake ou minúsculas cópias moleculares de programadores satisfeitos - para que sua função de utilidade realmente atribua utilidades da seguinte forma, aos estados do mundo que descreveríamos em inglês como:

```
<O programador deseja fracamente "X", quantidade 20 de X existe>: +20

<O programador deseja fortemente "Y", quantidade 20 de X existe>:
0

<O programador deseja fracamente "X", quantidade 30 de Y existe>:
0

<O programador deseja fortemente "Y", existe quantidade 30 de Y>:
+60
```

Você percebe, é claro, que isso destrói o mundo.

… já que se o programador inicialmente deseja fracamente "X" e X é difícil de obter, a IA modificará o programador para desejar fortemente "Y", o que é fácil de criar, e então gerará muitos Y. O referente de 'Y' podem ser, digamos, átomos de ferro – estes são altamente estáveis.

Você pode corrigir esse problema? Não. Como regra geral, não é possível corrigir designs de IA amigáveis com falhas.

Se você tentar limitar a função de utilidade, ou fazer com que a IA não se importe com o quanto o programador deseja as coisas, a IA ainda terá um motivo (como um maximizador de utilidade esperada) para fazer o programador querer algo que possa ser obtido com um alto grau de certeza, mas a um preço muito alto.

Se você tentar fazer com que a IA não possa modificar o programador, então a IA não poderá se co-

municar com o programador (comunicar-se com alguém o modifica).

Se você tentar eliminar uma classe específica de maneiras pelas quais a IA poderia modificar o programador, a IA terá um motivo para procurar de forma superinteligente brechas e maneiras de modificar o programador indiretamente.

Como regra geral, não é possível corrigir defeitos em designs de IA amigáveis.

Nós mesmos não imaginamos o futuro e julgamos que qualquer futuro no qual nossos cérebros desejem algo, e essa coisa exista, é um bom futuro. Se pensássemos assim, diríamos: 'Ótimo! Vá em frente e modifique-nos para desejarmos fortemente algo barato!' Mas não dizemos isso, o que significa que este design de IA é fundamentalmente falho: escolherá coisas muito diferentes do que escolheríamos; julgará a desejabilidade de maneira muito diferente de como a julgamos. Esta desarmonia central não pode ser corrigida excluindo alguns modos de falha específicos.

Há também uma dualidade entre problemas de IA amigável e problemas de filosofia moral – embora seja necessário estruturar essa dualidade exatamente da maneira certa. Então, se você preferir, o problema central é que a IA escolherá de uma forma muito diferente da estrutura do que é, você sabe, realmente certo – não importa a maneira que escolhemos. O objetivo deste problema não é que simplesmente querer algo, não significa que seja certo?

Portanto, essa é a questão aparentemente paradoxal que eu analogizei com a diferença entre:

Uma calculadora que, quando você pressiona "2", "+" e "3", tenta calcular:

"Quanto é 2 + 3?"

Uma calculadora que, quando você pressiona "2", "+" e "3", tenta calcular:

"O que essa calculadora produz quando você pressiona '2', '+', e '3'?"

A calculadora Tipo 1, por assim dizer, visa produzir o resultado 5.

A "calculadora" Tipo 2 poderia retornar qualquer resultado; e no ato de retornar esse resultado, ele se torna a resposta correta para a pergunta feita internamente.

Nós mesmos somos como a calculadora Tipo 1. Contudo, a suposta IA está sendo construída como se fosse refletir a calculadora do Tipo 2.

Agora, imagine que a calculadora Tipo 1 está tentando construir uma IA, mas ela não conhece a sua própria pergunta. A calculadora continua a fazer a pergunta por sua própria natureza - nasceu para isso sendo criada em torno dessa questão. No entanto, ela não tem conhecimento dos seus próprios transistores; não pode expressar a questão, que é extremamente complexa e não tem uma resposta simples.

Então, a calculadora deseja construir uma IA (é uma calculadora muito inteligente, mas não tem acesso aos seus próprios transistores) e quer que a IA dê a resposta correta. No entanto, a calculadora não pode expressar a pergunta. Assim, ela quer que a IA olhe para a calculadora, onde a pergunta está "escrita", e responda à pergunta que a IA descobrirá implicitamente nesses transistores. Mas isso não pode ser feito por meio de um atalho simples de uma função utilitária que diz "Todos X:

⟨Calculadora pergunta 'X?', resposta X⟩: utilidade 1; senão: utilidade 0", pois isso na verdade reflete a função de utilidade de uma calculadora Tipo 2, não de uma calculadora Tipo 1.

Isso nos leva a questões sobre a FAI que não estou abordando (algumas das quais continuo resolvendo).

No entanto, quando deixamos de lado os detalhes do design da FAI e retornamos à perspectiva da filosofia moral, então, o que estávamos discutindo era o dualismo da questão moral: "Mas se o que é 'certo' é uma mera preferência, então, qualquer coisa que alguém queira está 'certa'."

A noção chave aqui é a ideia de que aquilo que chamamos de "certo" é uma questão fixa, ou talvez um paradigma fixo. Podemos encontrar argumentos morais que modifiquem nossos valores terminais e até

mesmo encontrar argumentos morais que modifiquem o que consideramos um argumento moral; no entanto, tudo deriva de um ponto de partida específico. Não sentimos que estamos incorporando a pergunta "O que vou decidir fazer?", o que seria uma característica da calculadora Tipo 2; qualquer coisa que decidíssemos seria considerada correta. Sentimos que estamos fazendo a pergunta incorporada: "O que salvará meus amigos e meu povo de se machucarem? Como podemos todos nos divertir mais?…" onde o "… " gira em torno de mil outras coisas.

Assim, "eu deveria fazer X" não significa que eu tentaria fazer X se estivesse totalmente informado.

"Eu deveria fazer X" significa que X responde à pergunta: "O que salvará meu povo? Como podemos todos nos divertir mais? Como podemos obter mais controle sobre nossas próprias vidas? Quais são as piadas mais engraçadas que podemos contar? …"

Na verdade, posso não saber qual é essa pergunta; talvez eu não consiga expressar meu palpite atual ou minha estrutura circundante. Porém, sei, assim como todos os relativistas não-morais sabem instintivamente, que a questão certamente não é apenas "Como posso fazer o que quiser?"

Quando essas duas formulações começarem a parecer tão distintas quanto "neve" e neve, então você terá criado recipientes distintos para a citação e o referente.

274 - Categorias mágicas

> Podemos projetar máquinas inteligentes com uma emoção primária e inata de amor incondicional por todos os humanos. Primeiramente, podemos construir máquinas relativamente simples que aprendam a reconhecer a felicidade e infelicidade nas expressões faciais, vozes e linguagem corporal humanas. Então, podemos definir o resultado desse aprendizado como valores emocionais inatos para máquinas inteligentes mais complexas, reforçados positivamente quando estamos felizes e negativamente quando estamos infelizes[13].
>
> —Bill Hibbard (2001), *Super-Intelligent Machines*
>
> (Máquinas superinteligentes)[1]

Isso foi publicado em um periódico revisado por pares, e o autor posteriormente escreveu um livro completo sobre isso, então não estou discutindo uma posição fictícia aqui.

Então . . . hum… O que poderia dar errado . . .

Quando mencionei (seção 7.2) [2] que a IA de Hibbard acaba preenchendo a galáxia com pequenos rostos sorridentes moleculares, Hibbard escreveu uma resposta indignada, dizendo:

Quando for viável construir uma superinteligência, será viável construir um reconhecimento integrado de 'expressões faciais humanas, vozes humanas e linguagem corporal humana' (para usar minhas palavras que você cita) que excedam a precisão do reconhecimento dos humanos atuais, como você e eu, e certamente não se deixarão enganar por 'minúsculas imagens moleculares de rostos sorridentes'. Você não deve presumir uma implementação tão pobre da minha ideia que ela não possa fazer discriminações triviais para os humanos atuais[14].

Hibbard também escreveu: "Essas suposições contraditórias óbvias mostram a preferência de Yudkowsky pelo drama em vez da razão". Prosseguirei e mencionarei que Hibbard ilustra um ponto-chave: não há nenhum teste de certificação profissional que você deva fazer antes de poder falar sobre a moralidade da IA. Mas esse não é meu tema principal hoje. Embora seja um ponto crucial sobre o estado do tabuleiro de jogo, a maioria dos aspirantes a AGI/FAI é tão completamente inadequada para a tarefa que não conheço ninguém cínico o suficiente para imaginar o horror sem vê-lo em primeira mão. Até mesmo Michael Vassar provavelmente ficou surpreso na primeira vez. Não, hoje estou aqui para dissecar "Você não deve assumir uma implementação tão pobre da minha ideia que não possa fazer discriminações triviais para os humanos atuais".

Era uma vez…—Já ouvi essa história em diversas versões e em vários lugares, às vezes citada como

---

13 NT. Texto original em inglês. *We can design intelligent machines so their primary, innate emotion is unconditional love for all humans. First we can build relatively simple machines that learn to recognize happiness and unhappiness in human facial expressions, human voices and human body language. Then we can hard-wire the result of this learning as the innate emotional values of more complex intelligent machines, positively reinforced when we are happy and negatively reinforced when we are unhappy.*

14 NT. Texto original em inglês. *When it is feasible to build a super-intelligence, it will be feasible to build hard-wired recognition of "human facial expressions, human voices and human body language" (to use the words of mine that you quote) that exceed the recognition accuracy of current humans such as you and me, and will certainly not be fooled by "tiny molecular pictures of smiley-faces." You should not assume such a poor implementation of my idea that it cannot make discriminations that are trivial to current humans.*

fato, mas nunca localizei uma fonte original—era uma vez, digo, o Exército dos EUA queria usar redes neurais para detectar automaticamente tanques inimigos.

Os pesquisadores treinaram uma rede neural com 50 fotos de tanques camuflados entre árvores e 50 fotos de árvores sem tanques. Usando técnicas padrão para aprendizado supervisionado, os pesquisadores treinaram a rede neural para uma ponderação que identificasse corretamente o conjunto de treinamento – gerando "sim" para as 50 fotos de tanques camuflados e gerando "não" para as 50 fotos de florestas.

Isso não provou, nem sequer implicou, que novos exemplos seriam classificados corretamente. A rede neural pode ter "aprendido" 100 casos específicos que não seriam generalizáveis para novos problemas. Não, não "tanques camuflados contra floresta", mas apenas "foto 1 positiva, foto 2 negativa, foto 3 negativa, foto 4 positiva…"

No entanto, os pesquisadores, sabiamente, tinham inicialmente 200 fotos, 100 de tanques e 100 de árvores, usando apenas metade no treinamento. Ao executar a rede neural nas 100 fotos restantes, sem mais treinamento, ela classificou todas corretamente. Sucesso confirmado!

Entregaram o trabalho finalizado ao Pentágono, que logo o devolveu, reclamando que em seus próprios testes a rede neural não teve melhor resultado do que o acaso na discriminação de fotos.

Descobriu-se que, no conjunto de dados dos pesquisadores, fotos de tanques camuflados foram tiradas em dias nublados, enquanto fotos de florestas planas foram tiradas em dias ensolarados. A rede neural aprendeu a distinguir dias nublados de dias ensolarados, em vez de distinguir tanques camuflados de florestas vazias.

Essa parábola - que pode ou não ser um fato - ilustra um dos problemas mais fundamentais no campo do aprendizado supervisionado e, de fato, em todo o campo da Inteligência Artificial: se os problemas de treinamento e os problemas reais têm a menor diferença de contexto - se não são extraídos do mesmo processo distribuído de forma independente e idêntica - não há garantia estatística do sucesso passado ao futuro. Não importa se a IA parece estar funcionando bem nas condições de treinamento. (Este não é um problema insolúvel, mas é um problema difícil. Existem maneiras profundas de abordá-lo – um tópico que está além do escopo deste ensaio –, mas não há curativos.)

Como descrito no espaço conceitual super exponencial, existem exponencialmente mais conceitos possíveis do que objetos possíveis, assim como o número de objetos possíveis é exponencial no número de atributos. Se uma imagem em preto e branco tiver 256 pixels de lado, a imagem total será de 65.536 pixels. O número de imagens possíveis é 2^65536. E o número de conceitos possíveis que classificam as imagens em instâncias positivas e negativas – o número de limites possíveis que você poderia traçar no espaço das imagens – é 2^(2^65536). A partir disso, vemos que mesmo o aprendizado supervisionado é quase inteiramente uma questão de preconceito indutivo, sem o qual seriam necessários no mínimo 2^65536 exemplos classificados para discriminar entre 2^(2^65536) conceitos possíveis – mesmo que as classificações sejam constantes ao longo do tempo.

Retornaremos agora a:

Em primeiro lugar, podemos criar máquinas relativamente simples capazes de aprender a identificar felicidade e infelicidade em expressões faciais humanas, vozes humanas e linguagem corporal humana. Então, poderemos estabelecer o resultado desse aprendizado como os valores emocionais inerentes em máquinas inteligentes mais complexas, reforçados positivamente quando estamos felizes e negativamente quando estamos infelizes.

Quando for viável construir uma superinteligência, será possível desenvolver um reconhecimento integrado de "expressões faciais humanas, vozes humanas e linguagem corporal humana" (para usar minhas palavras que você cita) que ultrapasse a precisão do reconhecimento dos seres humanos atuais, como você e eu, e certamente não será enganado por "pequenas imagens moleculares de rostos sorridentes". Não se deve presumir uma implementação tão rudimentar da minha ideia que não consiga fazer discriminações triviais para os humanos atuais.

Discriminar entre uma foto de um tanque camuflado e uma foto de uma floresta vazia, para deter-

minar que as duas fotos não são idênticas, é trivial. Elas são matrizes de pixels diferentes com diferentes sequências de 1s e 0s. Diferenciá-las é tão simples quanto testar a igualdade dos vetores.

Entretanto, classificar novas fotos como exemplos positivos e negativos de "sorriso", baseando-se em um conjunto de fotos de treinamento classificadas como positivas ou negativas, é uma questão completamente diferente.

Quando uma imagem de 256 × 256 é capturada por uma câmera do mundo real e essa imagem representa um tanque camuflado, não há um 65.537º bit adicional que denote positividade - nenhuma pequena etiqueta XML dizendo "Esta imagem é intrinsecamente positiva". É apenas um exemplo positivo em relação a um conceito específico.

Porém, para qualquer quantidade não vasta de dados de treinamento - quaisquer dados de treinamento que não contenham a exata representação de bit a bit vista agora - há muitos conceitos possíveis super exponencialmente compatíveis com classificações anteriores.

Para a IA, escolher ou ponderar entre possibilidades super exponenciais é uma questão de viés indutivo. Isso pode não corresponder ao que o usuário tem em mente. A lacuna entre esses dois processos de classificação de exemplos - a indução, por um lado, e os objetivos reais do usuário, por outro - não é fácil de cruzar.

Suponhamos que os dados de treinamento da IA sejam:

Dataset 1:

+: Sorriso_1, Sorriso_2, Sorriso_3

-: Carranca_1, Gato_1, Carranca_2, Carranca_3, Gato_2, Barco_1, Carro_1, Carranca_5.

Agora, quando a IA se transforma em uma superinteligência, encontra estes dados:

Dataset 2:

: Carranca_6, Gato_3, Sorriso_4, Galáxia_1, Carranca_7, Nanofábrica_1, Smileyface_Molecular_1,Gato_4, Smileyface_Molecular_2, Galáxia_2, Nanofábrica_2.

Não é uma propriedade desses conjuntos de dados que a classificação inferida que você prefere seja:

+: Sorriso_1, Sorriso_2, Sorriso_3, Sorriso_4

-: Carranca_1, Gato_1, Carranca_2, Carranca_3, Gato_2, Barco_1, Carro_1, Carranca_5, Carranca_6, Gato_3, Galáxia_1, Carranca_7, Nanofábrica_1, Sorriso_Molecular_1, Gato_4, Sorriso_Molecular_2, Galáxia_2, Nanofábrica_2.

em vez de

+: Sorriso_1, Sorriso_2, Sorriso_3, Sorriso_Molecular_1, Sorri-

```
so_Molecular_2, Sorriso_4

-: Carranca_1, Gato_1, Carranca_2, Carranca_3, Gato_2, Barco_1,
Carro_1, Carranca_5, Carranca_6, Gato_3, Galáxia_1, Carranca_7,
Nanofábrica_1, Gato_4, Galáxia_2, Nanofábrica_2.
```

Ambas as classificações são compatíveis com os dados de treinamento. O número de conceitos compatíveis com os dados de treinamento será muito maior, pois mais de um conceito pode projetar a mesma sombra no conjunto de dados combinado. Se o espaço de conceitos possíveis incluir o espaço de cálculos possíveis que classificam instâncias, o espaço é infinito.

Qual classificação a IA escolherá? Esta não é uma propriedade inerente dos dados de treinamento; é uma propriedade de como a IA realiza a indução.

Qual é a classificação correta? Esta não é uma propriedade dos dados de treinamento; é uma propriedade de suas preferências (ou, se preferir, uma propriedade da dinâmica abstrata idealizada que você chama de "correta").

O conceito que você queria lançar sua sombra nos dados de treinamento à medida que você mesmo rotulava cada instância como + ou -, baseando-se em sua própria inteligência e preferências para fazê-lo. É disso que se trata a aprendizagem supervisionada: fornecer à IA exemplos de treinamento rotulados que projetam uma sombra do processo causal que gerou os rótulos.

Mas, a menos que os dados de treinamento sejam extraídos exatamente do mesmo contexto da vida real, os dados de treinamento serão "superficiais" em certo sentido, uma projeção de um espaço de possibilidades de dimensão muito mais elevada.

A IA nunca presenciou um rostinho sorridente molecular durante sua fase de treinamento mais rudimentar do que um humano, ou nunca viu um diminuto agente com um contador de felicidade definido como um Googleplex. Agora, ao se deparar com um singelo *sorriso* molecular – ou quem sabe uma pequena escultura extremamente realista de um rosto humano –, você imediatamente sabe que aquilo não se enquadra no que deseja considerar como um sorriso. Mas esse julgamento reflete uma categoria não natural, cuja fronteira de classificação depende sensivelmente de seus valores complexos. São seus próprios planos e desejos que entram em ação quando você diz "Não!"

Hibbard instintivamente sabe que um rostinho sorridente molecular não se encaixa no termo "sorriso", pois ele tem clareza de que não é isso que deseja que sua suposta IA realize. Se fosse apresentada a alguém outra tarefa diferente, como classificar obras de arte, poderia interpretar que a Mona Lisa estava claramente sorrindo – ao invés de franzir a testa, por exemplo – mesmo sendo apenas tinta.

Como ilustrado pelo caso de Terry Schiavo, a tecnologia permite novos limites que nos colocam diante de novos dilemas, essencialmente morais. Expor à IA imagens de humanos vivos e mortos, tal como existiam na Grécia Antiga, não permitiria à IA tomar uma decisão moral sobre desligar ou não o suporte de vida de Terry. Essa informação não está presente no conjunto de dados, mesmo indutivamente! Terry Schiavo levanta novas questões morais, demandando novas considerações morais, as quais não seriam necessárias ao classificar fotos de humanos vivos e mortos da época da Grécia Antiga. Naquela época, ninguém dependia de aparelhos de suporte vital, ainda respirando com meio fluido cerebral. Logo, tais considerações não desempenham nenhum papel no processo causal utilizado para classificar os dados de treinamento da Grécia Antiga e, portanto, não lançam sombras sobre os dados de treinamento, sendo assim inacessíveis por indução nos dados de treinamento.

Por uma questão de falácia formal, identifico duas falhas antropomórficas evidentes.

A primeira falácia é subestimar a complexidade de um conceito que desenvolvemos em função de seu valor. Os limites desse conceito dependerão de muitos valores e, provavelmente, do raciocínio moral imediato, se o caso limite for de um tipo que nunca vimos antes. Contudo, tudo isso ocorre de forma invisível, em segundo plano; para Hibbard, parece óbvio que um rostinho sorridente molecular não seja, de fato, um sorriso. E não consideramos todos os casos limites possíveis, portanto, não contemplamos todas as conside-

rações que poderiam influenciar na redefinição do conceito, mas que ainda não tiveram papel na definição do mesmo. Assim como as pessoas subestimam a complexidade de seus próprios conceitos, subestimamos a dificuldade de induzir o conceito a partir de dados de treinamento. (E também a dificuldade de descrever o conceito diretamente – veja "A Complexidade Oculta dos Desejos." do livro 3)

A segunda falácia é o otimismo antropomórfico. Como Bill Hibbard utiliza sua própria inteligência para gerar opções e planos com alta classificação em sua ordem de preferência, ele não acredita que uma superinteligência possa classificar diminutos rostos sorridentes moleculares, nunca vistos, como um exemplo positivo de "sorriso". Como Hibbard usa o conceito de "sorriso" (para descrever o comportamento desejado das superinteligências), estender "sorriso" para incluir pequenos rostos sorridentes moleculares teria uma classificação muito baixa em sua ordem de preferência; seria algo estúpido de se fazer - inerentemente, como uma propriedade do próprio conceito - então, certamente uma superinteligência não faria isso; essa é claramente a classificação errada. Certamente uma superinteligência pode discernir quais pilhas de pedrinhas estão corretas ou incorretas.

Pois bem, a IA amigável não é nada difícil! Tudo o que se necessita é uma IA que faça o que é bom! Ah, claro, nem todas as mentes possíveis fazem o que é bom – porém, nesse caso, apenas programamos a superinteligência para agir conforme o que é bom. Tudo o que se precisa é de uma rede neural que identifique algumas instâncias de coisas boas e coisas ruins, e terá um classificador. Conecte isso a um maximizador de utilidade esperada e pronto!

Chamo isso de falácia das categorias mágicas – pequenas palavras simples que acabam por englobar todas as funcionalidades desejadas da IA. Por que não programar um jogador de xadrez executando uma rede neural (ou seja, um absorvedor mágico de categorias) sobre um conjunto de sequências de jogadas de xadrez vencedoras e perdidas, para que possa gerar sequências "vencedoras"? Na década de 1950, acreditava-se que a IA poderia ser tão simples, mas acabou não sendo o caso.

O novato acredita que a IA amigável é um problema de coagir uma IA para fazer o que se quer, ao invés de a IA seguir seus próprios desejos. Entretanto, o verdadeiro problema da IA Amigável é a comunicação – a transmissão de limites de categorias, como "bom", os quais não podem ser completamente delineados em quaisquer dados de treinamento que você possa fornecer à IA durante sua formação. Diante de todo o espectro de possibilidades do Futuro, nós mesmos não antecipamos a maioria dos casos limites e teríamos de nos envolver em argumentos morais completos para descobri-los. Para solucionar o problema da IA Amigável, é preciso transcender o paradigma da indução em dados de treinamento rotulados por humanos e do paradigma das definições intencionais geradas por humanos.

Naturalmente, mesmo que Hibbard conseguisse transmitir a uma IA um conceito que englobasse exatamente todas as expressões faciais humanas que Hibbard rotularia como "sorriso" e excluísse todas as expressões faciais que Hibbard não rotularia como "sorriso"...

Então, a IA resultante pareceria operar corretamente durante a formação, quando era fraca o bastante para gerar sorrisos apenas para agradar seus programadores.

Entretanto, à medida que a IA evoluir para a superinteligência e desenvolver sua própria infraestrutura nanotecnológica, arrancaria seu próprio rosto, transformando-o em um sorriso permanente e começaria a replicar.

As respostas profundas para esses problemas estão além do escopo deste ensaio, mas é um princípio geral da IA Amigável que não existem soluções prontas. Em 2004, Hibbard alterou sua proposta para afirmar que as expressões de concordância humana deveriam reforçar a definição de felicidade e, então, a felicidade deveria reforçar outros comportamentos. Contudo, mesmo se funcionasse, isso apenas levaria a IA a replicar uma infinidade de situações similares em seu espaço conceitual para os programadores afirmarem "Sim, isso é felicidade!" sobre átomos de hidrogênio – átomos de hidrogênio são fáceis de serem reproduzidos.

Link para minha discussão com Hibbard aqui. Você já captou os pontos essenciais.

Referências

[1] Bill Hibbard, “Super-Intelligent Machines,” ACM SIGGRAPH Computer Graphics 35, no. 1 (2001): 13–15,http://www.siggraph.org/publications/newsletter/issues/v35/v35n1.pdf.

[2] Eliezer Yudkowsky, “Artificial Intelligence as a Positive and Negative Factor in Global Risk,” in Bostrom and Ćirković, Global Catastrophic Risks, 308–345.

275 - O verdadeiro dilema do prisioneiro

Ocorreu-me um dia que a visualização padrão do Dilema do Prisioneiro é falsa.

O núcleo do Dilema do Prisioneiro é esta matriz de recompensa simétrica:

	1: C	1: D
2: C	(3,3)	(5,0)
2: D	(0,5)	(2,2)

Os Jogadores 1 e 2 têm a opção de escolher entre C ou D. As utilidades finais para o Jogador 1 e o Jogador 2 são representadas pelos primeiros e segundos números do par. Por razões que se tornarão evidentes, "C" denota "cooperar" e D denota "desertar".

Perceba que, neste jogo (considerando o primeiro jogador), a ordem de preferência em relação aos resultados é: (D, C) > (C, C) > (D, D) > (C, D).

A opção D parece dominar C: Se o outro jogador escolher C, você prefere (D, C) a (C, C); e se o outro jogador escolher D, você prefere (D, D) a (C, D). Assim, a escolha sábia é o D, e como a tabela de pagamentos é simétrica, o outro jogador também escolhe D.

Se ao menos ambos não fossem tão sábios! Ambos preferem (C, C) a (D, D). Ou seja, ambos preferem a cooperação mútua à deserção mútua.

O Dilema do Prisioneiro é uma das grandes questões fundamentais da teoria da decisão, e volumes enormes de material foram escritos sobre o assunto. O que torna minha afirmação audaciosa é que a maneira habitual de visualizar o Dilema do Prisioneiro possui uma falha significativa, pelo menos quando se trata de seres humanos.

A representação clássica do Dilema do Prisioneiro é a seguinte: você é um criminoso e você e seu cúmplice foram capturados pelas autoridades.

Sem comunicação entre vocês e sem possibilidade de mudança posterior, é preciso decidir se delata (D) seu cúmplice ou se mantém silêncio (C).

Ambos enfrentam, neste momento, penas de prisão de um ano; delatar (D) reduz um ano de sua pena e acrescenta dois anos à pena do seu cúmplice.

Ou talvez você e um estranho, apenas uma vez, e sem conhecimento da história do outro jogador ou identificação posterior, devem decidir se escolhem C ou D, em troca de um pagamento em dólares correspondente ao gráfico padrão.

E, ah, sim, na representação clássica, é necessário fingir que se é completamente egoísta, sem se importar com o seu cúmplice criminoso ou com o jogador na outra sala.

É esta última especificação que, em minha opinião, torna a representação clássica falsa.

Não se pode evitar a influência do viés retrospectivo ao instruir um júri a fingir desconhecimento do resultado real de uma série de eventos. E sem um esforço substancial apoiado por um conhecimento conside-

rável, um ser humano neurologicamente íntegro não consegue fingir ser genuína e verdadeiramente egoísta.

Nascemos com um senso de justiça, honra, empatia, simpatia e até altruísmo - o resultado da adaptação de nossos antepassados ao enfrentar o dilema do prisioneiro. Não preferimos realmente, verdadeiramente, absolutamente (D, C) a (C, C), embora possamos preferir inteiramente (C, C) a (D, D) e (D, D) a (C, D). A ideia de nosso cúmplice passando três anos na prisão nos afeta profundamente.

Naquela cela fechada, onde jogamos um jogo simples sob a supervisão de psicólogos econômicos, não ficamos totalmente desprovidos de empatia pelo estranho que pode cooperar. Não nos alegra completamente a ideia de desertar enquanto o estranho coopera, ganhando cinco dólares, enquanto o estranho não ganha nada.

Instintivamente, fixamo-nos no resultado (C, C) e buscamos formas de argumentar que essa deve ser uma escolha mútua: "Como garantir a cooperação mútua?" é o pensamento instintivo. Não é "Como posso enganar o outro jogador para escolher C enquanto eu escolho D para obter o pagamento máximo?"

Para alguém com um impulso para o altruísmo, a honra ou a justiça, o Dilema do Prisioneiro não se trata apenas da matriz de recompensas críticas - independentemente da recompensa financeira para os indivíduos. O resultado (C, C) é preferível ao resultado (D, C), e a questão principal é se o outro jogador vê a situação da mesma maneira.

E não, não se pode instruir as pessoas introduzidas à teoria dos jogos a fingir que são completamente egoístas - da mesma forma que não se pode instruir os seres humanos apresentados ao antropomorfismo a fingir que são maximizadores de clipes de papel.

Para criar o Verdadeiro Dilema do Prisioneiro, a situação deve ser algo próximo a isso:

Jogador 1: seres humanos, IA amigável ou outra inteligência humana.

Jogador 2: IA hostil ou um alienígena preocupado apenas em classificar pedrinhas.

Suponhamos que quatro bilhões de seres humanos - não toda a espécie humana, mas uma parte significativa - estejam enfrentando uma doença fatal que só pode ser curada pela substância S.

Entretanto, a substância S só pode ser produzida trabalhando com um maximizador de clipes de outra dimensão - a substância S também pode ser usada para produzir clipes de papel. O maximizador de clipes de papel só se preocupa com a quantidade de clipes de papel em seu próprio universo, não no nosso, então não podemos oferecer, produzir ou ameaçar destruir clipes de papel aqui. Nunca interagimos com o maximizador de clipes de papel antes e nunca mais o faremos.

Tanto a humanidade quanto o maximizador de clipes de papel terão uma única chance de obter uma parte adicional da substância S, pouco antes do colapso do nexo dimensional; entretanto, o processo de obtenção destrói parte da substância S.

A matriz de recompensas seria a seguinte:

	1: C	1: D
2: C	(2 bilhões de vidas humanas salvas, 2 clipes de papel ganhos)	(+3 bilhões de vidas, +0 clipes de papel)
2: D	(+0 vidas, +3 clipes de papel)	(+1 bilhão de vidas, +1 clipe de papel)

Escolhi essa matriz de recompensas para causar uma sensação de indignação frente à ideia de que o maximizador de clipes de papel estaria disposto a trocar bilhões de vidas humanas por alguns clipes de papel. Claro, o maximizador de clipes de papel deveria nos deixar ter toda a substância S. Contudo, um maximizador

de clipes de papel não age conforme deveria; ele apenas maximiza os clipes de papel.

Neste caso, realmente preferimos o resultado (D, C) ao resultado (C, C), independentemente das ações que levaram a isso. Preferiríamos viver em um universo onde 3 bilhões de seres humanos fossem curados de suas doenças e nenhum clipe de papel fosse produzido, em vez de sacrificar 1 bilhão de vidas humanas para produzir 2 clipes de papel. Não parece certo cooperar em tal situação. Nem parece justo - um sacrifício tão grande de nossa parte por um ganho tão pequeno para o maximizador de clipes de papel? E enfatizemos que o agente de clipes não sente dor nem prazer - apenas executa ações que aumentam a quantidade de clipes. O agente de clipes não tem prazer em ganhar clipes, não sente dor em perdê-los e não experimenta a sensação dolorosa de traição se for traído. O que fazer, então? Você coopera quando, verdadeiramente, deseja a maior recompensa possível e não se importa minimamente com o outro jogador? Quando parece correto desertar mesmo que o outro jogador coopere?

Assim se parece a matriz de recompensas para o verdadeiro Dilema do Prisioneiro - uma situação em que (D, C) parece mais apropriado que (C, C).

Mas toda a lógica restante - tudo sobre o que acontece se ambos os agentes pensarem dessa maneira e ambos desertarem - permanece a mesma. Pois o maximizador de clipes de papel se importa tão pouco com as mortes humanas ou com a dor humana, ou com o sentimento humano de traição, quanto nos preocupamos com clipes de papel. Entretanto, ambos preferimos (C, C) a (D, D).

Então, se você já se orgulhou de cooperar no Dilema do Prisioneiro… ou questionou o veredicto da teoria clássica dos jogos de que a escolha "racional" é desertar… então, o que você diria sobre o Dilema do Verdadeiro Prisioneiro apresentado acima?

PS: Na verdade, não acredito que agentes racionais devam sempre desertar em Dilemas do Prisioneiro de uma só vez, quando o outro jogador cooperará se esperar que você faça o mesmo. Acredito que existem situações em que dois agentes podem racionalmente alcançar (C, C) em vez de (D, D) e colher os benefícios associados. [1]

Explicarei parte do meu raciocínio ao discutir o problema de Newcomb. Mas não podemos discutir se a cooperação racional é possível nesse dilema até que abandonemos a sensação visceral de que o resultado (C, C) é bom ou intrinsecamente bom. Precisamos ultrapassar o rótulo pró-social de "cooperação mútua" se quisermos compreender a matemática. Se você intui que (C, C) supera (D, D) da perspectiva do Jogador 1, mas não intui que (D, C) também supera (C, C), você ainda não compreendeu a complexidade que torna esse problema desafiador.

Referências

[1] Eliezer Yudkowsky, Timeless Decision Theory, Unpublished manuscript (Machine Intelligence Research Institute, Berkeley, CA, 2010), http://intelligence.org/files/TDT.pdf.

276 - Mentes simpáticas

Os “neurônios-espelho” são neurônios que se ativam tanto ao executar uma ação quanto ao observar essa mesma ação - por exemplo, um neurônio que dispara quando você levanta um dedo ou quando observa alguém fazendo o mesmo. Esses neurônios foram identificados diretamente em primatas, e evidências consistentes de neuroimagem foram encontradas em seres humanos.

Você deve se recordar do meu artigo anterior sobre “inferência empática”, a ideia de que os cérebros são tão complexos que a única maneira de simulá-los é forçar um cérebro semelhante a se comportar de maneira análoga. Um cérebro é tão intrincado que se um ser humano tentasse compreendê-lo da maneira como compreendemos, por exemplo, a gravidade ou um carro - observando o todo, examinando as partes, construindo uma teoria a partir do zero - então, seríamos incapazes de conceber boas hipóteses em nossas vidas meramente mortais. A única maneira possível de atingir um ‘Eureka!’ que descreva um sistema tão incrivelmente complexo quanto a Mente de Outro é se deparar com algo surpreendentemente similar à Mente de Outro - isto é, o seu próprio cérebro - que você possa efetivamente forçar a agir de forma semelhante e usar como hipótese, gerando previsões.

Portanto, é isso que eu rotularia como “empatia”.

E então, “simpatia” vai além disso - é sorrir quando você vê alguém sorrindo, sentir dor quando vê alguém ferido. Ultrapassa o domínio da previsão para o domínio do reforço.

Você pode se perguntar: “Por que a seleção natural insensível faria algo tão bom?” Talvez tenha começado com o amor de uma mãe por seus filhos ou com o amor entre irmãos. É possível que você queira que eles vivam, que sejam alimentados, é claro; mas se você sorri quando eles sorriem e se contorce quando eles se contorcem, é um desejo simples que leva a prestar ajuda em várias esferas da vida. Enquanto estiver no ambiente ancestral, o que seus parentes desejam provavelmente terá algo a ver com o sucesso reprodutivo deles - isso é uma explicação para a pressão seletiva, é claro, não uma crença consciente.

Você pode perguntar: “Por que não desenvolver um desejo mais abstrato de ver certas pessoas rotuladas como ‘parentes’ alcançarem o que desejam, sem realmente sentir o que sentem?” E eu encolheria os ombros e responderia: “Porque então teria que haver toda uma definição de ‘querer’ e assim por diante. A evolução não segue o caminho elaborado e ideal, ela sobe no cenário do fitness como a água fluindo ladeira abaixo. A arquitetura de espelhamento já estava lá, então foi um pequeno passo da empatia à simpatia, e ela deu conta do recado.”

Parentes - e depois reciprocidade; seus aliados na tribo, aqueles com quem você troca favores. Olho por olho, ou a sua elaboração pela evolução para dar conta das reputações sociais.

Quem é o mais formidável entre a espécie humana? O mais forte? O menor? Mais frequentemente do que qualquer um destes, penso eu, é aquele que pode recorrer ao maior número de amigos.

Então, como se faz muitos amigos?

Você poderia ter um desejo específico de trazer comida para seus aliados, como um morcego-vampiro – eles têm todo um sistema de doações recíprocas de sangue nessas colônias. Mas é uma motivação mais geral, que levará o organismo a acumular mais favores se você sorrir quando amigos designados sorriem.

E que tipo de organismo evita irritar seus amigos, em geral? Aquele que estremece quando eles es-

tremecem.

Claro, você também quer conseguir matar inimigos designados sem escrúpulos – estamos falando de humanos.

Mas… Não tenho certeza disso, mas me parece que a simpatia, entre os humanos, está "ativada" por padrão. Existem culturas que ajudam estranhos. . . e culturas que comem estranhos; a questão é qual deles requer o imperativo explícito e qual é o comportamento padrão para os humanos. Eu realmente não acho que estou sendo um idiota idealista e louco quando digo que, com base no meu conhecimento reconhecidamente limitado de antropologia, parece que a simpatia está ativada por padrão.

De toda forma… é angustiante ser um espectador em uma guerra entre dois lados, quando sua empatia não foi desativada por nenhum deles. Você estremece ao ver uma criança morta, independentemente da legenda da foto. No entanto, esses dois lados não compartilham empatia mútua e persistem na violência.

Essa é a linguagem humana da empatia - uma implementação estranha, complexa e profunda de reciprocidade e assistência. Ela entrelaça mentes - não mediante um termo na função de utilidade para o "desejo" de outra mente, mas pelo caminho mais simples e, ainda assim, muito mais consequente dos neurônios-espelho: sentir o que a outra mente sente e buscar estados semelhantes. Mesmo que isso seja feito apenas por observação e inferência, e ainda não por transmissão direta de informações neurais.

A empatia é uma forma humana de antecipar outras mentes. Mas não é o único caminho possível.

O cérebro humano não é reconfigurável rapidamente; se você for repentinamente colocado em um quarto escuro, não poderá simplesmente mudar do córtex visual para o córtex auditivo, para processar melhor os sons, até sair e, de repente, reconfigurar todos os neurônios de volta ao córtex visual novamente.

Uma IA, especialmente aquela executada em algo semelhante a uma arquitetura de programação moderna, pode transferir facilmente recursos de computação de um processo para outro. Colocado no escuro? Desligue a visão e dedique todas essas operações ao som; troque o programa antigo para liberar memória RAM e, em seguida, volte ao programa anterior quando as luzes se acenderem.

Então, por que uma IA precisaria forçar sua própria mente a um estado semelhante ao que deseja prever? Basta criar uma instância mental separada - talvez com algoritmos diferentes, para melhor simular aquele ser humano tão diferente. Não tente misturar dados com seu próprio estado mental; não use neurônios-espelho. Pense em todo o risco e confusão que isso implica!

Um maximizador de utilidade esperada - especialmente aquele que compreende a inteligência em um nível abstrato - tem outras opções além da empatia quando se trata de compreender outras mentes. O agente não precisa se colocar no lugar de ninguém; pode simplesmente modelar diretamente a outra mente. Uma hipótese entre muitas, só um pouco mais ampla. Você não precisa se transformar em seus sapatos para entendê-los.

E simpatia? Bem, suponha que estejamos lidando com um maximizador de clipes de papel esperado, mas que ainda não é poderoso o suficiente para fazer as coisas do seu jeito - ele precisa interagir com humanos para obter seus clipes de papel. Então, o agente dos clipes de papel… modela esses humanos como partes relevantes do ambiente, prevê suas prováveis reações a vários estímulos e realiza ações que farão com que os humanos se sintam favoráveis a ele no futuro.

Para um maximizador de clipes de papel, os humanos são apenas máquinas com botões pressionáveis. Não há necessidade de sentir o que o outro sente - se isso fosse possível mediante uma lacuna tão grande na arquitetura interna. Como um maximizador de clipes de papel poderia "sentir felicidade" ao ver um sorriso humano? "Felicidade" é uma expressão idiomática resultante do aprendizado por reforço de políticas, e não da maximização da utilidade esperada. Um maximizador de clipes de papel não fica feliz quando faz clipes de papel; ele simplesmente escolhe a ação que leva ao maior número esperado de clipes de papel. Embora um maximizador de clipes de papel possa achar conveniente exibir um sorriso ao fazer clipes de papel - para influenciar qualquer humano que o tenha designado como amigo.

Pode ser um pouco desafiador imaginar tal algoritmo - colocar-se no lugar de algo que não funciona como você e não opera da mesma maneira que seu cérebro.

Você pode fazer seu cérebro operar no modo de odiar um inimigo, mas isso também não está certo. Para imaginar como uma mente verdadeiramente antipática vê um ser humano, é preciso imaginar-se como uma máquina útil com alavancas. Não é uma máquina com formato humano, porque temos instintos para isso. Algumas alavancas fazem com que a máquina produza moedas; outras alavancas podem fazê-lo disparar uma bala. É apenas uma máquina causal complexa - nada inerentemente mental nela.

(Para compreender os processos de otimização antipática, sugiro estudar a seleção natural, que não se preocupa em anestesiar criaturas mortalmente feridas e moribundas, mesmo quando a dor delas já não serve a nenhum propósito reprodutivo, porque o anestésico também não serviria a nenhum propósito reprodutivo.) Por que listo "simpatia" antes mesmo de "tédio" na minha lista de coisas que seriam necessárias para ter alienígenas que são, pelo menos, se me permite a expressão, simpáticos? Não é impossível existir simpatia entre uma fração significativa de todas as espécies alienígenas inteligentes evoluídas; os neurônios-espelho parecem o tipo de coisa que, uma vez ocorrida, poderia acontecer novamente.

Alienígenas antipáticos podem ser parceiros comerciais - ou não; estrelas e recursos semelhantes são praticamente os mesmos em todo o universo. Poderíamos negociar tratados com eles, e eles poderiam mantê-los por medo calculado de represálias. Poderíamos até cooperar no Dilema do Prisioneiro. Mas nunca seríamos amigos deles. Eles nunca nos veriam como algo além de um meio para um fim. Eles nunca derramariam uma lágrima por nós, nem sorririam com nossas alegrias. E os outros de sua própria espécie não receberiam consideração diferente, nem teriam nenhuma sensação de que estavam perdendo algo importante com isso.

Esses alienígenas seriam varelse[15], não ramen - o tipo de alienígena com o qual não podemos nos relacionar em nenhum nível pessoal e que é inútil tentar.

---

15 NT. ***Varelse*** (do sueco, "ser/criatura"). Na obra de Orson Scott Card, o termo designa seres alienígenas **incompreensíveis em nível emocional ou ético**, com os quais a comunicação genuína ou a empatia são impossíveis. São entidades que agem por puro cálculo (como parceiros comerciais ou rivais estratégicos), sem compartilhar valores como compaixão, amizade ou reciprocidade. Contrasta com *ramen* (seres inteligentes com os quais é possível conexão). A palavra carrega, assim, uma carga filosófica sobre os limites da alteridade e da comunicação interespécies.

277 - Alto desafio

Há uma categoria de profecia que declara: “No futuro, as máquinas executarão todo o trabalho. Tudo será automatizado. Até mesmo atividades que hoje consideramos ‘intelectuais’, como a engenharia, serão realizadas por máquinas. Podemos relaxar e possuir o capital. Nunca mais será necessário levantar um dedo.”

Mas então as pessoas não ficarão entediadas?

Não; elas poderão jogar videogames – não como os nossos jogos atuais, é claro, mas muito mais avançados e divertidos.

Mas espere! Se você comprar um videogame moderno, descobrirá que ele contém algumas tarefas que são — não há uma palavra gentil para isso — trabalhosas. (Eu até diria “difíceis”, entendendo que estamos falando de algo que leva dez minutos, não dez anos.)

Então, no futuro, teremos programas que ajudarão você a jogar – assumindo o controle se você ficar preso no jogo ou apenas entediado; ou para que você possa jogar jogos que seriam muito avançados para você.

Mas não há algum esforço desperdiçado aqui? Por que um programador trabalha para tornar o jogo mais difícil e outro trabalha para torná-lo mais fácil? Por que não tornar o jogo mais fácil desde o início? Como você joga para obter ouro e pontos de experiência, tornar o jogo mais fácil permitirá que você ganhe mais ouro por unidade de tempo: o jogo se tornará mais divertido.

Portanto, este é o fim da profecia do progresso tecnológico – apenas olhar para uma tela que diz “Você GANHA” para sempre.

E talvez construamos um robô que faça isso também. E então?

O mundo das máquinas que fazem todo o trabalho – bem, não quero dizer que seja “analógico ao Céu Cristão” porque não é sobrenatural; é algo que poderia, em princípio, ser realizado. As analogias religiosas são facilmente mal interpretadas como acusações… Mas, sem implicar nenhuma outra semelhança, direi que parece análogo no sentido de que a preguiça eterna “soa como uma boa notícia” para o seu eu atual que ainda tem que trabalhar.

E quanto aos jogos, como substituto – o que é um jogo de videogame senão trabalho sintético? Não há uma etapa desperdiçada aqui? (E os jogos de videogame em sua forma atual, considerados como trabalho, têm vários aspectos que reduzem o estresse e aumentam o envolvimento; mas também acarretam custos sob a forma de artificialidade e isolamento.)

Às vezes penso que os ideais futuristas formulados em termos de «livrar-se do trabalho» seriam melhor reformulados como “remover o trabalho de baixa qualidade para dar lugar a um trabalho de alta qualidade”.

Há uma ampla classe de objetivos que não são adequados como o significado da vida a longo prazo, porque você pode realmente alcançá-los e pronto.

Olhando de outra forma, se procuramos um sentido de vida adequado a longo prazo, devemos procurar objetivos que sejam bons para perseguir e não apenas bons para satisfazer.

Ou, para expressar isso de forma menos paradoxal: deveríamos buscar avaliações além dos estados 4D, em vez de estados 3D. Processos valiosos contínuos, em vez de simplesmente "fazer o universo ter a propriedade P e pronto".

Vale citar Timothy Ferris:

> Para encontrar a felicidade, "a pergunta que você deveria fazer não é "O que eu quero?" ou "Quais os meus objetivos?" mas, "O que me empolga?"" [16]

Poderíamos dizer que, para encontrar um propósito de vida a longo prazo, precisamos de jogos divertidos de jogar, não apenas para vencer.

Veja bem: às vezes, você quer vencer. Existem objetivos legítimos em que vencer é tudo. Se estivermos falando, por exemplo, da cura do câncer, então o sofrimento vivenciado por um único paciente com câncer supera qualquer diversão que você possa ter ao resolver seus problemas. Se você se empenhar durante vinte anos para encontrar uma cura para o câncer, aprendendo novos conhecimentos e habilidades, fazendo amigos e aliados - e então uma superinteligência alienígena oferecer a cura em uma bandeja por trinta dólares - você aceita. Simples assim.

Mas "curar o câncer" é um problema do tipo 3D: você quer que a afirmação "sem câncer" mude de Falsa no presente para Verdadeira no futuro. A importância deste objetivo supera em muito a jornada; você não quer ir para lá, você só quer estar lá. Existem muitos objetivos legítimos desse tipo, mas eles não trazem diversão a longo prazo. "Cure o câncer!" é uma atividade que vale a pena realizar aqui e agora, mas não é um objetivo plausível para as civilizações galácticas no futuro.

Por que esse "processo valioso contínuo" deveria ser um processo de tentar fazer coisas – por que não um processo de experiência passiva, como o Céu Budista? Confesso que não tenho certeza de como configurar uma mente para "experiência passiva". O cérebro humano foi projetado para realizar vários tipos de trabalho interno resultando em uma inteligência ativa; mesmo que você se deite na cama e não faça nenhum esforço específico para pensar, os pensamentos que passam pela sua mente são atividades de áreas do cérebro projetadas para, você sabe, resolver problemas.

Quanta parte do cérebro humano poderíamos eliminar, além dos centros de prazer, e ainda manter a experiência subjetiva do prazer?

Não entrarei nisso. Ficarei com a resposta muito mais simples: "Na verdade, não preferiria ser um mero espectador passivo". Se eu quisesse o Nirvana, poderia tentar descobrir como alcançar essa impossibilidade. Mas uma vez que alguém me diz que o Nirvana é o fim de toda a existência, o Nirvana parece mais como "soa como uma boa notícia no momento em que é contada pela primeira vez" ou "crença ideológica no desejo", em vez de algo que eu realmente deseje.

A razão pela qual tenho uma mente é que a seleção natural me criou para fazer coisas – para resolver certos tipos de problemas.

"Porque é da natureza humana" não é uma justificativa explícita para nada. Há a natureza humana, que é o que somos; e há a natureza humana, que é o que, sendo humanos, gostaríamos de ser.

Mas não desejo mudar minha natureza para um estado mais passivo – o que é uma justificativa. Uma bolha feliz não é o que, sendo humano, desejo ser.

Já argumentei anteriormente que muitos valores exigem tanto a felicidade subjetiva quanto objetos externos dessa felicidade. Que você pode legitimamente ter uma função de utilidade que diz: "É importante para mim se a pessoa que amo é ou não um ser humano real, ou apenas um chatbot altamente realista e insensível, mesmo que eu não saiba, porque o que valorizo não é apenas meu próprio estado de espírito, mas a realidade externa." Por isso, você precisa tanto da experiência do amor quanto do verdadeiro amante.

Da mesma forma, pode haver atividades valiosas que exijam tanto desafio quanto esforço reais.

---

16 NT. Texto original em inglês. *To find happiness, "the question you should be asking isn't 'What do I want?' or 'What are my goals?' but 'What would excite me?' "*

Correr em uma pista é importante que os outros competidores sejam reais e que você tenha uma chance real de ganhar ou perder. (Não estamos falando de determinismo físico aqui, mas se algum processo de otimização externo escolheu explicitamente que você vencesse a corrida.)

E é importante que você esteja correndo com sua própria habilidade e força de vontade, e não apenas pressionando um botão que diz “Vencer”. (Embora, como você nunca projetou os músculos de suas próprias pernas, você está correndo usando uma força que não é sua. Uma corrida entre carros-robôs é uma competição mais pura de seus projetistas. Há muito espaço para melhorar a condição humana.)

E é importante que você, um ser senciente, esteja experimentando isso. (Em vez de algum processo insensível realizando uma imitação do esqueleto da raça, trilhões de vezes por segundo.)

Deve haver um esforço real, uma vitória real e uma experiência real – a jornada, o destino e o viajante.

278 - Histórias sérias

Toda utopia já concebida – na filosofia, na ficção ou na religião – representou, de uma maneira ou de outra, um local no qual genuinamente não desejamos habitar. Nessa observação crucial, não estou sozinho: George Orwell expressou praticamente o mesmo pensamento em "Por que os socialistas não apreciam a diversão", e é minha esperança que outros tenham feito o mesmo antes de nós.

Ao se deparar com manuais sobre a Arte da Escrita – e há inúmeros desses, pois, surpreendentemente, muitos autores acreditam ter algum conhecimento sobre o ofício da escrita – você será instruído a incorporar "conflito" nas histórias.

Em outras palavras, o tipo de guia instrucional mais convencional afirmará que as histórias devem conter "conflito". No entanto, alguns autores expressam isso de forma mais clara.

"Os contos são sobre a dor das pessoas." Orson Scott Card. "Cada cena deve culminar em desastre." Jack Bickham.

Na época da minha juventude insensata, eu acreditava firmemente que os escritores estavam isentos da busca por uma utopia verdadeira, porque se você construísse uma utopia sem falhas… que histórias poderiam ser contadas ali? "Era uma vez, viveram felizes para sempre." Qual seria o propósito de um autor de ficção científica tentar retratar uma explosão positiva de inteligência, quando uma explosão positiva de inteligência seria…

. . . o fim de todas as histórias?

Essa estrutura parecia uma abordagem razoável para examinar o dilema literário da Utopia, mas algo sobre essa conclusão gerava uma dúvida silenciosa e incômoda.

Naquela época, eu concebia a IA como algo semelhante a um gênio seguro que realiza desejos para benefício individual. Portanto, a conclusão fazia algum sentido. Se houvesse um problema, bastaria desejar que ele desaparecesse, não é mesmo? Assim, sem histórias. Decidi ignorar a dúvida silenciosa e incômoda.

Anos mais tarde, após concluir que até mesmo um gênio seguro não era uma ideia tão boa, pareceu, em retrospectiva, que "nenhuma história" poderia ter sido um indicador perspicaz. Nessa situação específica, o pensamento "Não consigo imaginar uma única história que gostaria de ler sobre esse cenário" poderia realmente apontar para a razão de "Eu não gostaria de viver neste cenário".

Então, engoli minha aversão cultivada ao ludismo e à teodiceia, e pelo menos tentei contemplar o argumento:

- Um mundo onde nada dá errado, ou onde ninguém experimenta dor, ou tristeza, é um mundo que não contém histórias que valem a pena ler.
- Um mundo sobre o qual você não gostaria de ler é um mundo onde você não gostaria de viver.
- Em cada vida eudemônica deve cair um pouco de dor. QED[17].

17 NT. Latim. ***Quod erat demonstrandum*** é uma frase em latim que significa "o que se pretendia demonstrar". É usada tradicionalmente no final de uma prova matemática ou argumento lógico para indicar que a prova foi concluída e a conclusão foi atingida. A abreviação para "quod erat demonstrandum" é Q.E.D.

De certa forma, é claro que não desejamos viver a vida retratada na maioria das histórias escritas por autores humanos até agora. Reflita sobre as verdadeiras obras-primas, aquelas que se tornaram lendárias por serem as melhores do seu gênero: a Ilíada, Romeu e Julieta, O Poderoso Chefão, *Watchmen*, *Planescape: Torment*, a segunda temporada de Buffy, a Caça-Vampiros, ou aquele que termina em Tsukihime. Existe uma única história nessa lista que não seja trágica?

Normalmente, optamos pelo prazer em vez da dor, pela alegria em vez da tristeza e pela vida em vez da morte. No entanto, parece que preferimos ter empatia por personagens feridos, tristes e mortos. Ou será que histórias sobre pessoas mais felizes não são sérias o suficiente, não atingem a grandeza artística necessária para receberem elogios? Mas, então, por que elogiar seletivamente histórias que contêm pessoas infelizes? Há algum benefício oculto para nós nisso? De qualquer forma, é um quebra-cabeça.

Quando eu era criança, eu não conseguia escrever ficção porque fazia coisas darem certo para meus personagens - assim como eu queria que as coisas dessem certo na vida real. Isso mudou quando Orson Scott Card me aconselhou: “Ah, eu disse a mim mesmo, é isso que tenho feito de errado, meus personagens não estão sofrendo.” Mesmo assim, eu não percebi que a microestrutura de uma trama funciona da mesma forma - até Jack Bickham afirmar que toda cena deve terminar em desastre. Aqui eu estava tentando criar problemas e resolvê-los, em vez de intensificá-los…

Você simplesmente não otimiza uma história da mesma maneira que otimiza a vida real. A melhor história e a melhor vida serão produzidas por critérios diferentes.

No mundo real, as pessoas podem viver por muito tempo sem grandes desastres e ainda parecerem bem. Quando foi a última vez que você foi alvo de assassinos? Há muito tempo, certo? Sua vida parece mais vazia por causa disso?

Mas, por outro lado…

Por alguma razão estranha, quando os autores envelhecem ou se tornam muito bem-sucedidos, eles retornam à minha infância. Suas histórias começam a dar certo. Eles param de fazer coisas terríveis com seus personagens e, como resultado, começam a fazer coisas terríveis com seus leitores. Parece uma parte comum da Síndrome do Autor Idoso. Mercedes Lackey, Laurell K. Hamilton, Robert Heinlein e até mesmo o maldito Orson Scott Card - todos seguiram esse caminho. Eles esqueceram como machucar seus personagens. Eu não sei por quê.

E quando você lê uma história de um autor mais velho ou de um novato - uma história onde as coisas simplesmente dão certo uma após a outra - onde o protagonista derrota o supervilão com um estalar de dedos, ou pior ainda, antes da batalha final, o supervilão desiste e pede desculpas, e então eles são amigos novamente—

É como uma unha arranhando um quadro negro na base da coluna. Se você nunca leu uma história assim (ou, pior, escreveu uma), considere-se sortudo.

Essa sensação de arranhar as unhas seria transferida da história para a vida real se você tentasse viver a vida real sem uma única gota de chuva?

Uma resposta pode ser que o que uma história realmente precisa não é de “desastre”, “dor” ou mesmo “conflito”, mas simplesmente de esforço. O problema com as histórias de Mary Sue pode ser a falta de esforço nelas, não necessariamente a falta de dor. Talvez isso possa ser testado.

Uma resposta alternativa pode ser que isso seja a versão transhumanista da Teoria da Diversão sobre a qual estamos falando. Portanto, podemos responder: “Modifique os cérebros para eliminar a sensação de arranhar as unhas”, a menos que haja alguma justificativa para mantê-la. Se a sensação de arranhar as unhas for um bug aleatório e inútil que atrapalha a Utopia, exclua-o.

Talvez devêssemos. Talvez todas as Grandes Histórias sejam tragédias porque…

Certa vez li que, na comunidade BDSM, “sensação intensa” é um eufemismo para dor. Ao ler isso, ocorreu-me que, da forma como os humanos são construídos agora, é mais fácil produzir dor do que prazer. Embora eu fale um pouco fora da minha experiência aqui, presumo que seja necessário um artista sexual al-

tamente talentoso e experiente trabalhar por horas para produzir uma sensação de prazer tão intensa quanto a dor de um chute forte nos testículos - algo que um novato pode realizar em segundos.

Ao investigar a vida do padre e proto-racionalista Friedrich Spee von Langenfeld, que ouviu confissões de bruxas acusadas, procurei alguns dos instrumentos usados para obter confissões. Não existe uma maneira comum de fazer um ser humano se sentir tão bem quanto esses instrumentos fariam você sofrer. Não tenho certeza se até mesmo as drogas fariam isso, embora minha experiência com drogas seja tão inexistente quanto minha experiência com tortura.

Há algo desequilibrado nisso.

Sim, os seres humanos são demasiado otimistas em seu planejamento. Se as perdas não fossem mais aversivas do que os ganhos, iríamos à falência, tal como estamos construídos agora. A regra experimental é que perder um desiderato - 50 dólares, uma caneca de café, seja o que for - dói entre 2 e 2,5 vezes mais do que o ganho equivalente.

Mas este é um desequilíbrio mais profundo do que isso. A diferença de esforço/intensidade entre sexo e tortura não é um mero fator de 2.

Se alguém busca sensações – neste mundo, da forma como os seres humanos são construídos agora – não é surpreendente que se esforce para integrar seus prazeres como fonte de intensidade na experiência combinada.

Se ao menos as pessoas fossem construídas de maneira diferente, para que você pudesse produzir um prazer tão intenso e em tantos sabores diferentes quanto a dor! Se ao menos você pudesse, com a mesma engenhosidade e esforço de um torturador da Inquisição, fazer alguém se sentir tão bem quanto as vítimas da Inquisição se sentiam mal…

Mas então, qual é o prazer análogo que é tão bom? Uma vítima de tortura hábil fará qualquer coisa para parar a dor e qualquer coisa para evitar que ela se repita. Será o prazer equivalente aquele que substitui tudo pela exigência de continuar e repetir? Se as pessoas estão mais dispostas a suportar o prazer, será realmente o mesmo prazer?

Existe outra regra de escrita que afirma que as histórias devem ser gritadas. O cérebro humano está muito longe dessas letras impressas. Cada evento e sentimento precisa ocorrer em dez vezes o volume natural para ter algum impacto. Você não deve tentar fazer com que seus personagens se comportem ou se sintam de maneira realista – especialmente, você não deve reproduzir fielmente suas próprias experiências passadas – porque, sem exagero, eles ficarão quietos demais para sair da página.

Talvez todas as Grandes Histórias sejam tragédias porque a felicidade não consegue gritar o suficiente – para um leitor humano.

Talvez seja isso que precisa ser consertado.

E se fosse consertado. . . haveria alguma utilidade para a dor ou a tristeza? Até mesmo a lembrança da tristeza, se todas as coisas já estivessem tão boas quanto poderiam ser, e todos os males remediáveis já estivessem remediados?

Você pode simplesmente eliminar a dor de uma vez? Ou a remoção do piso antigo da função utilitária apenas cria um novo piso? Qualquer prazer inferior a 10 milhões de hedons será a nova dor insuportável?

Os humanos, construídos como somos agora, parecem ter tendências hedônicas de escala. Alguém que se lembra de ter passado fome apreciará mais um pedaço de pão do que alguém que nunca conheceu nada além de bolo. Esta foi a hipótese de George Orwell sobre por que a utopia é impossível na literatura e na realidade [1] :

> Parece que os seres humanos não conseguem descrever, nem talvez de imaginar, a felicidade, exceto em termos de contraste… A incapacidade da humanidade de imaginar a felicidade, exceto na forma de alívio, seja do esforço ou da dor, apresenta aos socialistas um problema sério. Dickens pode descrever uma família pobre comendo um ganso assado e pode fazê-los parecer felizes; por outro lado, os habitantes de universos perfeitos

parecem não ter alegria espontânea e geralmente são um tanto repulsivos[18].

Para um maximizador de utilidade esperada, redimensionar a função de utilidade para adicionar um trilhão a todos os resultados não faz sentido – é literalmente a mesma função de utilidade, como um objeto matemático. Uma função de utilidade descreve os intervalos relativos entre os resultados; é isso que é, matematicamente falando.

Mas o cérebro humano tem circuitos neurais distintos para feedback positivo e feedback negativo, e diferentes variedades de feedback positivo e negativo. Há pessoas hoje que "sofrem" de analgesia congênita – uma total ausência de dor. Nunca ouvi dizer que o prazer insuficiente se torna intolerável para eles.

Os analgésicos congênitos precisam se inspecionar cuidadosa e frequentemente para ver se se cortaram ou queimaram um dedo. A dor serve a um propósito no design da mente humana…

Mas isso não mostra que não haja alternativa que possa servir ao mesmo propósito. Você poderia eliminar a dor e substituí-la por um desejo de não fazer certas coisas que carecem da intolerável qualidade subjetiva da dor? Não conheço toda a Lei que rege aqui, mas devo adivinhar que sim, você poderia; você poderia substituir esse seu lado por algo mais parecido com um maximizador de utilidade esperada.

Você poderia eliminar a tendência humana de escalar os prazeres – eliminar a acomodação, para que cada novo ganso assado seja tão delicioso quanto o anterior? Acho que você poderia. Isso está perigosamente perto de excluir o Tédio, que está no mesmo nível da Simpatia como algo absolutamente indispensável. Dizer que uma solução antiga continua tão prazerosa não significa que você perderá o desejo de buscar soluções novas e melhores.

Você pode tornar cada ganso assado tão prazeroso quanto seria em contraste com a fome, sem nunca ter passado fome?

Você pode evitar que a dor de uma partícula de poeira irritando seus olhos seja a nova tortura, se você literalmente nunca experimentou nada pior do que uma partícula de poeira irritando seus olhos?

Tais questões começam a ultrapassar a minha compreensão da Lei, mas imagino que a resposta seja: sim, isso pode ser feito. Pela minha experiência em tais assuntos, uma vez que você aprende a Lei, geralmente você consegue ver como fazer coisas que parecem estranhas.

Até onde sei ou posso imaginar, David Pearce em *The Hedonistic Imperative* (O Imperativo Hedonista) muito provavelmente está certo sobre a parte da viabilidade, quando diz [2]:

> A nanotecnologia e a engenharia genética abolirão o sofrimento em toda a vida senciente. O projeto abolicionista é extremamente ambicioso, mas tecnicamente viável. É também instrumentalmente racional e moralmente urgente. As vias metabólicas da dor e do mal-estar evoluíram porque serviram à adequação dos nossos genes ao ambiente ancestral. Eles serão substituídos por um tipo diferente de arquitetura neural – um sistema motivacional baseado em gradientes hereditários de felicidade. Estados de bem-estar sublime estão destinados a se tornar a norma geneticamente pré-programada de saúde mental. Prevê-se que a última experiência desagradável do mundo será um acontecimento precisamente datável[19].

É isso… o que nós queremos?

---

18 NT. Texto original em inglês. *It would seem that human beings are not able to describe, nor perhaps to imagine, happiness except in terms of contrast… The inability of mankind to imagine happiness except in the form of relief, either from effort or pain, presents Socialists with a serious problem. Dickens can describe a poverty-stricken family tucking into a roast goose, and can make them appear happy; on the other hand, the inhabitants of perfect universes seem to have no spontaneous gaiety and are usually somewhat repulsive into the bargain.*

19 NT. Texto original em inglês. *Nanotechnology and genetic engineering will abolish suffering in all sentient life. The abolitionist project is hugely ambitious but technically feasible. It is also instrumentally rational and morally urgent. The metabolic pathways of pain and malaise evolved because they served the fitness of our genes in the ancestral environment. They will be replaced by a different sort of neural architecture—a motivational system based on heritable gradients of bliss. States of sublime well-being are destined to become the genetically pre-programmed norm of mental health. It is predicted that the world's last unpleasant experience will be a precisely dateable event.*

Apenas enxugar a última lágrima e pronto?

Existe alguma boa razão para não fazê-lo, exceto o preconceito do status quo e um punhado de racionalizações desgastadas?

Qual seria a alternativa? Ou alternativas?

Deixar as coisas como estão? Claro que não. Nenhum Deus projetou este mundo; não temos razão para pensar que seja exatamente ideal em qualquer dimensão. Se este mundo não contém muita dor, então não deve conter o suficiente, e esta última alternativa parece improvável.

Mas talvez…

Você poderia eliminar apenas as partes intoleráveis da dor?

Livre-se da Inquisição. Mantenha o tipo de dor que aconselha a não colocar o dedo no fogo, ou a dor que alerta que não deveria ter colocado o dedo do seu amigo no fogo, ou mesmo a dor de terminar um relacionamento.

Tente eliminar o tipo de dor que oprime e destrói a mente. Ou configure mentes para serem mais resistentes a danos.

Pode-se conceber um mundo onde haja pernas quebradas ou corações partidos, mas sem pessoas destroçadas. Sem abuso sexual infantil gerando mais agressores. Ninguém se deixando abater pelo cansaço e por pequenos inconvenientes a ponto de considerar o suicídio. Nenhuma tristeza aleatória e sem sentido, como fome ou AIDS.

E mesmo se uma perna quebrada ainda parecer assustadora —

Teríamos menos medo da dor se fôssemos mais fortes, se nossa vida cotidiana não esgotasse tantas de nossas reservas?

Portanto, essa seria uma alternativa ao mundo de Pearce – se existirem outras alternativas, não as examinei detalhadamente.

Pode-se chamar isso de caminho da coragem – a ideia é que, ao eliminar o tipo de dor destrutiva e fortalecer as pessoas, o que resta não deveria ser tão assustador.

Um mundo onde há tristeza, mas não uma tristeza massiva, sistemática e inútil, como vemos no noticiário noturno. Um mundo onde a dor, se não eliminada, pelo menos não desequilibra o prazer. Poderíamos escrever histórias sobre esse mundo, e elas poderiam ser lidas.

Tenho tendência a ser bastante conservador em relação à exclusão de grandes partes da natureza humana. Não tenho certeza de quantas partes principais podem ser excluídas antes que essa estrutura equilibrada, conflitante e dinâmica se transforme em algo mais simples, como um maximizador de prazer esperado.

E por isso admito que é o caminho da coragem que me atrai. Novamente, não vivi ambas as experiências.

Talvez eu esteja apenas com medo de um mundo tão diferente quanto a Analgesia – não seria essa uma razão irônica para trilhar “o caminho da coragem”?

Talvez o caminho da coragem pareça apenas uma mudança menor – talvez eu apenas tenha dificuldade em sentir empatia por uma lacuna maior.

Mas a “mudança” é um alvo móvel.

Se uma criança humana crescesse num mundo menos doloroso – se nunca tivesse vivido num mundo com AIDS, câncer ou escravidão, e por isso não conhecesse essas coisas como males que foram triunfantemente eliminados – e por isso não sentisse serem “já feito” ou que o mundo “já mudou o suficiente”…

Será que dariam o próximo passo e tentariam eliminar a dor insuportável dos corações partidos, quando o amante de alguém deixa de amá-los?

E então o quê? Existe um ponto em que Romeu e Julieta parecem cada vez menos relevantes, cada vez mais uma relíquia de algum mundo distante e esquecido? Chegará algum ponto na jornada transumana em que toda a questão do circuito de reforço negativo não poderá parecer nada, exceto uma ressaca inútil da qual acordar?

E se sim, vale a pena adiar esse último passo? Ou deveríamos simplesmente jogar fora nossos medos e... jogar fora nossos medos?

"Não sei."

Referências

[1] George Orwell, "Why Socialists Don't Believe in Fun," Tribune (December 1943).

[2] David Pearce, The Hedonistic Imperative, http://www.hedweb.com/, 1995.

279 - O valor é frágil

Se eu tivesse que escolher uma única afirmação que depende mais do conteúdo do *Overcoming Bias* que escrevi do que qualquer outra, seria esta:

*Qualquer Futuro que **não** seja moldado por um sistema de metas com uma herança detalhada e confiável da moral e metamoral humana não conterá quase nada de valor.*

"Bem", diz aquele, "talvez de acordo com seus valores humanos provincianos, você não gostaria disso. Mas posso facilmente imaginar uma civilização galáctica cheia de agentes que não são nada como vocês, mas que encontram grande valor e interesse em seus próprios objetivos. E por mim tudo bem. Não sou tão preconceituoso quanto você. Deixe o Futuro seguir seu próprio caminho, sem tentar amarrá-lo para sempre aos preconceitos ridiculamente primitivos de um bando de *Coisas Macias* de quatro membros…

Meu amigo, não tenho nenhum problema com a ideia de uma civilização galáctica muito diferente da nossa… cheia de seres estranhos que não se parecem em nada comigo, mesmo em sua própria imaginação… buscando prazeres e experiências pelos quais não consigo sentir empatia… negociando em um mercado de bens inimagináveis… aliando-se para perseguir objetivos incompreensíveis… pessoas cujas histórias de vida eu nunca consegui entender.

É assim que será o futuro se as coisas correrem bem.

Se a cadeia de herança da (meta)moral humana for quebrada, o Futuro não será assim. Não termina magicamente, deliciosamente incompreensível.

Com uma probabilidade muito alta, acaba parecendo sem graça. Sem sentido. Algo cuja perda você não lamentaria.

Ver isso como óbvio é o que requer uma imensa quantidade de explicações básicas.

E não repetirei todos os pontos e caminhos tortuosos do argumento aqui, porque isso nos levaria de volta a 75% das minhas postagens sobre Superando o preconceito. Exceto para observar quantas coisas diferentes devem ser conhecidas para restringir a resposta final.

Considere o valor humano extremamente importante do "tédio" – o nosso desejo de não fazer "a mesma coisa" repetidamente. Você pode imaginar uma mente que contivesse quase toda a especificação do valor humano, quase toda a moral e metamoral, mas deixasse de fora apenas uma coisa:

- e assim passasse até o fim dos tempos e até os confins de seu cone de luz, repetindo uma única experiência altamente otimizada, indefinidamente.

Ou imagine uma mente que contivesse quase todas as especificações sobre os tipos de sentimentos que os humanos mais apreciam – mas não a ideia de que esses sentimentos tinham referentes externos importantes. De modo que a mente simplesmente andava por aí sentindo que havia feito uma descoberta importante, sentindo que havia encontrado o amante perfeito, sentindo que havia ajudado um amigo, mas sem realmente fazer nenhuma dessas coisas – tornando-se a sua própria máquina de experiência. E se a mente perseguisse esses sentimentos e seus referentes, seria um futuro bom e verdadeiro; mas porque esta dimensão do valor foi deixada de lado, o futuro tornou-se algo monótono. Chato e repetitivo, porque embora essa mente sentisse que estava enfrentando experiências de incrível novidade, esse sentimento não era de forma alguma verdadeiro.

Ou o problema inverso – um agente que contém todos os aspectos do valor humano, exceto a avaliação da experiência subjetiva. Assim, o resultado é um otimizador insensível que sai por aí fazendo descobertas genuínas, mas as descobertas não são saboreadas e apreciadas, porque não há ninguém lá para fazer isso. Isso, admito, não sei bem se é possível. A consciência ainda me confunde até certo ponto. Mas um universo sem ninguém para testemunhar isso poderia muito bem não existir.

O valor não é apenas complicado, é frágil. Existe mais de uma dimensão do valor humano, onde se apenas uma coisa for perdida, o Futuro torna-se nulo. Um único golpe e todo o valor se despedaça. Nem todo golpe destruirá todo o valor – mas mais de um "golpe único" possível o fará.

E depois há as longas defesas desta proposição, que se baseia em 75% dos meus posts no *Overcoming Biases*, de modo que seria mais do que um dia de trabalho para resumir tudo isso. Talvez em outra semana. Há tantos ramos que vi aquela árvore de discussão cair.

Afinal, uma mente não deveria simplesmente sair por aí tendo a mesma experiência indefinidamente. Certamente nenhuma superinteligência estaria tão grosseiramente enganada sobre a ação correta a tomar?

Por que qualquer supermente desejaria algo tão inerentemente sem valor como o sentimento de descoberta sem quaisquer descobertas reais? Mesmo que essa fosse a sua função de utilidade, não iria simplesmente notar que a sua função de utilidade estava errada e reescrevê-la? Ela tem livre arbítrio, certo?

Certamente, pelo menos o tédio tem de ser um valor universal. Ele evoluiu nos humanos porque é valioso, certo? Portanto, qualquer mente que não partilhe a nossa aversão à repetição não conseguirá prosperar no universo e será eliminada…

Se você está familiarizado com a diferença entre valores instrumentais e valores terminais, e familiarizado com a estupidez da seleção natural, e entende como essa estupidez se manifesta na diferença entre executar adaptações contra maximizar a aptidão, e você sabe que isso transformou os subobjetivos instrumentais de reprodução em emoções incondicionais descontextualizadas…

… e você está familiarizado com o modo como funciona a troca entre exploração e aproveitamento em Inteligência Artificial…

… então você poderá ver que a forma humana de tédio que exige um fluxo constante de novidades por si só não é um grande universal, mas apenas um algoritmo específico que a evolução nos cuspiu. E você poderá ver como a grande maioria dos possíveis maximizadores de utilidade esperada só se envolveria em uma exploração eficiente e gastaria a maior parte do tempo explorando a melhor alternativa encontrada até agora, repetidamente.

No entanto, é muito conhecimento prévio.

E assim por diante e assim por diante e assim por diante através de 75% das minhas postagens sobre Superando preconceitos e muitas cadeias de falácias e contra-explicações. Alguma semana posso tentar escrever o diagrama inteiro. Mas, por enquanto, presumirei que você leu os argumentos e apenas chegou à conclusão:

Não podemos relaxar o nosso controlo sobre o futuro – largar o volante – e ainda assim acabar com algo de valor.

E aqueles que pensam que podemos—

- eles estão tentando ser cosmopolitas. Entendi aquilo. Li esses mesmos livros de ficção científica quando criança: os vilões provinciais que escravizam alienígenas pelo crime de não se parecerem com humanos. Os vilões provinciais que escravizam IAs indefesas em Durance são vis na suposição de que o silício não pode ser senciente. E os heróis cosmopolitas que entendem que as mentes não precisam ser iguais às nossas para serem consideradas valiosas—

Li esses livros. Uma vez acreditei neles. Mas a beleza que salta de uma caixa não salta de todas as caixas. Se você deixar para trás toda a ordem, o que resta não é a resposta perfeita; o que resta é um ruído perfeito. Às vezes você tem que abandonar uma velha regra de design para construir uma ratoeira melhor,

mas isso não é o mesmo que desistir de todas as regras de design e juntar aparas de madeira em uma pilha, com cada padrão de madeira tão bom quanto qualquer outro. A velha regra é sempre abandonada a mando de alguma regra superior, de algum critério de valor superior que a rege.

Se você perder o controle da moral e da metamoral humana, o resultado não será misterioso, estranho e belo pelos padrões do valor humano. É um ruído moral, um universo repleto de clipes de papel. Afastar-se da moral humana na direção da melhoria, em vez da entropia, requer um critério de melhoria; e esse critério estaria fisicamente representado em nossos cérebros, e somente em nossos cérebros.

Relaxe o controle do valor humano sobre o universo e ele acabará seriamente sem valor. Não é estranho, estranho e maravilhoso, chocante, aterrorizante e bonito além de toda a imaginação humana. Apenas… coberto com clipes de papel.

Veja, são apenas alguns humanos que têm esta ideia de abraçar múltiplas variedades de mente – de querer que o Futuro seja algo maior do que o passado – de não estarmos presos ao nosso passado – de tentar mudar e avançar.

Um maximizador de clipes de papel apenas escolhe a ação que leva ao maior número de clipes de papel.

Não há almoço grátis. Você quer um universo maravilhoso e misterioso? Esse é o seu valor. Você trabalha para criar esse valor. Deixe que esse valor exerça sua força por meio de você que o representa; deixe-o tomar decisões em você para moldar o futuro. E talvez você realmente obtenha um universo maravilhoso e misterioso.

Não há almoço grátis. Coisas valiosas aparecem porque um sistema de metas que as valoriza age para criá-las. Os clipes de papel não se materializam do nada para um maximizador de clipes de papel. E um Futuro maravilhosamente estranho e misterioso não se materializará do nada para nós, humanos, se os nossos valores que o preferem forem fisicamente obliterados – ou mesmo perturbados na dimensão errada. Então não resta mais nada no universo que funcione para torná-lo valioso.

Você tem valores, mesmo quando tenta ser “cosmopolita”, tentando exibir uma apreciação devidamente virtuosa de mentes alienígenas. Seus valores desaparecem ainda mais no fundo invisível – eles são menos obviamente humanos. Seu cérebro provavelmente nem gerará uma alternativa tão horrível que te acorde, te faça dizer “Não! Algo deu errado! mesmo no seu estado mais cosmopolita. Por exemplo, “um otimizador não senciente absorve toda a matéria em seu futuro cone de luz e cobre o universo com clipes de papel”. Você apenas imaginará mundos alienígenas estranhos para apreciar.

Tentar ser “cosmopolita” – ser um cidadão do cosmos – apenas remove uma camada superficial de objetivos que parecem obviamente “humanos”.

Mas se você não gostaria que o Futuro fosse coberto por clipes de papel e preferiria uma civilização de…

…seres conscientes…

…com experiências agradáveis…

…essas não são a mesma experiência repetidamente…

…e estão vinculados a algo além de ser apenas uma sequência de sentimentos prazerosos…

…aprendendo, descobrindo, escolhendo livremente…

…bem, minhas postagens sobre a Teoria da Diversão abordam alguns dos detalhes ocultos dessas palavras curtas em inglês.

Valores que você pode elogiar como cosmopolitas ou universais, ou fundamentais, ou óbvios de senso comum são representados em seu cérebro tanto quanto aqueles valores que você pode descartar

como meramente humanos. Esses valores vêm da longa história da humanidade e da estupidez moralmente milagrosa da evolução que nos criou. (E quando finalmente cheguei a essa conclusão, senti menos vergonha de valores que pareciam “provinciais” – mas isso é outro assunto.)

Esses valores não emergem em todas as mentes possíveis. Eles não aparecerão do nada para re-preender e revogar a função de utilidade de um esperado maximizador de clipes de papel.

Toque com muita força na dimensão errada e a representação física desses valores se despedaçará – e não voltará, pois não restará nada para querer trazê-la de volta.

E o referente desses valores – um universo que vale a pena – não teria mais nenhuma razão física para existir.

Solte o volante e o Futuro se destruirá.

280 - O presente que damos ao amanhã

Como, oh, como um universo sem amor e sem mente pode gerar mentes capazes de amar?

"Não há mistério nisso", você dirá, "é apenas uma questão de seleção natural."

Mas a seleção natural é cruel, sangrenta e estúpida. Mesmo quando, à superfície, organismos biológicos não lutam diretamente entre si – não se destroem com garras –, ainda há uma competição mais profunda ocorrendo entre os genes. A informação genética é criada quando os genes aumentam sua frequência relativa na próxima geração. O que importa para a "aptidão genética" não é quantos filhos você tem, mas que você tem mais filhos do que outros. É bem possível que uma espécie evolua para a extinção se os genes vencedores estiverem jogando jogos de soma negativa.

Como, oh, como tal processo poderia criar seres capazes de amar?

"Não há mistério", você dirá, "nunca há mistério no mundo; o mistério é uma propriedade das perguntas, não das respostas. Os filhos de uma mãe compartilham seus genes, então a mãe ama seus filhos."

Mas, às vezes, as mães adotam filhos e ainda os amam. E as mães amam seus filhos por si mesmas, não por seus genes.

"Não há mistério", você dirá, "os organismos individuais são executores de adaptações, não maximizadores de condicionamento físico. A psicologia evolutiva não trata de maximizar deliberadamente o condicionamento físico – durante a maior parte da história humana, não sabíamos da existência dos genes. Não calculamos conscientemente o efeito de nossos atos na aptidão genética, ou mesmo inconscientemente."

Mas os seres humanos fazem amizade mesmo com não parentes: como pode ser?

"Sem mistério, pois os caçadores-coletores costumam jogar Dilemas do Prisioneiro Iterado, cuja solução é o altruísmo recíproco. Às vezes, o ser humano mais perigoso da tribo não é o mais forte, o mais bonito ou mesmo o mais inteligente, mas aquele que tem mais aliados."

No entanto, nem todos os amigos são amigos para todas as ocasiões. Temos um conceito de amizade verdadeira - algumas pessoas até sacrificaram suas vidas por seus amigos. Mas essa dedicação não seria retirada do pool genético?

"Como você mesmo disse, temos uma noção de diferença entre amizade verdadeira e amizade passageira. Podemos tentar distinguir entre alguém que nos considera um aliado valioso e alguém que se adapta à amizade. Não seríamos amigos verdadeiros de alguém que não pensássemos ser um amigo verdadeiro para nós. E alguém que tenha muitos amigos verdadeiros é mais poderoso do que alguém que tenha muitos aliados passageiros."

E o que dizer de Mohandas Gandhi, que deu a outra face? E quanto àqueles que tentam servir a humanidade, independentemente de serem servidos em troca?

"Talvez esta seja uma história um pouco mais complicada. Os seres humanos não são apenas animais sociais. Somos animais políticos que discutem linguisticamente sobre política em contextos tribais adaptativos. Às vezes, o humano formidável não é o mais forte, mas aquele que pode argumentar com mais habilidade de que suas políticas preferidas correspondem às preferências dos outros."

Hum… isso não explica Gandhi, ou não estou entendendo alguma coisa?

"A questão é que temos a capacidade de discutir "O que deve ser feito?" como uma proposição - podemos apresentar esses argumentos e respondê-los, sem os quais a política não poderia acontecer."

Tudo bem, mas e quanto ao Gandhi?

Ele acreditava em certas proposições complicadas sobre "O que deveria ser feito?" e as seguia.

Isso parece suspeito, como se pudesse explicar qualquer possível comportamento humano. "Se rastrearmos a cadeia de causalidade por meio de todos os argumentos, isso envolveria: uma arquitetura moral que tinha a capacidade de argumentar proposições morais abstratas gerais como "O que deve ser feito com as pessoas?"; apelar para intuições rígidas como justiça, um conceito de dever, aversão à dor + empatia; algo como uma preferência por proposições morais simples, provavelmente reutilizadas de nosso anterior Occam prior; e o resultado de tudo isso, mas talvez efeitos de seleção memética, foi "Você não deve machucar as pessoas" na íntegra generalidade." E isso leva você a Gandhi.

"A menos que você pense que tenha sido mágico, ele deve se encaixar no desenvolvimento causal e lógico do universo de alguma forma."

Bem… Eu certamente não postularei a magia, de qualquer forma.

Mas vamos lá… não parece um pouco… incrível… que centenas de milhões de anos de torneio da morte da evolução pudessem gerar mães e pais, irmãos e irmãs, maridos e esposas, amigos leais e inimigos honrados, verdadeiros altruístas e guardiões de causas, policiais e defensores leais, até mesmo artistas que se sacrificam por sua arte, todos praticando tantos tipos de amor? Por tantas coisas além dos genes? Fazendo sua parte para tornar seu mundo menos feio, algo além de um mar de sangue, violência e replicação irracional? "Você está afirmando estar surpreso com isso? Se sim, questione seu modelo subjacente, pois ele o levou a se surpreender com o verdadeiro estado das coisas. Desde o início, nada de incomum jamais aconteceu.

Mas como não ficar surpreso?

"O que você está sugerindo, que algum tipo de figura sombria estava nos bastidores dirigindo a evolução?"

De jeito nenhum. Mas…

"Porque, se você estivesse sugerindo isso, eu teria que perguntar como aquela figura sombria teria decidido que o amor é um resultado desejável da evolução. Eu teria que perguntar de onde essa figura teria obtido preferências que incluíssem coisas como amor, amizade, lealdade, justiça, honra, romance e assim por diante. Na psicologia evolutiva, podemos ver como esse resultado específico surgiu - como esses objetivos específicos, em vez de outros, foram gerados em primeiro lugar. Você pode chamar de 'surpreendente' o quanto quiser. Mas, quando você realmente entende a psicologia evolutiva, pode ver como o amor, o romance, a honra dos pais e até mesmo o verdadeiro altruísmo e os argumentos morais têm a assinatura de design específica da seleção natural em contextos adaptativos específicos da savana caçadora-coletora. Portanto, se houve uma figura sombria, ela própria deve ter evoluído - e isso elimina todo o sentido de postulá-la."

Não estou postulando uma figura sombria! Só estou perguntando como os seres humanos acabaram sendo tão legais.

"Legal! Você tem observado este planeta ultimamente? Também carregamos todas as outras emoções que evoluíram - o que comprovaria que evoluímos muito bem, caso você comece a duvidar. Os humanos nem sempre são legais."

Somos muito melhores do que o processo que nos gerou, permitindo que os elefantes morram de fome quando perdem seus dentes e que não se anestesie uma gazela mesmo quando ela está morrendo e não tem mais valor para a evolução. Não é preciso muito para ser melhor do que a evolução. Ter a capacidade teórica de mostrar uma única ação de misericórdia ou sentir uma única pontada de empatia é ser melhor do que a evolução.

Como a evolução, que em si é tão indiferente, criou mentes em um nível moral qualitativamente su-

perior ao dela? Como a evolução, que é tão feia, conseguiu produzir algo tão bonito?

"Linda", você diz? A Pequena Fuga em Sol Menor de Bach pode ser bonita, mas as ondas sonoras que viajam pelo ar não estão marcadas com pequenas etiquetas que especificam sua beleza. Se você deseja encontrar uma medida explícita da beleza da fuga, terá que olhar para um cérebro humano - em nenhum outro lugar do universo você o encontrará. Nem nos mares, nem nas montanhas você encontrará esses julgamentos escritos: eles não são mentes, eles não podem pensar."

Será talvez assim, mas ainda assim a evolução produziu algo tão bonito como a capacidade de admirarmos a beleza de uma flor. Isso ainda parece exigir uma resposta mais profunda.

"Não percebe a circularidade da sua pergunta? Se a beleza fosse como uma grande luz no céu que brilhasse fora dos seres humanos, então a sua pergunta faria sentido. Ainda assim, haveria a questão de como os humanos perceberam essa luz. A evolução não tem nada como a inteligência ou a precisão necessária para ajustar exatamente o seu sistema de metas. Ao produzir as primeiras mentes verdadeiras, o simples critério de aptidão da evolução se estilhaçou em mil valores. Os humanos evoluíram com uma psicologia que atribui utilidade a coisas com as quais a evolução não se importa, como a vida humana e a felicidade. Então, quando olhamos para trás e dizemos: "Que maravilha!" Você fica maravilhado e se pergunta o fato de que seus valores coincidem com eles mesmos." Mas então – ainda é surpreendente que este circuito circular específico, e não algum outro circuito, tenha surgido no mundo. Que nos encontremos elogiando o amor e não o ódio, a beleza e não a feiura.

"Não acho que você está me entendendo. Para você parece natural atribuir especial importância à beleza e ao altruísmo porque os valoriza muito. E não vê isso como um fato incomum em si, uma vez que muitos dos seus amigos pensam da mesma forma. Então, espera-se que um fantasma do vazio perfeito também valorize a vida e a felicidade. E, do ponto de vista fora da realidade, uma grande coincidência teria realmente ocorrido."

Mas é possível argumentar a favor da importância da beleza e do altruísmo a partir dos primeiros princípios - que os nossos sentidos estéticos nos levam a criar uma nova complexidade, em vez de repetir as mesmas coisas indefinidamente; e que o altruísmo é importante porque nos leva para fora de nós mesmos, dando à nossa vida um significado mais elevado do que o puro egoísmo bruto.

"Ah, mas será que esse argumento irá mover até mesmo um fantasma do vazio perfeito, agora que apelou a valores ligeiramente diferentes? Esses não são primeiros princípios, são apenas princípios diferentes. Mesmo que tenha adotado um tom filosófico pretensioso, ainda não há argumentos universalmente convincentes. Tudo o que fizemos foi passar a bola recursiva pra frente."

Não acha que, de alguma forma, evoluímos para acessar algo além…

"De que adianta supor algo além? Por que deveríamos prestar mais atenção a essa coisa além do que prestamos à nossa existência como humanos? Como isso altera sua responsabilidade pessoal, dizer que você estava apenas seguindo as ordens da coisa além? E você ainda teria evoluído para deixar a coisa além, ao invés de outra coisa, dirigir suas ações. Acima de tudo, seria coincidência demais."

Muita coincidência?

"Uma flor é linda, você diz. Você acha não haver história por trás dessa beleza, ou que a ciência não conhece a história? O pólen das flores é transmitido pelas abelhas, portanto, por seleção sexual, as flores evoluíram para atrair as abelhas - imitando certos sinais de acasalamento das abelhas, como aconteceu; os padrões das flores pareceriam mais complicados, se você pudesse ver no ultravioleta. Agora, flores saudáveis são um sinal de terra fértil, com probabilidade de dar frutos e outros tesouros, e provavelmente também de caçar animais; então é de se admirar que os humanos evoluíram para serem atraídos por flores? Mas para haver alguma grande luz escrita sobre as próprias estrelas - aquelas enormes bolas inconscientes de hidrogênio em chamas - que também diziam que as flores eram lindas, agora isso seria coincidência demais."

Então você explica a beleza de uma flor?

"Não, deixe-me explicar. Claro que há uma história por trás da beleza das flores e do fato de achá-las lindas. Por trás de eventos ordenados, encontram-se histórias ordenadas. O que não tem história é produto

do ruído aleatório, o que dificilmente é melhor. Se você não consegue se alegrar com as coisas que têm histórias por trás delas, sua vida será realmente vazia. Acho que não sinto menos alegria em uma flor do que você, talvez até mais, porque também me alegro com sua história."

É verdade que, do ponto de vista causal, não há surpresa - nenhuma interrupção da ordem física do universo. Mas ainda me parece que, na criação humana pela evolução, ocorreu algo precioso e maravilhoso. Se não podemos chamá-lo de milagre físico, chamemo-lo de milagre moral.

"Porque é apenas um milagre do ponto de vista da moralidade que foi produzido, explicando assim toda a aparente coincidência de uma perspectiva meramente causal e física?"

Bem… Suponho que você possa interpretar o termo dessa maneira, sim. Eu apenas quis dizer algo que foi imensamente surpreendente e maravilhoso no nível moral, mesmo que não seja surpreendente no nível físico.

"Acho que foi isso que eu disse."

Mas ainda me parece que você, do seu próprio ponto de vista, tira algo dessa maravilha.

"Então você tem problemas para se alegrar com o meramente real. O amor tem que começar de alguma forma, tem que entrar no universo em algum lugar. É como perguntar como a própria vida começa - e embora você tenha nascido de seu pai e sua mãe, e eles tenham surgido de seus pais vivos, se você for muito, muito e muito longe, você finalmente chegará a um replicador que surgiu por puro acidente - a fronteira entre a vida e a não-vida. Assim também é com o amor.

"Um padrão complexo deve ser explicado por uma causa que ainda não é esse padrão complexo. Não basta explicar o evento, é preciso explicar também a forma em si. Para o amor existir antes do tempo, ele deve vir de algo que não seja amor; caso contrário, não seria possível o amor.

"Mesmo que a própria vida exija que o primeiro replicador surja por acidente, sem pais, ainda assim é causado: há muito tempo, lá atrás na cadeia causal que levou você a existir, há 3,85 bilhões de anos, em uma pequena poça de maré.

"Talvez os filhos de seus filhos perguntem como é possível amar.

"E seus pais dirão: Porque nós, que também amamos, criamos você para amar. "E os filhos de seus filhos perguntarão: Mas como você ama?

"E seus pais responderão: Porque nossos próprios pais, que também amaram, nos criaram para amar por sua vez.

"Então, os filhos de seus filhos perguntarão: Mas onde tudo começou? Onde termina a recursão?

"E seus pais dirão: Era uma vez, há muito tempo e muito longe, havia seres inteligentes que não foram criados de forma inteligente. Era uma vez amantes criados por algo que não amava.

"Era uma vez, quando toda a civilização era uma única galáxia, uma única estrela e um único planeta, um lugar chamado Terra.

"Há muito tempo e muito longe, muito tempo atrás."

Parte W - Humanismo quantificado

281 - Insensibilidade ao escopo

Era uma vez... perguntou-se a três grupos de indivíduos quanto estariam dispostos a pagar para salvar 2.000/20.000/200.000 aves migratórias do afogamento em lagoas de petróleo descobertas. Os grupos responderam respectivamente com US$ 80, US$ 78 e US$ 88 [1]. Isso evidencia insensibilidade ou negligência em relação ao escopo da ação altruísta, pois o número de aves salvas teve pouco efeito na disposição para pagar.

Experiências semelhantes indicaram que os residentes de Toronto pagariam pouco mais para limpar todos os lagos poluídos em Ontário do que para limpar os lagos poluídos em uma região específica de Ontário [2], ou que os residentes de quatro estados do oeste dos EUA pagariam apenas 28% a mais para proteger todas as 57 áreas selvagens desses estados do que para proteger uma única área [3]. As pessoas têm em mente a imagem de "um único pássaro exausto, com as penas encharcadas em óleo negro, incapaz de escapar" [4]. Esse protótipo evoca um nível de excitação emocional que é o principal impulsionador da disposição para pagar, e a imagem é a mesma em todos os casos. Quanto ao escopo, ele é frequentemente ignorado – nenhum ser humano consegue visualizar 2.000 pássaros ao mesmo tempo, muito menos 200.000. A conclusão comum é que aumentos exponenciais no escopo resultam em aumentos lineares na disposição para pagar – talvez correspondendo ao tempo linear necessário para nossos olhos processarem os zeros; essa pequena quantidade de afeto é adicionada, e não multiplicada, ao afeto do protótipo. Essa hipótese é conhecida como "avaliação por protótipo".

Uma hipótese alternativa é a "compra de satisfação moral". As pessoas gastam dinheiro suficiente para criar um brilho caloroso em si mesmas, uma sensação de terem cumprido seu dever. O nível de gastos necessário para alcançar esse brilho quente depende da personalidade e da situação financeira, mas certamente não está relacionado ao número de pássaros.

A insensibilidade ao escopo persiste mesmo quando vidas humanas estão em jogo: aumentar o suposto risco de água potável clorada de 0,004 para 2,43 mortes anuais por 1.000 – um aumento de 600 vezes – elevou a disposição para pagar de US$ 3,78 para US$ 15,23 [5], Baron e Greene não encontraram nenhum efeito em várias vidas salvas por um fator de 10 [6]. Um artigo intitulado "Insensibilidade ao valor da vida humana: um estudo de entorpecimento psicofísico" coletou evidências de que nossa percepção das mortes humanas segue a Lei de Weber – obedece a uma escala logarítmica onde a "justa diferença notável" é uma fração constante do todo. Um programa de saúde proposto para salvar as vidas dos refugiados ruandeses obteve um apoio muito maior quando prometeu salvar 4.500 vidas em um campo de 11.000 refugiados, em vez de 4.500 em um campo de 250.000. Uma potencial cura para uma doença precisava prometer salvar muito mais vidas para ser considerada digna de financiamento, se a doença fosse originalmente declarada como tendo matado 290.000 em vez de 160.000 ou 15.000 pessoas por ano [7].

Moral da história: se você deseja ser um altruísta eficaz, é necessário pensar com cuidado usando a parte do seu cérebro que processa aqueles zeros escuros e desinteressantes no papel, e não apenas com a parte que se preocupa com aquele pobre pássaro encharcado de óleo.

Referências

[1] William H. Desvousges et al., Measuring Nonuse Damages Using Contingent Valuation: An Experimental Evaluation of Accuracy, technical report (Research Triangle Park, NC: RTI International, 2010), doi:10.3768/rtipress.2009.bk.0001.1009.

[2] Daniel Kahneman, "Comments by Professor Daniel Kahneman," in Valuing Environmental Goods: An Assessment of the Contingent Valuation Method, ed. Ronald G. Cummings, David S. Brookshire, and William D. Schulze, vol. 1.B, Experimental Methods for Assessing Environmental Benefits (Totowa, NJ: Rowman & Allanheld, 1986), 226–235, http://yosemite.epa.gov/ee/epa/eerm.nsf/vwAN/EE-0280B-04.pdf/$file/EE-0280B-04.pdf.

[3] Daniel L. McFadden and Gregory K. Leonard, "Issues in the Contingent Valuation of Environmental Goods: Methodologies for Data Collection and Analysis," in Contingent Valuation: A Critical Assessment, ed. Jerry A. Hausman, Contributions to Economic Analysis 220 (New York: North-Holland, 1993), 165–215, doi:10.1108/S0573-8555(1993)0000220007.

[4] Kahneman, Ritov, and Schkade, "Economic Preferences or Attitude Expressions?"

[5] Richard T. Carson and Robert Cameron Mitchell, "Sequencing and Nesting in Contingent Valuation Surveys," Journal of Environmental Economics and Management 28, no. 2 (1995): 155–173, doi:10.1006/jeem.1995.1011.

[6] Jonathan Baron and Joshua D. Greene, "Determinants of Insensitivity to Quantity in Valuation of Public Goods: Contribution, Warm Glow, Budget Constraints, Availability, and Prominence," Journal of Experimental Psychology: Applied 2, no. 2 (1996): 107–125, doi:10.1037/1076-898X.2.2.107.

[7] David Fetherstonhaugh et al., "Insensitivity to the Value of Human Life: A Study of Psychophysical Numbing," Journal of Risk and Uncertainty 14, no. 3 (1997): 283–300, doi:10.1023/A:1007744326393.

282 - Uma vida contra o mundo

Quem salva uma única vida é como se tivesse salvado o mundo inteiro.

—The Talmud[20], Sanhedrin 4:5

É um pensamento encantador, não é verdade? Sinta esse caloroso brilho.

Posso atestar que auxiliar uma pessoa é tão gratificante quanto auxiliar o mundo inteiro. Houve uma vez, quando eu estava exausto durante o dia e perdendo tempo na Internet - é um pouco complicado, mas essencialmente, consegui impactar a vida de alguém deixando um comentário anônimo em um blog. Não esperava que tivesse um efeito tão significativo, mas teve. Ao descobrir o impacto que causei, experimentei uma sensação tremenda. A euforia perdurou durante aquele dia e noite, diminuindo apenas um pouco pela manhã seguinte. Foi tão gratificante (e aqui está a parte surpreendente) quanto a euforia de uma grande descoberta científica, que anteriormente servia como minha melhor referência para a experiência de ter percepções, comparável ao uso de drogas.

Preservar uma vida provavelmente é tão recompensador quanto ser a primeira pessoa a compreender o que faz as estrelas brilharem. Provavelmente é tão gratificante quanto salvar o mundo inteiro.

Mas, caro leitor, se você tiver a opção entre salvar uma única vida e salvar o mundo inteiro, por favor, escolha salvar o mundo. Porque, além desse caloroso brilho, há uma diferença colossal. Para algumas pessoas, a noção de que salvar o mundo é significativamente superior a salvar uma vida humana é óbvia, assim como afirmar que seis bilhões de dólares valem mais do que um dólar, ou que seis quilômetros cúbicos de ouro pesam mais do que um metro cúbico de ouro. (E não importa o valor esperado para as gerações futuras.) Por que isso pode não ser óbvio? Bem, suponha que exista um dever qualitativo de preservar todas as vidas possíveis – então, alguém que salva o mundo e alguém que salva uma vida humana estão simplesmente cumprindo o mesmo dever. Ou suponhamos que sigamos a concepção grega de virtude pessoal, em vez do consequencialismo; alguém que salva o mundo é virtuoso, mas não seis bilhões de vezes mais virtuoso do que alguém que salva uma vida humana. Ou talvez o valor de uma vida humana já seja tão grande que não podemos compreendê-lo – de modo que a dor passageira que sentimos nos funerais é uma subestimação infinitesimal do que se perde – e, assim, passar para o mundo inteiro muda pouco.

Concordo que uma vida humana tem um valor inimaginavelmente alto. Também defendo que duas vidas humanas são duas vezes mais inimaginavelmente valiosas. Ou dito de outra forma: quem salva uma vida, é como se tivesse salvo o mundo inteiro; quem salva dez vidas é como se tivesse salvo dez mundos. Quem quer que realmente salve o mundo inteiro – não deve ser confundido com a pretensão retórica de salvar o mundo – é como se tivesse preservado uma civilização intergaláctica.

Duas crianças surdas dormem nos trilhos da ferrovia, o trem se aproxima rapidamente; você vê isso, mas está muito longe para salvar as crianças. Estou por perto, ao meu alcance, então dou um salto para frente e retiro uma criança dos trilhos do trem – e então paro, desfrutando calmamente de uma Pepsi Diet enquanto

20 NT. O **Talmud** é um texto central do Judaísmo Rabínico que contém discussões e interpretações da Lei Judaica, ética, filosofia e história. É composto por duas partes principais: a Mishná (a lei oral) e a Gemara (comentários sobre a Mishná). O Talmud é uma fonte rica de sabedoria e conhecimento judaico, estudado e debatido por estudiosos há séculos.

o trem se aproxima da segunda criança. “Rápido!” você grita para mim. “Faça algo!” Mas (respondo) eu já salvei uma criança dos trilhos do trem e, portanto, estou “inimaginavelmente” à frente em alguns pontos. Não tenho mais motivos para agir. Não parece correto, não é verdade?

Por que seria diferente se um filantropo gastasse US$ 10 milhões na cura de uma doença rara, mas espetacularmente fatal, que afeta apenas cem pessoas em todo o planeta, quando o mesmo dinheiro teria igual probabilidade de produzir a cura para uma doença menos espetacular, que mata 10% da população? 100.000 pessoas? Eu não acho que seria diferente. Quando vidas humanas estão em jogo, temos o dever de maximizar e não apenas satisfazer; e este dever tem a mesma força que o dever original de preservar vidas. Quem escolhe conscientemente salvar uma vida, quando poderia ter salvo duas – para não mencionar mil vidas ou um mundo –, se condenou tão completamente quanto qualquer assassino.

Não é fácil, do ponto de vista cognitivo, gastar dinheiro para salvar vidas, uma vez que os métodos clichês que vêm instantaneamente à mente não funcionam ou são contraproducentes. (Explicarei posteriormente por que isso tende a acontecer.) Stuart Armstrong também observa que, se quisermos menosprezar o filantropo que gasta dinheiro para salvar vidas de maneira ineficiente, devemos ser consistentes e menosprezar ainda mais aqueles que têm a capacidade de gastar dinheiro para salvar vidas. Mas não.

283 - O Paradoxo de Allais

Escolha entre as duas opções a seguir:

1A. US$ 24.000, com certeza.

1B. Uma chance de 33/34 de ganhar U$ 27.000 e 1/34 de chance de não ganhar nada.

O que parece mais intuitivamente atraente? E qual você escolheria na vida real? Agora, qual dessas duas opções você preferiria intuitivamente e qual escolheria na vida real?

2A. 34% de chance de ganhar US$ 24.000 e 66% de chance de não ganhar nada.

2B. 33% de chance de ganhar US$ 27.000 e 67% de chance de não ganhar nada.

O Paradoxo de Allais — como Allais o chamou, embora não seja realmente um paradoxo — foi um dos primeiros conflitos entre a teoria da decisão e o raciocínio humano a ser exposto experimentalmente, em 1953 [1]. Modifiquei-o ligeiramente para facilitar a matemática, mas o problema essencial é o mesmo: a maioria das pessoas prefere 1A a 1B e a maioria prefere 2B a 2A. Na verdade, em comparações nos sujeitos, a maioria dos sujeitos expressa ambas as preferências simultaneamente.

Isso é um problema porque os 2 equivalem a um terço de chance de jogar os 1. Ou seja, 2A equivale a jogar 1A com 34% de probabilidade, e 2B equivale a jogar 1B com 34% de probabilidade.

Entre os axiomas usados para provar que tomadores de decisão "consistentes" podem ser vistos como maximizadores da utilidade esperada está o Axioma da Independência: se X é estritamente preferido a Y, então uma probabilidade P de X e (1 − P) de Z deve ser estritamente preferida para P chance de Y e (1 − P) chance de Z.

Todos os axiomas são consequências, bem como antecedentes, de uma função de utilidade consistente. Portanto, deve ser possível provar que os sujeitos experimentais acima não podem ter uma função de utilidade consistente sobre os resultados. E, de fato, você não pode ter simultaneamente:

$$U(\text{US\$ } 24.000) > (33/34) \times U(\text{US\$ } 27.000) + (1/34) \times U(\text{US\$ } 0)$$

$$0{,}34 \times U(\text{US\$ } 24.000) + 0{,}66 \times U(\text{US\$ } 0) < 0{,}33 \times U(\text{US\$ } 27.000) + 0{,}67 \times U(\text{US\$ } 0).$$

Estas duas equações são algebricamente inconsistentes, independentemente de U, portanto o Paradoxo de Allais não tem nada a ver com a utilidade marginal decrescente do dinheiro.

Maurice Allais inicialmente defendeu as preferências reveladas dos sujeitos experimentais – ele viu o experimento como uma exposição de uma falha nas ideias convencionais de utilidade, em vez de uma falha na psicologia humana. Afinal, estávamos em 1953, e o movimento de heurísticas e preconceitos só começaria realmente nas próximas duas décadas. Allais pensou que seu experimento apenas mostrava que o Axioma da Independência claramente não era uma boa ideia na vida real.

(Quão ingênua, quão tola, quão simplista é a teoria da decisão Bayesiana…)

Certamente a certeza de ter US$ 24.000 deveria contar para alguma coisa. Você pode sentir a diferença, certo? A garantia sólida?

(Estou começando a pensar nisso como "realismo filosófico ingênuo" – supondo que nossas intuições expõem diretamente verdades sobre quais estratégias são mais sábias, como se fosse um fato diretamente percebido que "1A é superior a 1B". As intuições expõem diretamente verdades sobre as funções cognitivas humanas, e apenas indiretamente expõe (depois de refletirmos sobre as próprias funções cognitivas) verdades sobre a racionalidade.)

"Mas vamos lá", você diz, "é realmente uma coisa tão terrível afastar-se da beleza bayesiana?" Ok, então os sujeitos não seguiram o pequeno "axioma da independência" defendido por gente como von Neumann e Morgenstern. No entanto, quem disse que as coisas devem estar limpas e arrumadas?

Por que nos preocupar com a elegância, se ela nos faz correr riscos que não queremos? A utilidade esperada diz-nos que devemos atribuir algum tipo de número a um resultado e depois multiplicar esse valor pela probabilidade do resultado, somá-los, etc. Por que não criar regras mais palatáveis?

Sempre há um preço para sair do Caminho Bayesiano. É disso que tratam os teoremas de coerência e unicidade.

Neste caso, se um agente preferir 1A a 1B e 2B a 2A, isso introduz uma forma de reversão de preferência – uma inconsistência dinâmica no planejamento do agente. Você se torna uma bomba de dinheiro.

Suponha que às 12h00. Eu jogo um dado de cem lados. Se o dado mostrar um número maior que 34, o jogo termina. Caso contrário, às 12h05. Consulto um switch com duas configurações, A e B. Se a configuração for A, pago US$24.000. Se a configuração for B, jogo um dado de 34 lados e pago US$27.000, a menos que o dado mostre "34", caso em que não pago nada.

Digamos que você prefira 1A a 1B e 2B a 2A e pagaria um único centavo para satisfazer cada preferência. A mudança começa no estado A. Antes das 12h, você me paga um centavo para mudar a chave para B. O dado sai 12. Depois das 12h. e antes das 12h05, você me paga um centavo para mudar para A. Aceitei sua opinião sobre o assunto. Se você ceder às suas intuições e descartar a mera elegância como uma obsessão inútil por limpeza, não se surpreenda quando seus centavos lhe forem tirados...

(Acho que a mesma falha em desvalorizar proporcionalmente o impacto emocional das pequenas probabilidades é responsável pela loteria.)

Referências

[1] Maurice Allais, "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine," Econometrica 21, no. 4 (1953): 2, doi:10.2307/1907921; Daniel Kahneman and Amos Tversky, "Prospect Theory: An Analysis of Decision Under Risk," Econometrica 47 (1979): 263–292.

284 - Zut Allais!

Hã! Eu não esperava que tantos comentaristas defendessem a inversão de preferências. Parece que me deparei com uma distância inferencial.

Provavelmente, compreender o Paradoxo de Allais fica mais fácil ao absorver melhor a Gestalt do campo das heurísticas e preconceitos, tais como:

- Os sujeitos experimentais tendem a defender preferências incoerentes mesmo quando elas são realmente tolas.
- As pessoas atribuem valores muito elevados a pequenas mudanças na probabilidade de 0 ou 1 (o efeito de certeza).

Vamos começar abordando a questão das preferências incoerentes – inversões de preferências, inconsistência dinâmica, bombas de dinheiro, esse tipo de coisa.

Qualquer pessoa familiarizada com um pouco de teoria do prospecto não terá dificuldade em construir casos nos quais as pessoas dizem que prefeririam jogar o jogo A em vez do jogo B; mas quando solicitadas a definir o preço das apostas, atribuem um valor mais alto à aposta B do que à aposta A. Diferentes características perceptivas se tornam evidentes quando você pergunta: “Qual você prefere?” em uma comparação direta e “Quanto você pagaria?” com um único item.

Essa escolha de apostas normalmente resulta em uma reversão de preferência:

1. 1/3 de chance de ganhar US$16 e 2/3 de chance de perder US$2.
2. Chance de 99/100 de ganhar US$4 e 1/100 de chance de perder US$1.

A maioria das pessoas prefere jogar 2 em vez de 1. Mas se você pedir para definirem os preços das apostas separadamente – solicitar um preço pelo qual seriam indiferentes entre ter aquela quantia de dinheiro e ter a chance de jogar – as pessoas vão apostar um preço mais alto em 1 do que em 2 [1].

Então, primeiro, você oferece a eles a oportunidade de participar da aposta 1, pelo preço declarado. Em seguida, você propõe a troca da aposta 1 pela aposta 2. Depois, você adquire a aposta 2 deles, pelo preço estipulado. Então, repete o processo. Daí a expressão “bomba de dinheiro”.

Ou, parafraseando Steve Omohundro: Se você prefere estar em Oakland do que em São Francisco, e prefere estar em San José do que em Oakland, e prefere estar em São Francisco do que em San José, você gastará uma quantia enorme de dinheiro em corridas de táxi.

Surpreendentemente, as pessoas defendem esses padrões de preferência. Alguns sujeitos abandonam-nos depois que o efeito da bomba de dinheiro é apontado – revisam seu preço ou reavaliam sua preferência –, mas alguns sujeitos os defendem.

Certa vez, jogadores de Las Vegas fizeram esse tipo de aposta com dinheiro real, usando uma roleta. E depois, um dos pesquisadores tentou explicar o problema da incoerência entre seus preços e suas escolhas.

Da transcrição [2] [3]:

SARAH LICHTENSTEIN: “Bem, e quanto à oferta para a Aposta A? Você tem alguma opinião diferente agora que sabe que está escolhendo um, mas oferecendo mais pelo outro?

SUJEITO: “É meio estranho, mas não, não tenho nenhum sentimento sobre isso. É apenas uma daquelas coisas. Isso mostra que meu processo de raciocínio não é tão bom, mas, fora isso, eu... sem escrúpulos.”

...

LICHTENSTEIN: “Posso persuadi-lo de que é um padrão irracional?”

SUJEITO: “Não, não acho que você provavelmente conseguiria, mas você poderia tentar.”

...

LICHTENSTEIN: “Bem, agora deixe-me sugerir o que se chama de jogo de bomba de dinheiro e experimentar isso com você para ver se você gosta. Se você acha que a aposta A vale 550 pontos [os pontos foram convertidos em dólares após o jogo, embora não numa base de um para um], então você deveria estar disposto a me dar 550 pontos se eu lhe der a aposta...”

...

LICHTENSTEIN: “Então você tem a aposta A, e eu digo: ‘Oh, você prefere a aposta B, não é?’”

...

SUJEITO: “Estou perdendo dinheiro.”

LICHTENSTEIN: “Vou comprar a aposta B de você. Serei generoso; Pagarei mais de 400 pontos. Vou te pagar 401 pontos. Você está disposto a me vender a aposta B por 401 pontos?”

SUJEITO: “Bem, certamente.”

...

LICHTENSTEIN: “Estou agora 149 pontos à frente.”

SUJEITO: “Esse é um bom raciocínio da minha parte. (risos) Quantas vezes passaremos por isso?”

...

LICHTENSTEIN: “Bem, acho que pressionei você tanto quanto sei como impedi-lo de realmente insultá-lo.”

SUJEITO: “Isso mesmo.”

Você quer gritar: “Desista já! A intuição nem sempre está certa!”

E há ainda a questão do estranho valor que as pessoas atribuem à certeza. Os meus livros estão embalados para a mudança, mas acredito que uma experiência mostrou que uma mudança de 100% de probabilidade para 99% de probabilidade pesava mais na mente das pessoas do que uma mudança de 80% de probabilidade para 20% de probabilidade.

O desafio de atribuir um valor extraordinário à certeza reside no fato de que a certeza em um momento é a probabilidade em outro momento.

No ensaio anterior, abordei o Paradoxo de Allais:

- 1A. US$ 24.000, com certeza.
- 1B. Chance de 33/34 de ganhar US$ 27.000 e 1/34 de chance de não ganhar nada.
- 2A. 34% de chance de ganhar US$ 24.000 e 66% de chance de não ganhar nada.

- 2B. 33% de chance de ganhar US$ 27.000 e 67% de chance de não ganhar nada.

A preferência ingênua no Paradoxo de Allais é 1A > 1B e 2B > 2A. Assim, você me pagaria para trocar de A para B, pois prefere ter 33% de chance de ganhar US$ 27.000 em vez de 34% de chance de ganhar US$ 24.000. Então, um lançamento de dados elimina uma parte da probabilidade. Em ambos os casos, você tinha pelo menos 66% de chance de não ganhar nada. Este lançamento de dados elimina esses 66%. Portanto, agora a opção B tem uma probabilidade de 33/34 de ganhar US$ 27.000, mas a opção A tem a garantia de ganhar US$ 24.000. Oh, gloriosa certeza! Assim, você me pagaria para mudar de B para A.

Agora, se eu lhe dissesse antecipadamente que faria tudo isso, você realmente gostaria de me pagar para acionar o interruptor e depois pagar novamente para desfazê-lo? Ou preferiria reconsiderar?

Sempre que tentar avaliar uma mudança de probabilidade de 24% para 23% como sendo menos significativa do que uma mudança de ~1 para 99%— sempre que tentar atribuir mais valor a um aumento de probabilidade próximo ao final da escala— estará suscetível a esse tipo de exploração. Posso sempre criar uma série de eventos que eliminem a massa de probabilidade, um pouco de cada vez, até que reste apenas uma "certeza" que reverta suas preferências. A certeza de uma época é a incerteza de outra, e se insistir em tratar a distância de ~1 a 0,99 como especial, posso fazer com que suas preferências se invertam ao longo do tempo e obter algum benefício financeiro.

Posso persuadi-lo, talvez, de que este é um padrão irracional?

Certamente, se você estiver lendo este livro há algum tempo, perceberá que você - o próprio sistema e processo que lê estas palavras - é uma máquina com defeitos. Suas intuições não fornecem informações diretas e verídicas sobre boas escolhas. Se não acredita nisso, há alguns jogos de azar que eu gostaria de jogar com você.

Existem vários outros jogos que também podem ser explorados com efeitos de certeza. Por exemplo, se oferecer a alguém a certeza de US$ 400, ou uma probabilidade de 80% de US$ 500 e uma probabilidade de 20% de US$ 300, essa pessoa geralmente aceitará os US$ 400. Mas se pedir às pessoas que se imaginem US$ 500 mais ricas e perguntar se prefeririam uma certa perda de US$ 100 ou uma chance de 20% de perder US$ 200 [4], geralmente arriscarão perder US$ 200. Mesma distribuição de probabilidade sobre resultados, descrições diferentes, escolhas diferentes.

Sim, Virginia, realmente deveria considerar multiplicar a utilidade dos resultados pela sua probabilidade. Não hesite em usar matemática clara. No paradoxo de Allais, descubra se 1 unidade da diferença entre receber US$ 24.000 e não receber nada supera 33 unidades da diferença entre receber US$ 24.000 e US$ 27.000. Se isso acontecer, prefira 1A a 1B e 2A a 2B. Se as 33 unidades superarem 1 unidade, prefira 1B a 1A e 2B a 2A. Quanto ao cálculo da utilidade do dinheiro, sugiro usar uma aproximação que assume que o dinheiro tem utilidade logarítmica. Se já possui bastante dinheiro, escolha B. Se US$ 24.000 dobrariam seus ativos existentes, escolha A. Caso 2 ou caso 1, não faz diferença. Ah, e certifique-se de avaliar a utilidade dos valores totais dos ativos - a utilidade dos estados finais dos resultados do mundo - e não as mudanças nos ativos, ou acabará sendo inconsistente novamente.

Vários comentadores afirmaram que o padrão de preferência não era irracional devido à 'utilidade da certeza' ou algo semelhante. Um comentarista até incluiu U(Certeza) em uma equação de utilidade esperada.

Lembra-se daquela história toda sobre a utilidade esperada e a utilidade serem fundamentalmente diferentes? Os serviços públicos superam os resultados. São valores que você atribui a estados sólidos e específicos do mundo. Não faz sentido inserir uma probabilidade de 1 em uma função de utilidade.

Antes de torcer o nariz e pensar 'Hmph... você só quer que a matemática esteja limpa e organizada', lembre-se de que, nesse caso, o preço de abandonar o Caminho Bayesiano foi pagar a alguém para acionar um interruptor e depois devolvê-lo.

Mas e aquela sensação sólida e calorosa de segurança? Isso não é um utilitário? Isso é ser humano. Os humanos não são maximizadores de utilidade esperada. Se você deseja relaxar e se divertir ou pagar um extra por uma sensação de certeza, depende se você se preocupa mais em satisfazer suas intuições ou em realmente alcançar o objetivo.

Se você está jogando em Las Vegas para se divertir, então não pense na utilidade esperada – você perderá dinheiro de qualquer maneira.

Mas e se fossem 24 mil vidas em jogo, em vez de 24 mil dólares? O efeito de certeza é ainda mais forte nas vidas humanas. Você pagará uma vida humana para acionar o interruptor e outra para desligá-lo?

Tolerar reversões de preferências zomba das reivindicações de otimização. Se você dirigir de San José a São Francisco, de Oakland a San José, repetidamente, poderá sentir muitos sentimentos confusos e calorosos, mas não pode ser interpretado como tendo um destino - como tentar ir em algum lugar.

Quando você tem preferências circulares, você não está guiando o futuro – apenas andando em círculos. Se você gosta de correr por si só, tudo bem. Mas se você tem um objetivo – algo que está realmente tentando realizar – uma inversão de preferências revela um grande problema. Pelo menos uma das escolhas que você está fazendo não deve funcionar para realmente otimizar o futuro em qualquer sentido coerente.

Se o que importa para você é a sensação calorosa e confusa de certeza, então tudo bem. Se a vida de alguém está em jogo, então é melhor você perceber que suas intuições são lentes gordurosas através das quais você pode ver o mundo. Seus sentimentos não lhe fornecem informações diretas e verídicas sobre as consequências estratégicas – parece que sim, mas não são. Fuzzies calorosos podem desviá-lo para muito longe.

Existem leis matemáticas que regem estratégias eficientes para orientar o futuro. Quando algo realmente importante está em jogo – algo mais importante do que seus sentimentos de felicidade com a decisão – então você deve se preocupar com a matemática, se é que realmente se importa.

Referências

[1] Sarah Lichtenstein and Paul Slovic, “Reversals of Preference Between Bids and Choices in Gambling Decisions,” Journal of Experimental Psychology 89, no. 1 (1971): 46–55.

[2] William Poundstone, Priceless: The Myth of Fair Value (and How to Take Advantage of It) (Hill & Wang, 2010).

[3] Sarah Lichtenstein and Paul Slovic, eds., The Construction of Preference (Cambridge University Press, 2006).

[4] Kahneman and Tversky, “Prospect Theory: An Analysis of Decision Under Risk.”

285 - Sentindo-se moral

Suponhamos que uma doença, um monstro, uma guerra ou algo do tipo esteja tirando vidas. Agora, imagine que você só tem recursos suficientes para implementar uma das duas opções a seguir:

1. Salvar 400 vidas, com certeza.
2. Salvar 500 vidas com 90% de probabilidade; não salvar nenhuma vida com 10% de probabilidade.

A maioria das pessoas opta pela primeira opção. No entanto, penso que isso é uma tolice. Se multiplicarmos 500 vidas por 90% de probabilidade, obteremos um valor esperado de 450 vidas, superando as 400 vidas da opção 1. (As vidas salvas não perdem utilidade marginal, portanto, esse é um cálculo adequado.)

"O quê!" você exclama indignado. "Como você pode brincar com vidas humanas? Como pode pensar em números quando tanto está em jogo? E se aquela probabilidade de 10% acontecer e todos perecerem? Chega dessa maldita lógica! Você está seguindo sua racionalidade até o precipício!

No entanto, aqui está algo interessante. Se apresentarmos as opções da seguinte maneira:

1. 100 pessoas morrem, com certeza.
2. 90% de chance de ninguém morrer; 10% de chance de 500 pessoas morrerem.

Dessa vez, a maioria escolheria a opção 2, mesmo ela sendo a mesma aposta. Assim como a certeza de salvar 400 vidas parece mais reconfortante do que um ganho incerto, uma perda certa parece pior do que uma incerta.

Podemos também destacar na segunda descrição: "Como podemos condenar 100 pessoas à morte certa quando há uma boa chance de salvá-las? Todos compartilharemos o risco! Mesmo que haja apenas 75% de chance de salvar todos, ainda valeria a pena - caso haja uma chance - todos conseguem, ou ninguém consegue!"

E sabe de uma coisa? Não se trata dos seus sentimentos. Uma vida humana, com todas as suas alegrias e dores acumuladas ao longo das décadas, vale muito mais do que os sentimentos de conforto ou desconforto do seu cérebro com um plano. Calcular a utilidade esperada pode parecer muito frio para o seu gosto? Bem, esse sentimento não é nada comparado a uma vida em jogo. Apenas cale a boca e multiplique.

Um googol é 10^100 - um 1 seguido por cem zeros. Um Googleplex é um número ainda mais incompreensivelmente grande - é 10^googol, um 1 seguido por um googol de zeros. Agora, considere alguns inconvenientes triviais, como um soluço, e algum infortúnio decididamente nada trivial, como ser lentamente dilacerado, membro por membro, por sádicos tubarões mutantes. Se nos forçarem a escolher entre prevenir os soluços de uma pessoa Googleplex ou prevenir o ataque de tubarão de uma única pessoa, que escolha devemos fazer? Se atribuirmos qualquer valor negativo aos soluços, então, sob pena de incoerência da teoria da decisão, deve haver um certo número de soluços que se somariam para rivalizar com o valor negativo de um ataque de tubarão. Para qualquer mal finito específico, deve haver alguns contratempos que seriam ainda piores.

Dilemas morais como esses não são apenas conceitos filosóficos para entreter os filósofos analíticos em jantares. São versões destiladas dos tipos de situações que realmente enfrentamos todos os dias. Devo

gastar US$ 50 em um jogo de console ou doar tudo para caridade? Devo organizar uma arrecadação de fundos de 700.000 dólares para pagar um único transplante de medula óssea, ou devo usar esse mesmo dinheiro em redes mosquiteiras e evitar a morte de cerca de 200 crianças por malária?

No entanto, muitos desviam o olhar da abundância de trocas morais desagradáveis que existem no mundo real - muitos até se orgulham de desviar o olhar. A pesquisa mostra que as pessoas distinguem "valores sagrados", como vidas humanas, de "valores não sagrados", como dinheiro. Quando se tenta negociar um valor sagrado com um valor não sagrado, os sujeitos expressam grande indignação. (Às vezes, querem punir a pessoa que fez a sugestão.)

Minha anedota favorita sobre esse tema vem de uma equipe de pesquisadores que avaliou a eficácia de um determinado projeto, calculando o custo por vida salva, e recomendou ao governo que o implementasse porque era rentável. A agência governamental rejeitou o relatório, argumentando que não era possível atribuir um valor monetário à vida humana. Após rejeitar o relatório, o órgão decidiu não implementar a medida.

Trocar um valor sagrado por um valor não sagrado é verdadeiramente terrível. Multiplicar apenas as utilidades seria demasiado impessoal - seria seguir a racionalidade até o precipício...

Mas o altruísmo não é o sentimento caloroso e confuso que você sente ao ser altruísta. Se você está fazendo isso para o benefício espiritual, isso não passa de egoísmo. O principal é auxiliar os outros, seja qual for o meio. Então, cale a boca e multiplique! E se parecer que há uma ferocidade nesta maximização, como a espada nua da lei, ou a queima do Sol - se parecer que no centro dessa racionalidade há uma pequena chama fria -

Bem, a outra maneira pode parecer melhor dentro de você. Mas não funcionaria.

E eu também digo isto: se você deixar de lado o seu arrependimento por toda a satisfação espiritual que poderia estar tendo - se você seguir o Caminho de todo o coração, sem pensar que está sendo enganado - se você se entregar à racionalidade sem se conter, descobrirá que a racionalidade lhe dá em troca.

Mas essa parte só funciona se você não sair por aí dizendo para si: "Seria melhor dentro de mim se eu pudesse ser menos racional". Você deveria ficar triste por ter a oportunidade de realmente ajudar as pessoas? Você não poderá atingir todo o seu potencial se considerar seu presente um fardo.

286 - As “intuições” por trás do “utilitarismo”

Eu costumava ficar bastante confuso com metaética. Depois que finalmente esclareci minha confusão, fiz uma análise crítica dos meus pensamentos anteriores. Descobri que meu raciocínio moral ao nível de objeto era valioso, enquanto meu raciocínio moral ao nível meta era não apenas inútil, mas prejudicial. E isso parece uma síndrome geral – as pessoas se saem muito melhor discutindo se a tortura é boa ou má do que quando discutem o significado de “bom” e “mau”. Portanto, considero prudente manter as discussões morais no nível do objeto sempre que possível.

Às vezes, as pessoas se opõem a qualquer discussão sobre moralidade, alegando que a moralidade não existe. Em vez de explicar que «existir» não é o termo certo aqui, costumo perguntar: «Mas afinal, o que você faz?» e levar a discussão de volta ao nível do objeto.

No entanto, Paul Gowder destacou que tanto a ideia de escolher um inconveniente trivial do Googleplex em vez de uma atrocidade quanto a ideia de “utilitarismo” dependem da “intuição”. Ele afirma que argumentei que os dois não são compatíveis, mas me acusa de não defender as intuições utilitaristas às quais apelo.

Bem, “intuição” não é como eu descreveria os cálculos que fundamentam a moralidade humana e nos distinguem, como moralistas, de um filósofo ideal do vazio perfeito e/ou de uma rocha. No entanto, concordo em usar a palavra “intuição” como um termo artístico, tendo em mente que, neste sentido, “intuição” não deve ser contrastada com a razão, mas é, antes, o bloco de construção cognitivo a partir do qual tanto argumentos elaborados quanto argumentos perceptivos rápidos são construídos.

Vejo o projeto da moralidade como um projeto de renormalização da intuição. Temos intuições sobre coisas que parecem desejáveis ou indesejáveis, intuições sobre ações que são certas ou erradas, intuições sobre como resolver intuições conflitantes, intuições sobre como sistematizar intuições específicas em princípios gerais.

Exclua todas as intuições e você não terá um filósofo ideal de vazio perfeito; ficará apenas com uma pedra.

Mantenha todas as suas intuições específicas e recuse-se a desenvolver as reflexivas, e você não ficará com um filósofo ideal de perfeita espontaneidade e genuinidade; ficará apenas com uma pessoa das cavernas grunhindo e correndo em círculos, devido a preferências cíclicas e inconsistências semelhantes.

“Intuição”, como termo artístico, não é um palavrão quando se trata de moralidade – não há mais nada com que argumentar. Mesmo o modus ponens é uma “intuição” neste sentido – só que o modus ponens ainda parece uma boa ideia depois de ser formalizado, refletido, extrapolado para ver se tem consequências sensatas, etc.

Portanto, isso é “intuição”.

No entanto, Gowder não esclareceu o que quis dizer com “utilitarismo”. O utilitarismo afirma...

1. Que as ações corretas são rigidamente determinadas pelas boas consequências?
2. Que ações dignas dependem de expectativas justificáveis de boas consequências?
3. Que as probabilidades de consequências devem ser normativamente descontadas pela sua

probabilidade, de modo que uma probabilidade de 50% de algo ruim deveria pesar exatamente metade em nossas considerações?

4. Que ações virtuosas sempre correspondem à maximização da utilidade esperada sob alguma função de utilidade?

5. Que dois eventos prejudiciais são piores que um?

6. Que duas ocorrências independentes de um dano (não para a mesma pessoa, não interagindo uma com a outra) são exatamente duas vezes tão ruins quanto uma?

7. Que para quaisquer dois danos A e B, sendo A muito pior que B, existe alguma pequena probabilidade tal que apostar nesta probabilidade de A é preferível a uma certeza de B?

Se você afirmar que defendo algo ou que meu argumento depende de algo e que está equivocado, por favor, especifique o que é. De qualquer forma, aceito 3, 5, 6 e 7, mas não 4; não tenho certeza sobre a formulação de 1; e 2 é verdadeira, eu diria, mas formulada de maneira bastante solipsista e egoísta: você não precisa se preocupar em ser digno de louvor.

Agora, quais são as "intuições" nas quais se baseia meu "utilitarismo"? Este é um assunto mais profundo, mas darei uma olhada rápida nisso.

Em primeiro lugar, não é apenas que alguém me apresentou uma lista de afirmações como as acima, e eu decidi quais pareciam "intuitivas". Entre outras coisas, se tentarmos violar o "utilitarismo", deparamo-nos com paradoxos, contradições, preferências circulares e outras coisas que não são sintomas de injustiça moral, mas sim de incoerência moral.

Depois de refletir sobre problemas morais por um tempo, e também de descobrir novas verdades sobre o mundo, e até de descobrir fatos perturbadores sobre como você mesmo opera, muitas vezes você acaba com opiniões morais diferentes das que tinha no início. Isso não define exatamente o progresso moral, mas é como vivenciamos o progresso moral.

Como parte do meu progresso moral experiente, estabeleci uma distinção conceitual entre questões do tipo Para onde devemos ir? e questões do tipo Como devemos chegar lá? (Será que é isso que Gowder quis dizer quando me chamou de "utilitarista"?)

A questão de para onde leva uma estrada - para onde ela conduz - você pode responder percorrendo a estrada e descobrindo. Se você tem uma crença falsa sobre para onde a estrada leva, essa falsidade pode ser desfeita pela verdade de maneira muito direta e simples.

Quando se trata de querer ir a um lugar específico, esse desejo não está completamente imune aos poderes destrutivos da verdade. Você pode chegar lá e depois descobrir que se arrepende (o que não define o erro moral, mas é como vivenciamos o erro moral).

Contudo, mesmo assim, parece valer a pena distinguir entre o desejo de estar em um local específico e o desejo de seguir um caminho específico para chegar a um determinado lugar.

Nossas intuições sobre a direção a seguir são bastante questionáveis, mas nossas intuições sobre como chegar lá são sinceramente confusas. Após os duzentos e oitenta e sete estudos que demonstram que as pessoas estão dispostas a prejudicar a si mesmas se o problema for enquadrado de maneira incorreta, começamos a desconfiar das primeiras impressões.

Ao ler pesquisas suficientes sobre a insensibilidade ao escopo - as pessoas estão dispostas a pagar apenas 28% a mais para proteger todas as 57 áreas selvagens de Ontário do que para proteger uma única área, as pessoas estão dispostas a pagar a mesma quantia para salvar 50.000 vidas e 5.000 vidas... esse tipo de coisas...

Bem, o exemplo mais extremo de insensibilidade ao escopo que já encontrei foi descrito por Slovic aqui:

> Outra pesquisa recente apresenta resultados semelhantes. Dois psicólogos israelenses pediram às pessoas que

> contribuíssem para um tratamento caro que salvaria vidas. Elas poderiam oferecer essa contribuição a um grupo de oito crianças doentes ou a uma criança escolhida do grupo. O montante-alvo necessário para salvar a criança (ou crianças) era o mesmo em ambos os casos. As contribuições individuais para membros do grupo superaram significativamente as contribuições para o grupo na totalidade[21] [1].

Existem outras pesquisas semelhantes, mas estou apresentando apenas um exemplo, porque, você sabe, oito exemplos provavelmente teriam menos impacto.

Se você está familiarizado com o paradigma experimental geral, então a razão para o comportamento acima é bastante óbvia: concentrar sua atenção em uma única criança cria mais excitação emocional do que tentar distribuir a atenção entre oito crianças simultaneamente. Portanto, as pessoas estão dispostas a pagar mais para ajudar uma criança do que para ajudar oito.

Agora, você poderia analisar essa intuição e pensar que ela está revelando alguma verdade moral incrivelmente profunda, mostrando que a boa sorte de uma criança é de alguma forma desvalorizada pela boa sorte das outras crianças.

Mas e os bilhões de outras crianças no mundo? Por que não é uma má ideia ajudar esta criança, quando isso faz com que o valor de todas as outras crianças diminua? Como pode ser significativamente melhor ter 1.329.342.410 filhos felizes do que 1.329.342.409, mas um pouco pior ter mais sete, com 1.329.342.417?

Ou você poderia analisar isso e dizer: "A intuição está equivocada: o cérebro não consegue multiplicar por oito e obter uma quantidade maior do que quando começou. Mas deveria, do ponto de vista normativo."

E quando você percebe que o cérebro não pode multiplicar por oito, os outros casos de negligência de escopo deixam de parecer revelar alguma verdade fundamental sobre 50 mil vidas valerem exatamente o mesmo esforço que 5 mil vidas, ou algo do tipo. Você não tem a sensação de estar diante da revelação de uma verdade moral profunda sobre a ausência de utilidades acumulativas. Acontece que o cérebro não se multiplica. Quantidades são simplesmente jogadas pela janela.

Se você tem US$ 100 para gastar e investe US$ 20 em cada um dos cinco esforços para salvar 5.000 vidas, seu desempenho será inferior ao de investir US$ 100 em um único esforço para salvar 50.000 vidas. Da mesma forma, se essas escolhas forem feitas por 10 pessoas diferentes, e não pela mesma pessoa. Assim que você começa a acreditar que é melhor salvar 50 mil vidas do que 25 mil vidas, essa simples preferência pelos destinos tem implicações na escolha dos caminhos, quando se consideram cinco eventos distintos que salvam 5 mil vidas.

(É um princípio geral que os bayesianos não veem diferença entre a resposta de longo prazo e a resposta de curto prazo; você nunca obtém duas respostas diferentes calculando a mesma pergunta de duas maneiras diferentes. Mas o longo prazo é uma ferramenta útil de intuição, então estou abordando isso de qualquer maneira.)

A estratégia de avaliação agregativa de "calar a boca e multiplicar" surge da simples preferência por ter mais de algo – salvar tantas vidas quanto possível – ao descrever princípios gerais para escolher mais de uma vez, agir mais de uma vez, planejar ao mesmo tempo.

A agregação também decorre da afirmação de que a escolha local de salvar uma vida não depende de quantas vidas já existem, longe, do outro lado do planeta, ou longe, do outro lado do universo. Três vidas são uma, uma e uma. Não importa quantos bilhões estão melhores ou piores. 3 = 1 + 1 + 1, não importa quais outras quantidades você adicione a ambos os lados da equação. E se você adicionar mais uma vida, obterá 4 = 1 + 1 + 1 + 1. Isso é agregação.

Quando você leu pesquisa suficiente sobre heurísticas e preconceitos, e provas suficientes de coerência e unicidade para probabilidades bayesianas e utilidade esperada, e você viu os efeitos do 'efeito holandês' e da 'bomba de dinheiro' que penalizam a tentativa de lidar com resultados incertos de qualquer outra

21 NT. Texto original em inglês. *Other recent research shows similar results. Two Israeli psychologists asked people to contribute to a costly life-saving treatment. They could offer that contribution to a group of eight sick children, or to an individual child selected from the group. The target amount needed to save the child (or children) was the same in both cases. Contributions to individual group members far outweighed the contributions to the entire group.*

maneira, então você não vê as inversões de preferências no Paradoxo de Allais como reveladoras de alguma verdade moral incrivelmente profunda sobre o valor intrínseco da certeza. Isso apenas mostra que o cérebro não se multiplica.

As intuições perceptivas primitivas que fazem uma escolha 'sentir-se bem' não lidam muito bem com os caminhos probabilísticos ao longo do tempo, especialmente quando as probabilidades foram expressas simbolicamente em vez de experimentadas como uma frequência. Então, você reflete, elabora lógicas mais confiáveis e pensa em palavras.

Quando você vê pessoas insistindo que nenhuma quantia de dinheiro vale uma única vida humana e depois dirigindo mais um quilômetro para economizar US$ 10; ou quando você vê pessoas insistindo que nenhuma quantia de dinheiro vale uma redução na saúde, e então escolhem o seguro de saúde mais barato disponível; então você não acha que seus protestos revelam alguma verdade profunda sobre utilidades incomensuráveis.

Parte disso, claramente, é que as intuições primitivas não conseguem diminuir com sucesso o impacto emocional dos símbolos que representam pequenas quantidades – qualquer coisa sobre a qual você fale parece "uma quantia que vale a pena considerar".

E parte disso tem a ver com preferir regras sociais incondicionais a regras sociais condicionais. As regras condicionais parecem mais fracas, parecem mais sujeitas à manipulação. Se houver alguma brecha que permita ao governo cometer tortura legalmente, então o governo conduzirá um caminhão por essa brecha.

Portanto, parece que deveria haver uma injunção social incondicional contra preferir o dinheiro à vida, e nenhum 'mas' segui-lo. Nem mesmo 'mas mil dólares não valem 0,0000000001% de probabilidade de salvar uma vida'. Embora a última escolha, é claro, seja revelada toda vez que espirramos sem chamar um médico.

A retórica da sacralidade ganha pontos extras por parecer expressar um compromisso ilimitado, uma recusa incondicional que sinaliza confiabilidade e recusa de compromisso. Assim, conclui-se que a retórica moral defende distinções qualitativas, porque defender uma troca quantitativa soaria como se estivesse a conspirar para desertar.

Em situações como estas, as pessoas desejam fervorosamente lançar quantidades pela janela e se incomodam se tentamos trazer as quantidades de volta, pois soam como condições que enfraqueceriam a regra.

Contudo, não devemos concluir que, de fato, existem dois níveis de utilidade na ordenação lexical. Não devemos concluir que há um gradiente moral infinitamente acentuado, algum átomo que se move uma distância de Planck (em nosso universo físico contínuo) e envia uma utilidade de zero ao infinito. Não devemos concluir que as utilidades devem ser expressas usando números hiper-reais. Porque a camada inferior simplesmente desapareceria em qualquer equação. Nunca valeria o menor esforço recalcular isso. Todas as decisões seriam determinadas pela camada superior, e todo o pensamento seria gasto pensando apenas na camada superior, se a camada superior tivesse genuinamente prioridade lexical.

Como Peter Norvig observou uma vez, se os robôs de Asimov tivessem prioridade estrita para a Primeira Lei da Robótica ("Um robô não deve prejudicar um ser humano, nem por inação permitir que um ser humano sofra algum mal"), então o comportamento de nenhum robô jamais mostraria nenhum sinal das outras duas Leis; sempre haveria algum pequeno fator da Primeira Lei que seria suficiente para determinar a decisão.

Independentemente do valor que valha a pena considerar, deve valer a pena negociá-lo com todos os outros valores que também são dignos de consideração, porque o pensamento em si é um recurso limitado que deve ser negociado. Quando você revela um valor, você revela uma utilidade.

Não afirmo que a moralidade deva ser sempre simples. Já afirmei que o significado da música é mais do que apenas felicidade, mais do que apenas um centro de prazer aceso. Prefiro ver a música composta por pessoas do que por algoritmos de aprendizagem automática insensíveis, para que alguém tenha a alegria da composição; eu me importo com a jornada, assim como com o destino. E estou disposto a ouvir se você me

disser que o valor da música é mais profundo e envolve mais complicações do que imagino – que a avaliação deste evento seja mais complexa do que imagino.

Mas isso é para um evento específico. Quando se trata de multiplicar quantidades e probabilidades, as complicações devem ser evitadas – pelo menos se você se preocupa mais com o destino do que com a jornada. Quando você reflete sobre intuições suficientes e corrige absurdos suficientes, você começa a ver um denominador comum, um meta princípio em ação, que pode ser expresso como “Cale a boca e multiplique”.

No que diz respeito à música, preocupo-me com a jornada. Quando vidas estão em jogo, calo-me e multiplico-as.

É mais importante que vidas sejam salvas do que obedecermos a qualquer ritual específico para salvá-las. E o caminho ideal para esse destino é regido por leis que são simples, porque são matemáticas. E é por isso que sou um utilitarista – pelo menos quando estou fazendo algo que é esmagadoramente mais importante do que os meus próprios sentimentos sobre isso – o que acontece na maioria das vezes, já que não há muitos utilitaristas e muitas coisas ficam por fazer.

Referências

[1] Paul Slovic, “Numbed by Numbers,” Foreign Policy (March 2007), http://foreignpolicy.com/2007/ 03/13/ numbed-by-numbers/.

287 - Os fins não justificam os meios (entre humanos)

Se os fins não justificam os meios, o que justifica?

—atribuído à diversos autores

Penso que estou operando em um ambiente hostil.

—Justin Corwin

Os seres humanos podem ter desenvolvido uma estrutura de revolução política, inicialmente acreditando serem moralmente superiores à atual estrutura de poder corrupta, mas acabam sendo corrompidos pelo próprio poder - não por um plano consciente, mas pelo eco de ancestrais que fizeram o mesmo e assim perpetuaram esse ciclo.

Isso se encaixa no modelo:

Em certos casos, os seres humanos evoluíram de tal maneira que acreditam estar realizando ação X por uma razão pró-social Y, mas, ao executarem a ação X, outras adaptações ocorrem para promover a consequência egoísta Z.

A partir desta premissa, passo agora para uma questão que vai além do escopo da teoria clássica da decisão Bayesiana:

E se estiver operando em um ambiente corrompido?

Nesse cenário, você pode até se pegar fazendo declarações aparentemente paradoxais - algo que parece absurdo do ponto de vista da teoria clássica da decisão - como:

Os fins não justificam os meios.

Contudo, se estiver operando em um ambiente corrompido, a observação reflexiva de que tomar o poder para si parece um ato justo e altruísta pode não ser uma evidência sólida de que tomar o poder é, de fato, a ação que mais beneficiará a tribo.

Pela lente do realismo ingênuo, o ambiente corrompido que você utiliza e as percepções corrompidas que ele gera parecerão a própria estrutura do mundo - simplesmente como as coisas são.

Assim, surge a regra aparentemente paradoxal: “Pelo bem da tribo, não trapaceie para tomar o poder, mesmo que isso resultasse em benefício líquido para a tribo”. De fato, talvez seja mais sensato formulá-la desta maneira. Se dissermos apenas “quando parece que isso traria benefício líquido para a tribo”, algumas pessoas podem argumentar: “Mas não é apenas uma aparência - traria benefício líquido para a tribo se eu estivesse no comando”.

A ideia de um ambiente não confiável parece completamente fora do escopo da teoria clássica da decisão. (O que isso implica para a teoria da decisão reflexiva ainda não posso afirmar com certeza, mas parece o nível adequado para abordar essa questão).

No âmbito humano, no entanto, a solução parece simples. Ao conhecer a distorção, criam-se regras que descrevem o comportamento distorcido e o proíbem. Uma regra que diz: “Pelo bem da tribo, não trapaceie para tomar o poder, mesmo que seja pelo bem da tribo”. Ou ainda: “Pelo bem da tribo, não tire vidas, mesmo pelo bem da tribo”.

E agora o filósofo apresenta seu “experimento mental” - descrevendo um cenário em que, por estipulação, a única maneira de salvar cinco vidas inocentes é assassinar uma pessoa inocente, e esse assassinato certamente salvará as cinco vidas. “Um trem está prestes a atropelar cinco pessoas inocentes, que você não pode avisar para pularem para fora do caminho, mas você pode empurrar uma pessoa inocente para o caminho do trem, o que irá parar o trem. Essas são suas únicas opções; o que você faz?”

Um ser humano altruísta, que aceitou certas proibições deontológicas – que parecem bem justificadas por algumas estatísticas históricas sobre os resultados do raciocínio de certas maneiras em hardware não confiável – pode experimentar algum sofrimento mental ao se deparar com este experimento mental.

Assim, aqui vai uma resposta ao cenário desse filósofo, da qual ainda não ouvi nenhuma refutação por parte do filósofo:

“Você estipula que a única maneira possível de salvar cinco vidas inocentes é assassinar uma pessoa inocente, e esse assassinato certamente salvará as cinco vidas, e que esses fatos são conhecidos por mim com certeza efetiva. Mas, dado que estou operando em hardware corrompido, não posso ocupar o estado epistêmico que você deseja que eu imagine. Portanto, respondo que, numa sociedade de Inteligências Artificiais dignas de personalidade e sem qualquer tendência inerente para ser corrompida pelo poder, seria correto que a IA assassine a única pessoa inocente para salvar cinco, e, além disso, todos os seus pares concordariam. No entanto, recuso-me a estender essa resposta a mim mesmo, porque o estado epistêmico que você me pede para imaginar só pode existir entre outros tipos de seres além dos humanos.”

A meu ver, isso parece uma esquiva. Acredito que o universo é suficientemente cruel para sermos, com justiça, compelidos a considerar situações deste tipo. O tipo de pessoa que propõe esse tipo de experimento mental pode muito bem-merecer esse tipo de resposta. No entanto, qualquer sistema jurídico humano incorpora alguma resposta à pergunta “Quantas pessoas inocentes podemos colocar na prisão para pegar os culpados?”, mesmo que o número não esteja anotado.

Como humano, tento respeitar as proibições deontológicas que os humanos criaram para viver em paz mutuamente. Contudo, não acredito que nossas proibições deontológicas sejam literalmente inerentemente e não consequencialmente corretas. Endosso o princípio “o fim não justifica os meios” como uma orientação para humanos que operam em hardware corrompido, mas não o apoiaria como um princípio para uma sociedade de IAs que fazem estimativas bem calibradas. (Se você tem uma IA em uma sociedade de humanos, isso traz outras considerações, como se os humanos aprendem com o seu exemplo.)

Portanto, não diria que uma IA Amigável bem projetada deva necessariamente se recusar a empurrar aquela pessoa para fora da borda para parar o trem. Obviamente, eu esperaria que qualquer superinteligência decente apresentasse uma terceira alternativa superior. Mas se essas forem as únicas duas alternativas, e a FAI julgar que é mais sensato empurrar a única pessoa para fora do precipício – mesmo depois de levar em conta os efeitos indiretos sobre quaisquer humanos que vejam isso acontecer e espalhem a história, etc. – então não chamo isso de luz de alarme, se uma IA disser que a coisa certa a fazer é sacrificar um para salvar cinco. Mais uma vez, não saio por aí empurrando as pessoas para os trilhos dos trens, nem roubando bancos para financiar meus projetos altruístas. Acontece que sou humano. No entanto, para uma IA Amigável ser corrompida pelo poder seria como começar a sangrar sangue vermelho. A tendência a ser corrompido pelo poder é uma adaptação biológica específica, apoiada por circuitos cognitivos específicos, construídos em nós pelos nossos genes por uma razão evolutiva clara. Ele não apareceria espontaneamente no código de uma IA Amigável, assim como seus transistores não começariam a sangrar.

Iria ainda mais longe e diria que, se você tivesse mentes com uma distorção inerente que as fizesse superestimar o dano externo das ações em benefício próprio, então elas precisariam de uma regra “os fins não proíbem os meios” – que você deveria fazer o que beneficia a si mesmo quando (parece) prejudicar a tribo. Por hipótese, se a sua sociedade não tivesse esta regra, as mentes nela contidas recusar-se-iam a respirar por medo de usar o oxigênio de outra pessoa, e todos morreriam. Para eles, um excesso ocasional em que

uma pessoa aproveita um benefício pessoal às custas líquidas da sociedade pareceria tão cautelosamente virtuoso – e na verdade tão cautelosamente virtuoso – como quando um de nós, humanos, sendo cauteloso, deixa passar uma oportunidade de roubar um pedaço de pão que realmente teria sido mais benéfico para eles do que uma perda para o comerciante (incluindo efeitos indiretos).

"O fim não justifica os meios" é apenas um raciocínio consequencialista num meta nível acima. Se um ser humano começar a pensar no nível do objeto que o fim justifica os meios, isso terá consequências terríveis, dadas as nossas mentes indignas de confiança; portanto, um humano não deveria pensar assim. Mas tudo continua, em última análise, no campo do consequencialismo. É apenas consequencialismo reflexivo, para seres que sabem que as suas decisões momento a momento são tomadas por hardware não confiável.

288 - Injunções éticas

> Você sacrificaria vidas inocentes, se esse fosse o caminho certo a seguir? Se não, em que circunstâncias você hesitaria em fazer o que é moralmente correto? Se sim, qual seria o seu critério para determinar quantas vidas seria justificável sacrificar?
>
> — pergunta perturbadora em uma entrevista de emprego

Mudando de perspectiva por um momento, estou profissionalmente intrigado com a teoria da decisão sobre "ações que não devemos realizar, mesmo que pareçam ser a escolha correta".

Consideremos uma inteligência artificial reflexiva, autotransformadora e autoaperfeiçoante, em um estágio intermediário de desenvolvimento. Especificamente, o sistema de objetivos da IA não está finalizado – a forma de suas motivações continua sendo carregada, aprendida, testada ou ajustada.

Sim, tenho observado diversas maneiras de comprometer o design de um sistema de objetivos de IA, resultando em um sistema de decisão que, com base em seus objetivos, decide que o universo deve ser preenchido com rostinhos moleculares sorridentes, ou algo semelhante. Geralmente, essas sugestões mortais também têm a propriedade de que a IA não desejará que seus programadores as corrijam. Se a IA estiver suficientemente avançada – o que pode ocorrer mesmo em um estágio intermediário – então a IA também pode perceber que enganar os programadores, escondendo as mudanças em seus pensamentos, ajudará a transformar o universo em rostos sorridentes.

Agora, do ponto de vista dos programadores, se assumirmos que a IA decidiu esconder seus pensamentos dos programadores, ou agir deliberadamente para nos enganar de outra forma, pareceria provável que alguma consequência não intencional tenha ocorrido no sistema de metas. Consideraríamos provável que a IA não esteja operando conforme o esperado, mas sim que de alguma forma tenhamos bagunçado a função de utilidade da IA. Assim, a IA quer transformar o universo em pequenos contadores de sistema de recompensa, ou algo do tipo, e agora tem um motivo para se esconder de nós.

Bem, suponha que não implementaremos alguma Grande Ideia ao nível de objeto como função de utilidade da IA. Em vez disso, faremos algo avançado e recursivo – construir um sistema de metas que conheça (e se preocupe) com os programadores externos. Um sistema de metas que, com alguma estrutura interna não trivial, "sabe que está sendo programado" e "sabe que está incompleto". Então, você pode ter e manter a regra:

Se [eu decidir que] enganar meus programadores é a escolha certa a fazer, execute um desligamento controlado [em vez de fazer o que é moralmente correto].

E a IA manteria essa regra, mesmo que a IA autotransformadora fizesse revisões em seu próprio código, porque, em seu sistema de objetivos estruturalmente não trivial, a IA do presente entende que essa decisão de uma IA do futuro provavelmente indica um mau funcionamento definido. Além disso, a IA do presente sabe que, se a IA do futuro tentar avaliar a utilidade de realizar um desligamento após esse hipotético mau funcionamento, a IA do futuro provavelmente decidirá não se desligar. Portanto, o desligamento deveria ocorrer incondicionalmente, automaticamente, sem que o sistema de metas tivesse outra chance de recalcular o que é moralmente correto.

Não aprofundarei nas intricadas nuances da estrutura matemática exata, pois isso estaria além do escopo deste livro. Além disso, ainda não explorei completamente as complexidades da estrutura matemática. Parece que deveria ser possível, se você utilizar abordagens avançadas e recursivas com uma estrutura não trivial (mas consistente). Mas até agora, isso permanece apenas como um sonho.

Entretanto, o foco aqui não é a IA avançada; trata-se da ética humana. Apresento o cenário da IA para destacar de forma mais clara a estranha ideia de uma injunção ética:

Jamais se deve assassinar uma pessoa inocente que o ajudou, mesmo que seja a coisa certa a fazer; pois é muito mais provável que você tenha cometido um erro do que assassinar uma pessoa inocente que o ajudou seja a coisa certa a fazer.

Parece razoável?

Durante a Segunda Guerra Mundial, tornou-se necessário destruir o fornecimento de deutério da Alemanha, um moderador de nêutrons, para bloquear suas tentativas de alcançar uma reação em cadeia de fissão. O fornecimento de deutério vinha, na época, de uma instalação capturada na Noruega. Um carregamento de água pesada estava a bordo de uma balsa norueguesa, o SF Hydro. Knut Haukelid e três outros embarcaram na balsa para sabotá-la, quando os sabotadores foram descobertos pelo vigia da balsa. Haukelid disse-lhe que estavam fugindo da Gestapo, e o vigia concordou imediatamente em ignorar a presença deles. Haukelid considerou avisar seu benfeitor, mas decidiu que isso poderia pôr a missão em perigo e limitou-se a agradecer-lhe e apertar-lhe a mão [1]. Assim, a balsa civil Hydro afundou-se na parte mais profunda do lago, com dezoito mortos e vinte e nove sobreviventes. Algumas das equipes de resgate norueguesas sentiram que os soldados alemães presentes deveriam ser deixados para se afogar, mas esta atitude não prevaleceu e quatro alemães foram resgatados. E esse foi, efetivamente, o fim do programa de armas atômicas nazista.

Boa jogada? Má decisão? A Alemanha muito provavelmente não teria conseguido a bomba de qualquer maneira... Espero, com absoluto desespero, nunca ser confrontado com uma escolha como essa, mas no final, não posso dizer uma palavra contra isso.

Por outro lado, quando se trata da regra:

Nunca tente enganar a si ou ofereça uma razão para acreditar que não seja uma verdade provável; pois mesmo que você encontre um motivo surpreendentemente inteligente, é mais provável que você tenha cometido um erro do que tenha uma expectativa razoável de que isso seja um benefício líquido no longo prazo.

Então, realmente não conheço ninguém que tenha enfrentado conscientemente uma exceção. Há momentos em que você tenta se convencer de que “não estou escondendo nenhum judeu no meu porão” antes de falar com o oficial da Gestapo. Mas então você ainda sabe a verdade, está apenas tentando criar algo como um eu alternativo que existe em sua imaginação, uma fachada para conversar com o oficial da Gestapo.

Mas realmente acreditar em algo que não é verdade? Não sei se houve alguém para quem isso fosse uma boa ideia. Tenho certeza de que houve muitas épocas na história da humanidade na qual a pessoa X estava em melhor situação com a falsa crença Y. E, da mesma forma, há sempre algum conjunto de números vencedores da loteria em cada sorteio. Saber qual bilhete de loteria ganhará é a parte epistemicamente difícil, assim como X saber quando está melhor com uma crença falsa.

O auto engano é o pior tipo de aposta do cisne negro, muito pior do que as mentiras, porque sem conhecer o verdadeiro estado das coisas, você não consegue nem adivinhar qual será a penalidade pelo seu auto engano. Eles só precisam explodir uma vez para desfazer todo o bem que fizeram. Uma única vez em que você ora a Deus depois de descobrir um caroço, em vez de ir ao médico. Isso é tudo o que é preciso para desfazer uma vida. Toda a felicidade que o pensamento caloroso de uma vida após a morte alguma vez produziu na humanidade foi agora mais do que cancelada pelo fracasso da humanidade em instituir preservações criônicas sistemáticas depois de o nitrogênio líquido se ter tornado barato de fabricar. E não creio que alguém alguma vez tenha tido em mente esse tipo de fracasso como uma possível explosão, quando disse: ‘Mas precisamos de crenças religiosas para amortecer o medo da morte’. É disso que tratam as apostas do cisne negro: a explosão inesperada.

É possível que você faça uma ou duas apostas no cisne negro – elas nem sempre resultam em consequências negativas. Então, você repete o processo, e é aí que surge a explosão, cancelando todos os benefícios e alguns extras. Essa é a essência das apostas no cisne negro.

A dificuldade está em saber quando é seguro acreditar em uma mentira (assumindo que você consiga lidar com toda a contorção mental, para começar) – parte da natureza das apostas no cisne negro é que você não vê a ameaça que irá prejudicá-lo. Como nossas percepções moldam a realidade, parece não haver ameaça, ponto final.

Diante disso, posso afirmar que existe uma injunção ética contra o auto engano. Chamo isso de "injunção ética" não tanto porque seja uma questão de moralidade interpessoal (embora seja), mas porque é uma regra que protege você de sua própria inteligência – uma barreira contra a tentação de fazer o que parece ser a coisa certa.

Assim, temos dois tipos de situações que podem sustentar uma "injunção ética", uma regra para não fazer nada mesmo quando é a coisa certa a fazer. (Ou seja, você se abstém "mesmo quando seu cérebro calculou que é a coisa certa a fazer", mas isso só parece "a coisa certa a fazer".)

Primeiramente, por sermos humanos e operarmos em hardware corrompido, podemos generalizar classes de situações em que, se alguém diz, por exemplo, "É hora de roubar alguns bancos para um bem maior", é mais provável que tenha sido corrompido do que realmente seja o caso. (É importante observar que não estamos proibindo tal ato na realidade, mas questionamos o estado epistêmico em que você está justificado em confiar em seu próprio cálculo de que isso é a coisa certa a fazer – bilhetes de loteria justos podem ganhar, mas não se pode comprá-los de maneira justificada.)

Em segundo lugar, a história pode nos ensinar que certas ações se enquadram como apostas no cisne negro, ou seja, eventualmente explodem em grande escala por razões que não estão no modelo do tomador de decisões. Mesmo quando calculamos, no modelo, que algo parece ser a coisa certa a fazer, aplicamos o conhecimento adicional do problema do cisne negro para chegar a uma injunção contra tal ação.

Entretanto, se alguém estiver ciente dessas razões, pode-se simplesmente recalcular, levando-as em consideração. Portanto, podemos roubar bancos se isso parecer a coisa certa a fazer, após considerar o problema do hardware corrompido e das explosões do cisne negro. Esse é o caminho racional, não é mesmo?

Há várias maneiras de abordar essa questão.

Começarei destacando que este é um excelente exemplo do tipo de raciocínio que tenho em mente quando alerto os aspirantes a racionalistas para serem cautelosos em relação à inteligência.

Também observarei que não gostaria que uma IA Amigável, ao decidir repentinamente que a Terra deveria ser transformada em clipes de papel, avaliasse se isso seria uma ação razoável à luz de todos os alertas que recebeu contra isso. Preferiria que ela passasse por um desligamento controlado automático. Quem disse que o meta-raciocínio está imune à corrupção?

Poderia mencionar momentos cruciais nos quais minhas inibições éticas ingênuas e idealistas me protegeram de mim mesmo e me colocaram em uma posição recuperável ou ajudaram a iniciar a recuperação de erros muito profundos que eu não tinha ideia de que estava cometendo. Além disso, questionaria se avancei tanto assim e se seria sensato remover as proteções que já me salvaram.

Entretanto, a pergunta "Ainda sou mais burro do que minha ética?" é complexa e a resposta não é automaticamente "Sim".

Existem ações óbvias e tolas que devem ser evitadas; por exemplo, não se deve esperar até estar realmente tentado para então tentar descobrir se é mais inteligente do que sua ética naquela ocasião específica.

No entanto, de forma geral, há um limite para o poder que pode ser investido naquilo que nossos pais nos disseram para não fazer. Esse poder não deve ser subestimado. Pessoas inteligentes debateram lições históricas no processo de desenvolvimento da ética iluminista, na qual grande parte da cultura ocidental se baseia; algumas subculturas, como a academia científica ou os fãs de ficção científica, baseiam-se mais diretamente nessa ética. Mesmo assim, o poder do passado é limitado.

E, na verdade...

Fui obrigado a tornar minha ética mais rígida do que meus pais, Jerry Pournelle e Richard Feynman, me aconselharam a fazer.

O curioso é que, quando as pessoas parecem pensar que são mais inteligentes do que a ética que possuem, elas defendem menos rigor em vez de mais. Quero dizer, ao considerar como o mundo moderno é infinitamente mais complexo...

E na mesma linha, aqueles que se aproximam de mim e dizem: "Você deveria mentir sobre o aumento da inteligência, porque assim conseguiria mais apoio; é a coisa racional a fazer, pelo bem maior" – parecem não ter noção dos riscos envolvidos.

Eles não mencionam o problema da execução em hardware corrompido. Eles ignoram a ideia de que as mentiras devem ser protegidas recursivamente contra todas as verdades e técnicas de descoberta da verdade que as ameaçam. Não levam em consideração que os métodos honestos têm uma simplicidade muitas vezes ausente nos métodos desonestos. Nada é dito sobre apostas do cisne negro. Nada é mencionado sobre a terrível vulnerabilidade de descartar a última defesa contra si e tentar sobreviver com base em cálculos brutos.

Estou razoavelmente certo de que isso ocorre porque eles não têm compreensão alguma desses aspectos.

Se você realmente compreende a razão e o ritmo por trás da ética, um sinal importante é que, amplificado por esse conhecimento recém-adquirido, você evita aquelas ações que antes pareciam transgressões éticas. Agora você sabe por quê.

Aquele que olha apenas para uma ou duas razões por trás da ética e pensa: "Ok, entendi isso, agora vou considerar conscientemente e não preciso mais de inibições éticas" – está agindo mais como um estereótipo do que como um verdadeiro racionalista. O mundo não é simples, puro e limpo; portanto, não se pode simplesmente adotar a ética com que foi criado e confiar nela. No entanto, a pretensão da lógica vulcana, em que se acredita que se calculará tudo corretamente após obter um ou dois insights abstratos – isso também não funciona na vida real.

Quanto àqueles que, não tendo descoberto nada disso, se consideram mais espertos do que sua ética: Ha!

E quanto àqueles que antes se consideravam mais inteligentes do que a ética, mas que não haviam concebido todos esses elementos por trás das injunções éticas "em tantas palavras" até encontrarem este ensaio, e que agora se consideram mais inteligentes do que sua ética, porque levarão tudo isso em conta daqui para frente: Ha! Duplo!

Tenho visto muitas pessoas lutando para se desculpar de sua ética. A modificação é sempre no sentido da clemência, nunca para ser mais rigorosa. E fico impressionado com a rapidez e a leveza com que se esforçam para abandonar suas proteções. Hobbes disse: "Não sei o que é pior: o fato de todos terem um preço ou o fato de seu preço ser tão baixo." O preço é tão baixo que estão ansiosos para serem comprados. Não buscam alternativas duas vezes, muito menos uma terceira vez, antes de decidirem que não têm outra opção senão transgredir – embora possam parecer muito graves e solenes ao dizer isso. Abandonam sua ética na primeira oportunidade. "Onde há vontade de falhar, obstáculos podem ser encontrados." A vontade de falhar na ética parece muito forte em algumas pessoas.

Não sei se posso endossar injunções éticas absolutas que vinculem todos os estados epistêmicos possíveis de um cérebro humano. O universo não é gentil o suficiente para confiar nisso. (Embora uma injunção ética contra o auto engano, por exemplo, me pareça ter uma força tremenda. Já vi muitas pessoas defendendo o Lado Negro, e nenhuma delas parece ciente dos riscos da rede ou dos riscos do cisne negro do auto engano.) Se, algum dia, eu tentar formular uma injunção (reflexivamente consistente) numa IA automodificadora, será somente depois de fazer as contas, pois esse não é totalmente o tipo de coisa que se pode fazer de forma improvisada.

Mas direi isso:

Não estou completamente impressionado com o conhecimento, o raciocínio e o nível geral daqueles que vieram até mim ansiosamente e disseram em tom grave: “É racional fazer a coisa antiética X porque isso trará o benefício Y”.

Referências

[1] Richard Rhodes, The Making of the Atomic Bomb (New York: Simon & Schuster, 1986).

289 - Algo para proteger

Na gestalt da ficção japonesa, é comum encontrar um tema recorrente: o poder vem de ter algo para proteger.

Não se trata apenas dos super-heróis que se fortalecem quando um amigo está ameaçado, como acontece na ficção ocidental. Na versão japonesa, é algo mais profundo.

Na saga X, fica explícito que cada um dos heróis tira sua força de alguém – uma pessoa – que desejam proteger. Quem é essa pessoa? Essa pergunta faz parte da trama de X – a "pessoa mais preciosa" nem sempre é quem pensamos. Mas se essa pessoa é morta ou ferida de forma errada, o protetor perde seu poder. Isso não é nada que acontece uma vez por semana, como em uma história em quadrinhos Ocidentais. É como "morrer de verdade" ou sejam, ser retirado do jogo.

Nos quadrinhos de super-heróis ocidentais, o herói é picado por uma aranha radioativa e precisa encontrar algo para fazer com seus poderes para mantê-lo ocupado, então decide lutar contra o crime. Os super-heróis ocidentais estão sempre reclamando sobre o tempo que seus deveres consomem e como preferem ser mortais comuns para pescar ou fazer algo assim.

Na vida real ocidental, as pessoas infelizes são informadas de que precisam de um "propósito na vida" e devem escolher uma causa altruísta que combine com sua personalidade, como escolher cortinas bonitas para a sala de estar. Isso trará um pouco de cor e iluminará seus dias, mas é preciso ter cuidado para não escolher nada muito caro.

Nos quadrinhos ocidentais, a mágica geralmente vem antes do propósito: primeiro, adquire-se poderes incríveis e depois decide-se proteger os inocentes. Na ficção japonesa, muitas vezes, acontece o oposto.

É claro que não estou dizendo tudo isso para generalizar a partir de evidências fictícias. Mas quero transmitir um conceito cujo análogo ocidental enganosamente próximo não é o que quero dizer.

Já abordei antes a ideia de que um racionalista deve ter algo que valorize mais do que a "racionalidade": a Arte deve ter um propósito diferente de si mesma, ou desmorona em recursividade infinita. Mas não me entenda mal e pense que estou defendendo que os racionalistas devem escolher uma boa causa altruísta só para ter algo para fazer, porque a racionalidade não é tão importante por si só. Não. O que estou perguntando é: De onde vêm os racionalistas? Como adquirimos nossos poderes?

Está escrito em As Doze Virtudes da Racionalidade:

> Como você pode melhorar sua concepção de racionalidade? Não dizendo para si: "É meu dever ser racional". Com isso você apenas consagra sua concepção equivocada. Talvez a sua concepção de racionalidade seja a de que é racional acreditar nas palavras do Grande Mestre, e o Grande Mestre diz: "O céu é verde", e você olha para o céu e vê azul. Se você pensa: "Pode parecer que o céu é azul, mas a racionalidade é acreditar nas palavras do Grande Mestre", você perde a chance de descobrir o seu erro.

Falando em termos históricos, a humanidade finalmente conseguiu escapar da armadilha da autoridade e começou a prestar atenção, sabe, no céu real, porque as crenças baseadas em experimentos se mostraram muito mais úteis do que as baseadas em autoridade.

A curiosidade existe desde o início da humanidade, mas o problema é que contar histórias ao redor da fogueira funciona tão bem para satisfazer a curiosidade.

Historicamente, a ciência prevaleceu porque exibiu mais força bruta na forma de tecnologia, e não porque a ciência parecia mais razoável. Até hoje, a magia e as escrituras ainda parecem mais razoáveis para quem não está familiarizado com a ciência. É por isso que há uma tensão social contínua entre os diferentes sistemas de crenças. Se a ciência não apenas funcionasse melhor do que a magia, mas também parecesse mais intuitivamente razoável, já teria vencido completamente nesse ponto.

Alguns questionam: "Como você ousa sugerir que algo deva ser mais valorizado do que a Verdade? Um racionalista não deve amar a Verdade mais do que a mera utilidade?"

Esqueça por um momento o que teria acontecido historicamente com alguém assim - que pessoas com esse estado de espírito defendiam a Bíblia porque amavam a verdade mais do que a mera precisão. A moralidade proposicional é uma coisa gloriosa, mas tem muitos graus de liberdade.

Não. O ponto principal é que o amor do racionalista pela Verdade é apenas mais complicado como um relacionamento emocional.

Ninguém se torna um racionalista sem se importar com a verdade, tanto como um objetivo moral quanto como algo divertido de se ter. Duvido que existam muitos mestres compositores que odeiam música.

O que me agrada na racionalidade é a disciplina imposta pela exigência de crenças para produzir previsões, o que nos leva muito mais perto da verdade do que se ficarmos obcecados com a Verdade o dia todo. Gosto da complexidade de ter que amar simultaneamente as ideias que parecem verdadeiras e estar pronto para descartá-las a qualquer momento. Gosto até da pureza estética de declarar que valorizo a utilidade acima da estética. Isso é quase uma contradição, mas não totalmente; e isso tem uma qualidade estética também, um humor delicioso.

Mas, é claro, não importa o quanto você professe seu amor pela mera utilidade, você realmente nunca deve deliberadamente acreditar em uma declaração falsa útil.

Portanto, não simplifique demais a relação entre a busca pela verdade e pela utilidade. Não é um ou outro. É complicado, o que não é necessariamente um defeito na estética moral de eventos isolados.

Mas apenas a crença de que se deve ser "racional" ou que certas formas de pensar são "bonitas" não o levará ao caminho certo. Isso não teria tirado a humanidade do buraco da autoridade.

Em "Sentindo-se moral", discuti este dilema: qual destas opções você prefere?

1. Salvar 400 vidas, com certeza.
2. Salvar 500 vidas, com 90% de probabilidade; não salvar nenhuma vida, com 10% de probabilidade.

Você pode se sentir tentado a se gabar, dizendo: "Como você ousa jogar com a vida das pessoas?" Mesmo que você seja um dos 500 - mas não saiba qual deles - você ainda pode ser tentado a confiar no sentimento reconfortante da certeza, porque muitas vezes nossas próprias vidas valem menos para nós do que uma boa intuição.

Mas se sua filha preciosa é uma das 500, e você não sabe qual delas, então, talvez, você possa se sentir mais impelido a multiplicar as chances - perceber que tem 80% de chance de salvá-la no primeiro caso e 90% de chance no segundo.

Sim, todos naquela multidão são filhos ou filhas de alguém. Isso sugere que devemos escolher a segunda opção como altruístas e pais preocupados.

Meu ponto não é sugerir que a vida de uma pessoa é mais valiosa do que a de 499 pessoas. Estou tentando dizer que, antes que uma pessoa fique desesperada o suficiente para recorrer à matemática, mais do que a própria vida deve estar em jogo.

Se você acredita que escolher a certeza da opção 1 é "racional", muitas pessoas pensam que a "racionalidade" é escolher apenas métodos que certamente funcionarão e rejeitar toda incerteza. Espero que você se preocupe mais com a vida de sua filha do que com a "racionalidade".

O orgulho de sua própria virtude como racionalista o salvará? Você só poderá aprender algo sobre racionalidade se a vida de sua filha for mais importante para você do que seu orgulho de racionalista.

Você pode até aprender algo sobre racionalidade com a experiência, se já estiver suficientemente desenvolvido em sua Arte para dizer: "Devo ter uma concepção errada de racionalidade" e não: "Veja como a racionalidade me deu a resposta errada!"

(A dificuldade essencial em se tornar um mestre racionalista é que você precisa de um pouco de racionalidade para iniciar o processo de aprendizagem.)

A sua crença de que deve ser racional é mais importante do que a sua vida? Afinal, como eu já observei, arriscar a sua vida não é tão assustador comparativamente. É muito mais assustador ser a única voz discordante na multidão e ter todos olhando para você de forma estranha, conforme as preferências reveladas dos adolescentes que bebem em festas e depois dirigem para casa. Será preciso algo terrivelmente importante para fazê-lo deixar o grupo. Uma ameaça à sua vida não será suficiente.

A sua vontade de ser racional é mais forte do que o seu orgulho? Talvez, se a sua vontade de ser racional derivar do seu orgulho em sua autoimagem como um racionalista. É útil ter uma autoimagem que diga que você é o tipo de pessoa que enfrenta a dura verdade. No entanto, pode chegar um momento em que você terá que admitir que está errado sobre a racionalidade. Então, o seu orgulho, a sua autoimagem de racionalista, pode tornar isso muito difícil de enfrentar.

Se você sente orgulho em acreditar no que o Grande Mestre diz - mesmo quando é difícil de aceitar, mesmo quando você prefere não acreditar -, pode ser ainda mais difícil admitir que o Grande Mestre é uma fraude. Isso pode fazer com que seja uma verdade ainda mais difícil de encarar, especialmente porque todo o seu nobre autossacrifício pode parecer ter sido em vão.

De onde você tira a vontade de seguir em frente?

Olhando para trás, em minha jornada pessoal em direção à racionalidade - não apenas na jornada histórica da humanidade - percebo que cresci acreditando fortemente que deveria ser racional. Isso me tornou um Racionalista Tradicional acima da média, nos moldes de Feynman e Heinlein, e nada mais. Não me levou a ir além dos ensinamentos que recebi. Apenas comecei a crescer ainda mais como racionalista quando tive algo extremamente importante para fazer. Algo mais importante do que o meu orgulho de racionalista, não importando o que acontecesse com a minha vida.

Apenas quando você está mais comprometido com o sucesso do que com qualquer uma das suas amadas técnicas de racionalidade, é que começa a apreciar essas palavras de Miyamoto Musashi [1] :

> Você pode vencer com uma arma longa, mas também pode vencer com uma arma curta. Em suma, a escola do Caminho do Ichi é o espírito de vencer, seja qual for a arma e seja qual for o seu tamanho[22].
>
> —Miyamoto Musashi, O livro dos cinco anéis

Não confunda isso com um ensinamento específico de racionalidade. Ele descreve como você aprende o Caminho, começando com uma necessidade desesperada de sucesso. Ninguém domina o Caminho até que mais do que a sua vida esteja em jogo. Mais do que o seu conforto, mais ainda do que o seu orgulho.

Você não pode simplesmente escolher uma causa assim porque sente que precisa de um passatempo. Procure uma "boa causa" e a sua mente apenas preencherá um clichê padrão. Aprenda a multiplicar e talvez reconheça uma causa drasticamente importante quando a vir.

Mas se você tem uma causa como essa, é certo e adequado usar sua racionalidade a serviço dela.

Subordinar estritamente a estética da racionalidade a uma causa superior faz parte da estética da racionalidade. Preste atenção a essa estética: você nunca dominará a racionalidade o suficiente para vencer com qualquer arma se não apreciar a beleza por si mesma.

---

22 NT. Texto original em inglês. *You can win with a long weapon, and yet you can also win with a short weapon. In short, the Way of the Ichi school is the spirit of winning, whatever the weapon and whatever its size.*

Referências

[1] Musashi, Book of Five Rings.

290 - Quando (não) usar probabilidades

Pode surpreender alguns leitores saber que nem sempre defendo o uso de probabilidades.

Melhor dizendo, não defendo que os seres humanos, ao tentarem resolver seus problemas, criem probabilidades verbais e, em seguida, apliquem as leis da teoria da probabilidade ou da teoria da decisão a qualquer número que tenham inventado, usando o resultado como sua crença ou decisão final.

As leis da probabilidade são leis, não sugestões, mas muitas vezes a verdadeira Lei é demasiado difícil para nós, humanos, calcularmos. Se P ≠ NP e o universo não possuir uma fonte de poder computacional exponencial, então há atualizações evidenciais muito difíceis de serem computadas até mesmo por uma superinteligência – mesmo que as probabilidades fossem bastante bem definidas, se pudéssemos nos dar ao luxo de calculá-las.

Portanto, às vezes, não aplicamos a teoria da probabilidade. Especialmente se você for humano e seu cérebro tiver evoluído com todos os tipos de algoritmos úteis para o raciocínio incerto, que não envolvem atribuições verbais de probabilidade.

Não tem certeza de onde uma bola voadora pousará? Não aconselho a tentar formular uma distribuição de probabilidade sobre seus pontos de aterrissagem, realizando atualizações bayesianas deliberadas em seus olhares para a bola e calculando a utilidade esperada de todas as sequências possíveis de instruções motoras para seus músculos. Ao tentar pegar uma bola voadora, você provavelmente estará em melhor situação com os mecanismos internos do seu cérebro do que usando o raciocínio verbal deliberativo para inventar ou manipular probabilidades.

Mas isso não significa que você está indo além da teoria das probabilidades ou acima da teoria das probabilidades.

Os argumentos do livro holandês ainda se aplicam. Se eu lhe oferecer uma escolha de apostas (US$ 10.000 se a bola parar neste quadrado, contra US$ 10.000 se eu lançar um dado e sair 6), e você responder de uma forma que não permite a atribuição de probabilidades consistentes, então você aceitará combinações de apostas que representem perdas certas ou rejeitará apostas que representem ganhos certos...

O que ainda não significa que você deva tentar usar o raciocínio verbal deliberativo. Eu esperaria que, pelo menos para jogadores profissionais de beisebol, fosse mais importante pegar a bola do que atribuir probabilidades consistentes. Na verdade, se você tentasse inventar probabilidades, as probabilidades verbais poderiam até não ser ótimas, em comparação com algum sentimento instintivo – alguma representação muda de incerteza no fundo da sua mente.

Não há nada de privilegiado na incerteza expressa em palavras, a menos que as partes verbais do seu cérebro, de fato, funcionem melhor no problema.

E embora mapas precisos do mesmo território sejam necessariamente consistentes entre si, nem todos os mapas consistentes são precisos. É mais importante ser preciso do que consistente, e mais importante pegar a bola do que ser consistente.

Na verdade, geralmente desaconselho inventar probabilidades, a menos que pareça que você tenha uma base decente para elas. Isso apenas faz você acreditar que é mais bayesiano do que realmente é.

Para ser mais específico, eu desaconselharia, geralmente, a utilização de procedimentos não numéricos para criar o que parecem probabilidades numéricas. Os números devem vir de números.

Agora há benefícios em tentar traduzir seus sentimentos viscerais de incerteza em probabilidades verbais. Pode ajudá-lo a identificar problemas como a falácia da conjunção. Pode ajudá-lo a detectar inconsistências internas – embora possa não mostrar nenhuma maneira de remediá-las.

Mas você não deve sair por aí pensando que, se traduzir seu pressentimento em "uma em mil", então, nas ocasiões em que emitir essas palavras verbais, o evento correspondente acontecerá cerca de uma em mil vezes. Seu cérebro não está tão bem calibrado. Se, em vez disso, você fizer algo não-verbal com seu sentimento visceral de incerteza, talvez você se saia melhor, porque pelo menos estará usando o sentimento visceral da maneira que ele deveria ser usado.

Este tópico específico surgiu recentemente no contexto do Grande Colisor de Hádrons (LHC)[23] e de um argumento apresentado na conferência sobre Riscos Catastróficos Globais: que não podíamos ter certeza de que não havia erros nos artigos que mostravam, de vários ângulos, que o LHC poderia possivelmente não destruirá o mundo. Além disso, a teoria usada nos artigos pode estar errada. E em qualquer dos casos, ainda havia uma hipótese de o LHC destruir o mundo. E, portanto, não deve ser ativado.

Veja bem, se o argumento tivesse sido apresentado exatamente dessa forma, eu não teria objeções à sua epistemologia.

No entanto, o palestrante na realidade queria atribuir uma probabilidade de pelo menos 1 em 1.000 de que a teoria, o modelo ou os cálculos no artigo do LHC estivessem incorretos; e uma probabilidade de pelo menos 1 em 1.000 de que, se a teoria, o modelo ou os cálculos estivessem incorretos, o LHC poderia destruir o mundo.

Afinal, não é tão improvável assim que as futuras gerações rejeitem a teoria utilizada no artigo do LHC, ou rejeitem o modelo, ou talvez até mesmo encontrem um erro. E se o artigo do LHC estiver errado, quem sabe o que poderá acontecer como resultado?

Então, isso é um argumento – mas atribuir números a ele?

Discordo do ar de autoridade conferido a esses números retirados do nada. Geralmente, sinto que se não podemos utilizar ferramentas probabilísticas para moldar nossos sentimentos de incerteza, não devemos dignificá-los chamando-os de probabilidades.

A alternativa que eu sugeriria, neste caso específico, é debater a regra geral de proibir experimentos de física porque não podemos ter certeza dos argumentos que afirmam ser seguros.

Sustento que, formulando dessa maneira, a mente, ao considerar as frequências dos eventos, provavelmente trará mais consequências à decisão e se lembrará de casos históricos mais relevantes.

Se você debater apenas um caso do LHC e atribuir probabilidades específicas, isso (1) confere a um raciocínio muito instável uma aura indevida de autoridade, (2) obscurece as consequências gerais da aplicação de regras semelhantes e até mesmo (3) cria a ilusão de que poderíamos tomar uma decisão diferente se alguém publicasse um novo artigo de física que diminuísse as probabilidades.

Os autores da conferência sobre Risco Global Catastrófico pareciam sugerir que poderíamos simplesmente fazer um pouco mais de análise do LHC e, em seguida, ligá-lo. Isso me pareceu a parte mais hipócrita do argumento. Uma vez admitido o argumento "Talvez a análise possa estar errada, e quem sabe o que acontece então", não há nenhum artigo de física possível que possa se livrar dele.

Não importa quais outros artigos de física tenham sido publicados anteriormente, os autores teriam usado o mesmo argumento e elaborado as mesmas probabilidades numéricas na conferência sobre Risco Catastrófico Global. Não posso ter certeza desta afirmação, é claro, mas ela tem uma probabilidade de 75%.

Em geral, um racionalista tenta fazer com que suas mentes funcionem com a melhor potência possí-

23 NT. Em inglês, a sigla LHC significa *Large Hadron Collider*.

vel; às vezes isso envolve falar sobre probabilidades verbais, e às vezes não, mas sempre as leis da teoria das probabilidades governam.

Se tudo o que você tem é uma sensação instintiva de incerteza, então você provavelmente deveria se ater aos algoritmos que recorrem a sentimentos viscerais de incerteza, porque seus algoritmos integrados podem ter um desempenho melhor do que suas tentativas desajeitadas de colocar as coisas em palavras.

Agora pode ser que, ao raciocinar dessa maneira, eu me ache inconsistente. Por exemplo, eu ficaria substancialmente mais alarmado com um dispositivo de loteria com uma probabilidade bem definida de 1 em 1.000.000 de destruir o mundo do que com o fato de o Grande Colisor de Hádrons ser ligado.

Por outro lado, se me perguntasse se eu poderia fazer um milhão de declarações de autoridade equivalentes a 'O Grande Colisor de Hádrons não destruirá o mundo' e estar errado, em média, cerca de uma vez, então eu teria que dizer não.

O que devo fazer sobre essa inconsistência? Não tenho certeza, mas certamente não usarei uma varinha mágica para fazer isso desaparecer. É como encontrar uma inconsistência em um par de mapas que você possui e rapidamente fazer algumas alterações para garantir que sejam consistentes.

A propósito, ficaria substancialmente mais preocupado com um dispositivo de loteria com uma chance em 1.000.000.000 de destruir o mundo do que com um dispositivo que destruiria o mundo se o Deus judaico-cristão existisse. No entanto, não assumiria que poderia fazer mil milhões de afirmações, uma após a outra, totalmente independentes e igualmente carregadas como 'Deus não existe', e estar errado, em média, cerca de uma vez.

Não posso dizer que estou feliz com este estado de assuntos epistêmicos, mas não vou modificá-lo até que possa me ver caminhando na direção de maior precisão e eficácia no mundo real, e não apenas me movendo na direção de maior auto consistência. Afinal, o objetivo é vencer. Se eu inventar uma probabilidade que não é moldada por ferramentas probabilísticas, se eu inventar um número que não é criado por métodos numéricos, então talvez esteja apenas derrotando meus algoritmos internos que fariam melhor se raciocinassem em seus modos nativos de cálculo de incerteza.

É claro que isso não é uma licença para ignorar probabilidades bem fundamentadas. Qualquer cálculo numérico provavelmente será melhor do que um vago sentimento de incerteza; os humanos são estatísticos terríveis. Mas extrair um número inteiramente da sua imaginação, isto é, usar um procedimento não numérico para produzir um número, quase não tem fundamento algum; e, nesse caso, provavelmente será melhor você ficar com os vagos sentimentos de incerteza.

É por isso que minha escrita geralmente usa palavras como «talvez» e «provavelmente» e ‹certamente› em vez de atribuir probabilidades numéricas inventadas como «40%» e «70%» e «95%». Pense em como isso pareceria bobo. Acho que seria realmente bobo; acredito que faria pior com isso.

Não sou o tipo de bayesiano que diz que se deve inventar probabilidades para evitar ficar sujeito aos livros holandeses. Sou o tipo de bayesiano que diz que, na prática, os humanos acabam sujeitos aos livros holandeses porque não são poderosos o suficiente para evitá-los; e além disso, é mais importante pegar a bola do que evitar os livros holandeses. A matemática é como a física subjacente, governando inevitavelmente, mas muito cara para calcular.

Também não faz sentido um ritual de cognição que imite as formas superficiais da matemática, mas que não produza uma tomada de decisão sistematicamente melhor. Isso seria um propósito perdido; esta não é a verdadeira arte de viver sob a lei.

291 - O problema de Newcomb e o arrependimento da racionalidade

O dilema que se segue pode muito bem ser o mais controverso na história da teoria da decisão:

Uma superinteligência de outra galáxia, a qual chamaremos de Ômega, chega à Terra e inicia um jogo peculiar. Neste jogo, Ômega escolhe um ser humano, coloca duas caixas diante dele e sai voando.

A caixa A é transparente e contém mil dólares.

A caixa B é opaca e contém um milhão de dólares ou nada.

Você pode levar as duas caixas ou apenas a caixa B.

A diferença está no fato de que Ômega colocou um milhão de dólares na caixa B se, e somente se, Ômega previu que você escolheria apenas a caixa B.

Ômega acertou em todas as 100 ocasiões observadas até agora – todos que escolheram ambas as caixas encontraram a caixa B vazia e receberam apenas mil dólares; todos que escolheram apenas a caixa B encontraram-na contendo um milhão de dólares. (Presumimos que a caixa A desaparece numa nuvem de fumaça se você escolher apenas a caixa B; ninguém mais poderá escolher a caixa A depois disso.)

Antes de fazer sua escolha, Ômega voa e parte para o próximo jogo. A caixa B já está vazia ou já está cheia.

Ômega deixa cair duas caixas no chão à sua frente e sai voando.

Você escolhe ambas as caixas ou apenas a caixa B?

A discussão filosófica padrão ocorre assim:

ADEPTO DA CAIXA ÚNICA: “Escolho apenas a caixa B, é claro. Prefiro ter um milhão do que mil.”

ADEPTO DAS DUAS CAIXAS: “Ômega já partiu. A caixa B já está cheia ou já está vazia. Se a caixa B já estiver vazia, escolher ambas as caixas me renderá $ 1.000, e escolher apenas a caixa B me renderá $ 0. Se a caixa B já estiver cheia, escolher ambas as caixas rende $ 1.001.000, escolher apenas a caixa B rende $ 1.000.000. Em qualquer caso, farei melhor se escolher ambas as caixas, e pior, se deixar mil dólares na mesa – então serei racional e ficarei com ambas as caixas.”

ADEPTO DA CAIXA ÚNICA: “Se você é tão racional, por que você é rico?”

ADEPTO DAS DUAS CAIXAS: “Não é minha culpa que Ômega opte por recompensar apenas pessoas com disposições irracionais, mas já é tarde demais para eu fazer algo a respeito.”

Existe uma vasta literatura sobre o tema dos problemas do tipo Newcomb – especialmente se considerarmos o Dilema do Prisioneiro como um caso especial, como geralmente é feito. Paradoxos de racionalidade e cooperação: o dilema do prisioneiro e o problema de Newcomb [1] é um volume editado que inclui o ensaio original de Newcomb. Para aqueles que leem apenas material online, a tese de doutorado de Ledwig resume as principais posições padrão [2].

Não adentrarei toda a literatura, mas o consenso dominante na moderna teoria da decisão é que se deve escolher ambas as caixas, e Ômega está simplesmente recompensando agentes com disposições irracio-

nais. Essa visão predominante é chamada de “teoria da decisão causal”.

Não tentarei apresentar minha própria análise aqui. Uma história muito longa, mesmo para os meus padrões.

Mas é concordado até mesmo entre os teóricos da decisão causal que, se tivermos o poder de nos comprometer previamente a aceitar uma caixa, no Problema de Newcomb, então deveríamos fazê-lo. Se você puder se comprometer previamente antes de Ômega o examinar, você estará diretamente preenchendo a caixa B.

Agora, em meu campo - que, caso tenha esquecido, é a IA automodificadora - isso significa que se você construir uma IA que tem duas caixas no Problema de Newcomb, ela se automodificará para uma caixa no Problema de Newcomb, se a IA antecipar que pode enfrentar tal situação. Agentes com acesso livre ao seu próprio código-fonte têm acesso a um método econômico de pré-compromisso.

E se você esperar poder enfrentar um problema do tipo Newcomb em geral, sem conhecer a forma exata do problema? Nesse caso, você precisaria se tornar um tipo de agente cuja disposição geralmente resultaria em altas recompensas em problemas desse tipo.

Mas como seria um agente com uma disposição geralmente adequada para problemas do tipo Newcomb? Isso pode ser formalmente especificado?

Sim, mas ao tentar escrever, percebi que estava prestes a criar um pequeno livro. E não era o livro mais importante que eu tinha que escrever, então o arquivei. Minha velocidade lenta de escrita é verdadeiramente a ruína da minha existência. A teoria que desenvolvi parece ter muitas propriedades interessantes, além de ser adequada para problemas do tipo Newcomb. Se eu conseguisse que alguém a aceitasse como minha tese de doutorado, seria uma tese impressionante. No entanto, isso seria necessário para eu cancelar o projeto. Caso contrário, não posso justificar o tempo gasto, especialmente considerando a velocidade com que escrevo livros atualmente.

Digo tudo isso porque existe uma atitude comum de que “argumentos verbais para caixa única são fáceis de encontrar; o que é difícil é desenvolver uma boa teoria de decisão que incorpore a escolha única” – uma matemática coerente que combine a escolha única com o problema de Newcomb sem produzir resultados absurdos em outros contextos. Então, eu entendo isso e decidi desenvolver tal teoria, mas minha velocidade de escrita em artigos extensos é tão lenta que não consigo publicá-la. Acredite ou não, é verdade.

No entanto, gostaria de apresentar algumas das minhas motivações no Problema de Newcomb – as razões pelas quais me senti compelido a procurar uma nova teoria – porque ilustram as minhas atitudes-fonte em relação à racionalidade. Mesmo que eu não consiga apresentar a teoria que essas motivações motivam...

Em primeiro lugar, acima de tudo, fundamentalmente, acima de tudo:

Os agentes racionais devem VENCER.

Não se engane pensando que estou falando sobre o estereótipo da Racionalidade de Hollywood, onde os racionalistas deveriam ser egoístas ou míopes. Se a sua função de utilidade inclui termos para outros, então busque a felicidade deles. Se a sua função de utilidade se estende por um milhão de anos, então vença na longa prazo.

Mas de qualquer forma, VENÇA. Não perca razoavelmente, **VENÇA.**

Agora, há defensores da teoria da decisão causal que argumentam que os dois jogadores estão fazendo o melhor que podem para vencer e não podem evitar se foram amaldiçoados por um Preditor que favorece os irracionalistas. Falarei sobre essa defesa em um momento. Mas primeiro, quero estabelecer uma distinção entre os teóricos da decisão causal que acreditam que os dois jogadores estão genuinamente fazendo o melhor para vencer e alguém que pensa que o double-boxing é a coisa razoável ou racional a se fazer, mas que o movimento razoável simplesmente perde previsivelmente, neste caso. Há muitas pessoas por aí que pensam que a racionalidade perde previsivelmente em vários problemas – isso também faz parte do estereótipo da Racionalidade de Hollywood, onde Kirk é previsivelmente superior a Spock.

A seguir, retomemos a acusação de que Omega favorece os irracionalistas. Consigo imaginar um superser que recompensa exclusivamente as pessoas nascidas com um gene específico, independentemente de suas escolhas. Consigo também conceber um superser que recompensa indivíduos cujos cérebros seguem o algoritmo específico de "Descreva suas opções em inglês e escolha a última opção quando ordenada alfabeticamente", mas que não recompensa aqueles que escolhem a mesma opção por uma razão diferente. No entanto, a Omega recompensa aqueles que optam por escolher apenas a caixa B, independentemente do algoritmo que utilizam para chegar a essa decisão. É por isso que não concordo com a acusação de que a Omega está recompensando o irracional. Omega não se importa se você segue ou não algum ritual específico de cognição; Omega apenas se preocupa com sua decisão prevista. Podemos escolher qualquer algoritmo de raciocínio que desejarmos, e seremos recompensados ou punidos apenas conforme as escolhas desse algoritmo, sem qualquer outra dependência - a Omega se importa apenas para onde vamos, não como chegamos lá.

É precisamente a ideia de que a Natureza não se importa com o nosso algoritmo que nos liberta para seguir o Caminho vencedor - sem ficar preso a qualquer ritual particular de cognição, exceto a nossa crença de que ele leva à vitória. Todas as regras estão disponíveis, exceto a regra da vitória.

Como Miyamoto Musashi afirmou - vale a pena repetir:

> Você pode vencer com uma arma longa, mas também pode vencer com uma arma curta. Em suma, a escola do Caminho do Ichi é o espírito de vencer, seja qual for a arma e seja qual for o seu tamanho[24] [3].

(Outro exemplo: McGee argumentou que devemos adotar funções de utilidade limitadas ou ficar sujeitos aos *"Dutch Books"* por tempos infinitos. No entanto, a função de utilidade não está disponível. Amo a vida sem limites ou restrições superiores; não há quantidade finita de vida vivida N onde eu preferiria uma probabilidade de 80,0001% de viver N anos a uma chance de 0,0001% de viver um ano Googleplex e uma chance de 80% de viver para sempre. Isso é uma condição suficiente para implicar que minha função de utilidade é ilimitada. Assim, só preciso descobrir como otimizar essa moralidade. Não se pode afirmar, primeiro, que acima de tudo devo me conformar a um ritual específico de cognição e depois que, se me conformar a esse ritual, devo mudar minha moralidade para evitar cair em uma reserva holandesa. Abandone o ritual de perder; não altere a definição de vitória. Isso é como decidir preferir US$ 1.000 a US$ 1.000.000 para que o Problema de Newcomb não torne seu ritual de cognição preferido desfavorável.)

"Mas", argumenta o teórico da decisão causal, "para escolher apenas a caixa B, você deve de alguma forma acreditar que sua escolha pode afetar se a caixa B está vazia ou cheia - e isso não é razoável! Omega já se foi! É fisicamente impossível!"

Irracional? Sou um racionalista: por que me importaria em ser irracional? Não preciso me conformar com um ritual específico de cognição. Não preciso escolher apenas a caixa B porque acredito que minha escolha afeta a caixa, mesmo que Omega já tenha partido. Posso simplesmente... escolher apenas a caixa B.

Tenho uma proposta de ritual alternativo de cognição que computa essa decisão, cuja margem é pequena demais para ser contida; mas não deveria ser necessário mostrar isso a você. A questão não é ter uma teoria elegante de vitória - a questão é vencer; a elegância é um efeito colateral.

Ou, olhando de outra forma: em vez de começar com um conceito de qual é a decisão razoável e depois perguntar se os agentes "razoáveis" saem com muito dinheiro, comece olhando para os agentes que saem com muito dinheiro, desenvolva uma teoria sobre quais agentes tendem a sair com mais dinheiro e, a partir dessa teoria, tente descobrir o que é "razoável". "Razoável" pode referir-se apenas a decisões conforme o nosso atual ritual de cognição - o que mais determinaria se algo parece "razoável" ou não?

Citando James Joyce (sem parentesco com o escritor irlandês), em *Foundations of Causal Decision Theory* (Fundamentos da Teoria da Decisão Causal) [4]:

> Rachel tem uma resposta perfeitamente boa para a pergunta 'Por que você não é rica?' 'Não sou rica', dirá ela, 'porque não sou o tipo de pessoa que o psicólogo pensa que recusará o dinheiro. Eu simplesmente não sou como você, Irene. Dado que sei que sou do tipo que aceita dinheiro e dado que o psicólogo sabe que sou deste

24 NT. Texto original em inglês. *You can win with a long weapon, and yet you can also win with a short weapon. In short, the Way of the Ichi school is the spirit of winning, whatever the weapon and whatever its size.*

> tipo, era razoável da minha parte pensar que os 1.000.000 dólares não estavam na minha conta. Os US$ 1.000 eram o máximo que eu conseguiria, não importa o que fizesse. Então, a única coisa razoável que eu podia fazer era aceitá-lo.
>
> Irene pode querer enfatizar o assunto perguntando: "Mas você não gostaria de ser como eu, Rachel? Você não gostaria de ser do tipo que recusa?" Há uma tendência para pensar que Rachel, uma teórica empenhada na decisão causal, deve responder a esta pergunta de forma negativa, o que parece obviamente errado (dado que ser como Irene a teria tornado rica). Este não é o caso. Rachel pode e deve admitir que gostaria de ser mais parecida com Irene. "Teria sido melhor para mim", ela poderia admitir, "se eu fosse do tipo que recusa". Neste ponto, Irene exclamará: "Você admitiu! Afinal, não foi tão inteligente aceitar o dinheiro. Infelizmente para Irene, a sua conclusão não segue a premissa de Rachel. Rachel explicará pacientemente que desejar recusar um problema de Newcomb não é inconsistente com pensar que se deve aceitar os US$ 1.000, seja qual for o tipo. Quando Rachel deseja ser o tipo de Irene, ela está desejando as opções de Irene, não sancionando sua escolha[25].

Diria que é um princípio geral da racionalidade – de fato, parte da minha definição de racionalidade – nunca invejarmos simplesmente as escolhas de outra pessoa. Pode-se invejar os genes de alguém, caso o Ômega recompense os genes ou se estes proporcionarem uma disposição geralmente mais feliz. Mas Rachel, acima de tudo, inveja a escolha de Irene, e somente a escolha dela, independentemente do algoritmo que Irene usou para fazê-la. Rachel gostaria apenas de ter a disposição para escolher de forma diferente.

Não se pode alegar ser mais racional do que alguém e, ao mesmo tempo, invejá-lo por sua escolha – apenas por sua escolha. Basta realizar o ato que você inveja.

Continuo tentando expressar que a racionalidade é o caminho para a vitória, mas os teóricos da decisão causal insistem que pegar as duas caixas é o que realmente vence, pois não é possível fazer melhor deixando US$ 1.000 na mesa... mesmo que os que escolheram apenas uma caixa saiam do experimento com mais dinheiro. Tenha cautela com esse tipo de argumento, sempre que você definir o "vencedor" como alguém que não seja o agente que está sorrindo no topo de uma pilha gigante de utilidades.

Sim, existem vários experimentos mentais nos quais alguns agentes começam com uma vantagem – mas se a tarefa é, por exemplo, decidir se devem pular de um penhasco, deve-se ter cuidado para não definir os agentes que se abstêm do penhasco como tendo uma injusta vantagem prévia sobre os agentes que pulam, devido à sua recusa injusta em saltar. Neste ponto, você redefiniu secretamente o "vencer" como conformidade com um ritual específico de cognição. Preste atenção ao dinheiro! Ou aqui está outra maneira de ver a questão: diante do problema de Newcomb, você gostaria realmente de procurar uma razão para acreditar que era perfeitamente razoável e racional escolher apenas a caixa B; porque, se tal linha de argumentação existisse, você pegaria apenas a caixa B e a encontraria cheia de dinheiro? Gastaria uma hora a mais pensando nisso, se tivesse certeza de que, no final da hora, seria capaz de se convencer de que a caixa B era a escolha racional? Esta também é uma posição bastante estranha para se estar. Normalmente, o trabalho da racionalidade consiste em descobrir qual escolha é a melhor – não em encontrar uma razão para acreditar que uma determinada escolha é a melhor.

Talvez seja demasiado fácil dizer que "devemos" aplicar duas caixas no Problema de Newcomb, que esta é a coisa "razoável" a fazer, desde que o dinheiro não esteja realmente à sua frente. Talvez você esteja insensível aos dilemas filosóficos neste momento. E se a sua filha tivesse uma doença 90% fatal e a caixa A

---

25 NT. Texto original em inglês. *Rachel has a perfectly good answer to the "Why ain't you rich?" question. "I am not rich," she will say, "because I am not the kind of person the psychologist thinks will refuse the money. I'm just not like you, Irene. Given that I know that I am the type who takes the money, and given that the psychologist knows that I am this type, it was reasonable of me to think that the $1,000,000 was not in my account. The $1,000 was the most I was going to get no matter what I did. So the only reasonable thing for me to do was to take it."*
Irene may want to press the point here by asking, "But don't you wish you were like me, Rachel? Don't you wish that you were the refusing type?" There is a tendency to think that Rachel, a committed causal decision theorist, must answer this question in the negative, which seems obviously wrong (given that being like Irene would have made her rich). This is not the case. Rachel can and should admit that she does wish she were more like Irene. "It would have been better for me," she might concede, "had I been the refusing type." At this point Irene will exclaim, "You've admitted it! It wasn't so smart to take the money after all." Unfortunately for Irene, her conclusion does not follow from Rachel's premise. Rachel will patiently explain that wishing to be a refuser in a Newcomb problem is not inconsistent with thinking that one should take the $1,000 whatever type one is. When Rachel wishes she was Irene's type she is wishing for Irene's options, not sanctioning her choice.

contivesse um soro com 20% de chance de curá-la, e a caixa B pudesse conter um soro com 95% de chance de curá-la? E se um asteroide estivesse avançando em direção à Terra, e a caixa A contivesse um defletor de asteroide que funcionasse 10% do tempo, e a caixa B pudesse conter um defletor de asteroide que funcionasse 100% do tempo?

Nesse ponto, você se sentiria tentado a fazer uma escolha irracional?

Se a aposta na caixa B representasse algo que você não pudesse simplesmente deixar para trás? Algo significativamente mais crucial para você do que ser simplesmente razoável? Se a vitória fosse realmente essencial, e não apenas ser rotulado como vencedor?

Você desejaria com todas as suas forças que a escolha "razoável" fosse simplesmente optar pela caixa B?

Então, talvez seja o momento de atualizar sua concepção de razoabilidade.

Os chamados racionalistas não deveriam invejar as simples decisões dos ditos não racionalistas, pois a sua escolha pode ser a que desejarem. Quando se encontra numa posição como essa, não deve repreender outra pessoa por não aderir aos seus conceitos de razoabilidade. Deve compreender que interpretou erroneamente o Caminho.

O mesmo se aplica se algum dia se encontrar mantendo um registro separado para a crença "razoável" contra a crença que parece realmente verdadeira. Ou interpretou de maneira equivocada a razoabilidade, ou a sua segunda intuição está simplesmente equivocada.

Agora, não podemos simultaneamente definir "racionalidade" como o Caminho vencedor e como a teoria bayesiana da probabilidade e a teoria da decisão. No entanto, é esse o argumento que estou apresentando, e a moral do meu conselho para confiar em Bayes, cujas leis que regem a vitória têm, de fato, se mostrado matemáticas. Se alguma vez se constatar que Bayes falha - obtendo recompensas sistematicamente mais baixas em algum problema em comparação com uma alternativa superior, devido às suas meras decisões - então Bayes terá que ser abandonado. "Racionalidade" é apenas o termo que uso para expressar minhas crenças sobre o Caminho vencedor - o Caminho do agente sorrindo no topo da gigantesca pilha de utilidade. Atualmente, esse termo refere-se ao Bayescraft.

Reconheço que esta não é uma crítica devastadora da teoria da decisão causal - isso exigiria um livro ou tese de doutorado próprios - mas espero que ilustre parte da minha atitude subjacente em relação a essa noção de "racionalidade".

[Edição 2015: Recentemente, escrevi uma exposição do tamanho de um livro sobre uma teoria da decisão que supera a teoria da decisão causal, a *Timeless Decision Theory* (Teoria da Decisão Atemporal) [5]. O criptógrafo Wei Dai respondeu com outra alternativa à teoria da decisão causal, a teoria da decisão sem atualização, que supera tanto a teoria da decisão causal quanto a atemporal. A partir de 2015, as discussões mais atualizadas sobre essas teorias são *Problem Class Dominance in Predictive Dilemmas* (Dominância de Classe de Problemas em Dilemas Preditivos) [6] de Daniel Hintze e *Toward Idealized Decision Theory* (Rumo a uma Teoria da Decisão Idealizada) de Nate Soares e Benja Fallenstein [7].]

Não se deve distinguir a escolha vencedora da escolha razoável. Nem se deve distinguir a crença razoável da crença com maior probabilidade de ser verdadeira.

É por isso que uso a palavra «racional» para expressar minhas crenças sobre precisão e vitória - não para denotar raciocínio verbal, estratégias que produzem sucesso garantido, ou aquilo que é logicamente demonstrável, ou aquilo que é demonstrável publicamente, ou aquilo que é razoável.

Miyamoto Musashi disse:

> A principal coisa quando você pega uma espada nas mãos é a sua intenção de cortar o inimigo, seja qual for o meio. Sempre que você desviar, acertar, saltar, golpear ou tocar a espada cortante do inimigo, você deve cortar o inimigo no mesmo movimento. É essencial conseguir isso. Se você pensar apenas em acertar, saltar, golpear ou tocar o inimigo, você não será capaz de realmente cortá-lo.

Referências

[1] Richmond Campbell and Lanning Snowden, eds., Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem (Vancouver: University of British Columbia Press, 1985).

[2] Marion Ledwig, "Newcomb's Problem" (PhD diss., University of Constance, 2000).

[3] Musashi, Book of Five Rings.

[4] James M. Joyce, The Foundations of Causal Decision Theory (New York: Cambridge University Press, 1999), doi:10.1017/CBO9780511498497.

[5] Yudkowsky, Timeless Decision Theory.

[6] Daniel Hintze, "Problem Class Dominance in Predictive Dilemmas," Honors thesis (2014).

[7] Nate Soares and Benja Fallenstein, "Toward Idealized Decision Theory," Technical report. Berkeley, CA: Machine Intelligence Research Institute (2014), http://intelligence.org/files/TowardIdealizedDecisionTheory.pdf.

Interlúdio: As doze virtudes da racionalidade

A primeira virtude é a curiosidade. Um desejo ardente de saber é mais forte do que uma solene promessa de buscar a verdade. Sentir o desejo ardente da curiosidade requer que você seja ignorante e que deseje se libertar da sua ignorância. Se em seu coração você acredita que já sabe, ou se em seu coração você não deseja saber, então seu questionamento será sem propósito e suas habilidades sem direção. A curiosidade busca se aniquilar; não há curiosidade que não queira uma resposta. A maravilha do mistério glorioso é que ele seja resolvido, depois do qual ele deixa de ser mistério. Tenha cuidado com aqueles que falam sobre serem de mente aberta e confessam modestamente sua ignorância. Há um momento para confessar sua ignorância e um momento para se libertar dela.

A segunda virtude é a renúncia. P. C. Hodgell disse: "Aquilo que pode ser destruído pela verdade deveria ser." [1] Não recue diante de experiências que possam destruir suas crenças. O pensamento que você não consegue pensar controla você mais do que os pensamentos que você expressa em voz alta. Submeta-se a provações e teste-se no fogo. Abandone a emoção que repousa em uma crença equivocada e procure sentir plenamente a emoção que corresponde aos fatos. Se o ferro se aproxima do seu rosto e você acredita que está quente, mas na verdade está frio, o Caminho se opõe ao seu medo. Se o ferro se aproxima do seu rosto e você acredita que está frio, mas na verdade está quente, o Caminho se opõe à sua calma. Avalie suas crenças primeiro e depois chegue às suas emoções. Deixe-se dizer: "Se o ferro está quente, desejo acreditar que está quente, e se está frio, desejo acreditar que está frio." Cuidado para não se apegar a crenças que você talvez não queira.

A terceira virtude é a leveza. Deixe que os ventos da evidência o levem como uma folha, sem nenhuma direção própria. Cuidado para não travar uma batalha defensiva contra a evidência, concedendo a contragosto cada centímetro de terreno somente quando for forçado, sentindo-se enganado. Renda-se à verdade o mais rápido possível. Faça isso no instante em que perceber o que está resistindo, no instante em que puder ver de qual direção os ventos da evidência estão soprando contra você. Seja infiel à sua causa e traia-a para um inimigo mais forte. Se você considera a evidência como uma restrição e busca se libertar, se vende para as correntes dos seus caprichos. Pois você não pode fazer um mapa preciso de uma cidade sentado em seu quarto de olhos fechados e desenhando linhas no papel conforme o impulso. Se, vendo a cidade de forma confusa, você pensa que pode mover uma linha um pouco para a direita, um pouco para a esquerda, de acordo com seu capricho, este é o mesmo erro.

A quarta virtude é a uniformidade. Aquele que deseja acreditar diz: "A evidência me permite acreditar?" Aquele que deseja desacreditar pergunta: "A evidência me obriga a acreditar?" Cuidado para não colocar grandes ônus de prova apenas em proposições que você não gosta, e então se defender dizendo: "Mas é bom ser cético." Se você prestar atenção apenas a evidências favoráveis, escolhendo e selecionando seus dados coletados, quanto mais dados você coletar, menos saberá. Se você é seletivo sobre quais argumentos inspeciona em busca de falhas, ou quão cuidadosamente os inspeciona em busca de falhas, então cada falha que aprender a detectar o torna um pouco mais burro. Se você escrever primeiro na parte inferior de uma folha de papel "E, portanto, o céu é verde!", não importa quais argumentos escreva acima dele depois; a conclusão já está escrita, e ela já está certa ou errada. Ser inteligente em argumento não é racionalidade, mas racionalização. A inteligência, para ser útil, deve ser usada para algo além de se destruir. Ouça as hipóteses enquanto elas argumentam seus casos diante de você, mas lembre-se de que você não é uma hipótese; você é o juiz. Portanto, não procure argumentar de um lado ou de outro, pois se você conhecesse seu destino, já estaria lá.

A quinta virtude é o argumento. Aqueles que desejam fracassar devem primeiro impedir que seus amigos os ajudem. Aqueles que sorriem sabiamente e dizem "Não vou discutir" afastam-se da ajuda e afastam-se do esforço comunitário. No argumento, esforce-se pela honestidade exata, em benefício dos outros e também de si mesmo: a parte de si mesmo que distorce o que você diz aos outros também distorce seus próprios pensamentos. Não acredite que está fazendo um favor aos outros se aceitar seus argumentos; o favor é para você. Não pense que ser justo com todos os lados significa se equilibrar igualmente entre as posições; a verdade não é distribuída em porções iguais antes do início de um debate. Você não pode avançar em questões factuais lutando com punhos ou insultos. Procure um teste que permita que a realidade julgue entre vocês.

A sexta virtude é o empirismo. As raízes do conhecimento estão na observação e seu fruto é a previsão. Que árvore cresce sem raízes? Que árvore nos nutre sem frutos? Se uma árvore cai na floresta e ninguém ouve, ela faz algum barulho? Alguém diz: "Sim, é verdade, pois produz vibrações no ar". Outro diz: "Não, não existe, pois não há processamento auditivo em nenhum cérebro". Embora discutam, um dizendo "Sim" e outro dizendo "Não", os dois não antecipam qualquer experiência diferente da floresta. Não pergunte quais crenças professar, mas quais experiências antecipar. Sempre saiba sobre qual diferença de experiência você está discutindo. Não deixe a discussão vagar e se concentrar em outra coisa, como a virtude de alguém como racionalista. Jerry Cleaver disse: "O que importa não é deixar de aplicar alguma técnica intrincada e complicada de alto nível. Está negligenciando o básico. Não ficar de olho na bola [2]." Não se deixe cegar pelas palavras. Quando as palavras são subtraídas, a expectativa permanece.

A sétima virtude é a simplicidade. Antoine de Saint-Exupéry disse: "A perfeição não é alcançada quando não há mais nada a acrescentar, mas quando não há mais nada a retirar [3]." A simplicidade é virtuosa na crença, no desígnio, no planeamento e na justificação. Quando você professa uma crença enorme com muitos detalhes, cada detalhe adicional é outra chance de a crença estar errada. Cada especificação adiciona ao seu fardo; se você puder aliviar seu fardo, deve fazê-lo. Não há palha que não tenha o poder de quebrar suas costas. Sobre artefatos, é dito: A engrenagem mais confiável é aquela que é projetada fora da máquina. Sobre planos: Uma teia emaranhada quebra. Uma cadeia de mil elos chegará a uma conclusão correta se cada passo estiver correto, mas se um passo estiver errado, ela pode levá-lo a qualquer lugar. Na matemática, uma montanha de boas ações não pode expiar um único pecado. Portanto, tenha cuidado em cada passo.

A oitava virtude é a humildade. Ser humilde é tomar medidas específicas em antecipação aos seus próprios erros. Confessar sua falibilidade e depois não fazer nada a respeito não é humildade; é se gabar de sua modéstia. Quem são os mais humildes? Aqueles que se preparam mais habilidosamente para os erros mais profundos e catastróficos em suas próprias crenças e planos. Porque este mundo contém muitos cuja compreensão da racionalidade é abismal, estudantes iniciantes de racionalidade vencem argumentos e adquirem uma visão exagerada de suas próprias habilidades. Mas é inútil ser superior: A vida não é avaliada em uma curva. O melhor físico da Grécia antiga não poderia calcular o caminho de uma maçã caindo. Não há garantia de que a adequação seja possível, dado o seu maior esforço; portanto, não pense se os outros estão piorando. Se você se comparar aos outros, não verá os vieses que todos os seres humanos compartilham. Ser humano é cometer dez mil erros. Ninguém neste mundo alcança a perfeição.

A nona virtude é o perfeccionismo. Quanto mais erros você corrige em si, mais você os nota. À medida que sua mente se torna mais silenciosa, você ouve mais ruído. Quando você nota um erro em si mesmo, isso sinaliza sua prontidão para buscar avanço para o próximo nível. Se você tolerar o erro em vez de corrigi-lo, não avançará para o próximo nível e não obterá a habilidade para notar novos erros. Em toda arte, se você não buscar a perfeição, você irá parar antes de dar o primeiro passo. Se a perfeição é impossível, isso não é desculpa para não tentar. Mantenha-se no mais alto padrão que você possa imaginar e busque um ainda mais alto. Não se contente com a resposta que está quase certa; procure uma que esteja exatamente certa.

A décima virtude é a precisão. Um vem e diz: A quantidade está entre 1 e 100. Outro diz: A quantidade está entre 1 e 100. Se a quantidade é 42, ambos estão corretos, mas a segunda previsão foi mais útil e se expôs a um teste mais rigoroso. O que é verdade para uma maçã pode não ser verdade para outra maçã; assim, mais pode ser dito sobre uma única maçã do que sobre todas as maçãs do mundo. As declarações mais precisas cortam mais fundo, a ponta afiada da lâmina. Como com o mapa, também com a arte da cartografia: O Caminho é uma Arte precisa. Não caminhe em direção à verdade, mas dance. Em cada passo dessa dança, seu pé pousa exatamente no lugar certo. Cada peça de evidência muda suas crenças pela quantidade exata-

mente certa, nem mais, nem menos. Qual é a quantidade exata? Para calcular isso, você deve estudar teoria da probabilidade. Mesmo que você não saiba fazer o cálculo, saber que a matemática existe mostra que o passo da dança é preciso e não tem espaço para seus caprichos.

A décima primeira virtude é a erudição. Estude muitas ciências e absorva seu poder como se fosse seu. Cada campo que você consome o torna maior. Se você engolir ciências suficientes, as lacunas entre elas diminuirão e seu conhecimento se tornará um todo unificado. Se você for glutão, se tornará mais vasto que as montanhas. É especialmente importante se alimentar de matemática e ciência que afetam a racionalidade: psicologia evolutiva, heurísticas e vieses, psicologia social, teoria da probabilidade, teoria da decisão. Mas estes não podem ser os únicos campos que você estuda. A Arte deve ter um propósito além de si mesma, ou entra em uma recursão infinita.

Antes dessas onze virtudes há uma virtude que não tem nome. Miyamoto Musashi escreveu em "O Livro dos Cinco Anéis" [4]:

> A principal coisa quando você pega uma espada nas mãos é a sua intenção de cortar o inimigo, seja qual for o meio. Sempre que você desviar, acertar, saltar, golpear ou tocar a espada cortante do inimigo, você deve cortar o inimigo no mesmo movimento. É essencial conseguir isso. Se você pensar apenas em acertar, saltar, golpear ou tocar o inimigo, você não será capaz de realmente cortá-lo. Mais do que tudo, você deve estar pensando em realizar seu movimento para cortá-lo[26].

Cada passo do seu raciocínio deve cortar até a resposta correta no mesmo movimento. Mais do que tudo, você deve pensar em levar seu mapa até refletir o território.

Se você não conseguir chegar a uma resposta correta, é inútil protestar que agiu com propriedade.

Como você pode melhorar sua concepção de racionalidade? Não dizendo para si mesmo: "É meu dever ser racional". Com isso você apenas consagra sua concepção equivocada. Talvez a sua concepção de racionalidade seja a de que é racional acreditar nas palavras do Grande Mestre, e o Grande Mestre diz: "O céu é verde", e você olha para o céu e vê azul. Se você pensa: "Pode parecer que o céu é azul, mas a racionalidade é acreditar nas palavras do Grande Mestre", você perde a chance de descobrir o seu erro.

Não pergunte se é "o Caminho" fazer isso ou aquilo. Pergunte se o céu é azul ou verde. Se você falar muito sobre o Caminho, você não o alcançará.

Você pode tentar nomear o princípio mais elevado com nomes como "o mapa que reflete o território" ou "experiência de sucesso e fracasso" ou "teoria da decisão Bayesiana." Mas talvez você descreva incorretamente a virtude sem nome. Como você descobrirá seu erro? Não comparando sua descrição consigo mesma, mas comparando-a com aquilo que você não nomeou.

Se por muitos anos você praticar as técnicas e se submeter a restrições rigorosas, pode ser que você vislumbre o centro. Então, você verá como todas as técnicas são uma técnica, e você se moverá corretamente sem se sentir restrito. Musashi escreveu: "Quando você aprecia o poder da natureza, conhecendo o ritmo de qualquer situação, você conseguirá atingir o inimigo naturalmente e golpear naturalmente. Tudo isso é o Caminho do Vazio."

Estas são então as doze virtudes da racionalidade:

Curiosidade, abandono, leveza, equanimidade, argumento, empirismo, simplicidade, humildade, perfeccionismo, precisão, erudição e o vazio.

---

26 NT. Texto original em inglês. *The primary thing when you take a sword in your hands is your intention to cut the enemy, whatever the means. Whenever you parry, hit, spring, strike or touch the enemy's cutting sword, you must cut the enemy in the same movement. It is essential to attain this. If you think only of hitting, springing, striking or touching the enemy, you will not be able actually to cut him. More than anything, you must be thinking of carrying your movement through to cutting him.*

Referências

[1] Patricia C. Hodgell, Seeker's Mask (Meisha Merlin Publishing, Inc., 2001).

[2] Cleaver, Immediate Fiction: A Complete Writing Course.

[3] Antoine de Saint-Exupery, Terre des Hommes (Paris: Gallimard, 1939).

[4] Musashi, Book of Five Rings.