LIVRO 3

A MÁQUINA NO FANTASMA

RACIONALIDADE

De A a Z

ELIEZER YUDKOWSKY

RACIONALIDADE DE A a Z

A MÁQUINA NO FANTASMA

LIVRO 3

por **ELIEZER YUDKOWSKY**

Tradução de Mariana Hungria

Revisão de Enéas Canavezzi Versehgi

Brasil, 2024

Sumário

Mentes: uma introdução

por Rob Besinger

Você é uma mente, e isso o coloca em uma situação bastante peculiar.

Poucas coisas podem ser consideradas mentes. Você é um elemento raro no universo. Você consegue fazer previsões e planos, refletir e revisar crenças, sofrer, sonhar, perceber detalhes como joaninhas ou experimentar um desejo súbito por manga. Dentro da sua própria mente, você até pode formar uma imagem completa dela. É possível raciocinar sobre o seu próprio processo de pensamento e trabalhar para alinhar suas operações com seus objetivos.

Você é uma mente implementada em um cérebro humano. E, apesar de toda a sua notável flexibilidade, o cérebro humano é algo sistemático. Ele é voltado para padrões e rotinas. Sua mente pode seguir uma rotina por toda a vida, sem sequer perceber que está presa nela. Essas rotinas podem ter consequências significativas.

Quando um padrão mental é útil para você, chamamos isso de "racionalidade"

Você existe como você é, programado para exibir certas formas de racionalidade e certas formas de irracionalidade, devido à sua ancestralidade. Toda a vida na Terra, incluindo você, descende de moléculas autorreplicantes antigas. Esse processo de replicação era inicialmente desajeitado e aleatório, gerando diferenças replicáveis entre os replicadores. "Evolução" é o nome que damos à mudança nessas diferenças ao longo do tempo.

Uma vez que algumas dessas diferenças replicáveis afetam a capacidade de reprodução — um fenômeno chamado "seleção" —, a evolução resultou em organismos adequados para se reproduzirem em ambientes semelhantes aos de seus ancestrais. Tudo em você é construído com base nos ecos das lutas e vitórias dos seus antepassados.

E aqui está você: uma mente moldada a partir de mentes mais fracas, buscando compreender seu próprio funcionamento interno, para que ele possa ser aprimorado — aprimorado em relação aos seus objetivos, e não aos do seu criador, a evolução. Que lições e percepções úteis podemos extrair ao saber que essa é a nossa situação fundamental?

Fantasmas e máquinas

Nossos cérebros, em sua estrutura e dinâmica em uma escala microscópica, assemelham-se a muitos outros sistemas mecânicos. No entanto, raramente pensamos em nossas mentes nos mesmos termos em que pensamos nos objetos em nosso ambiente ou nos órgãos em nosso corpo. Nossas categorias mentais básicas — crença, decisão, palavra, ideia, sentimento e assim por diante — possuem pouca semelhança com as nossas categorias físicas.

Filósofos do passado perceberam essa observação e a seguiram, argumentando que mentes e cérebros são fenômenos fundamentalmente distintos e separados. Essa visão é conhecida como "o dogma do fantasma na máquina"[1], conforme denominado pelo filósofo Gilbert Ryle. No entanto, cientistas e filósofos modernos que rejeitaram o dualismo, não necessariamente o substituíram por um modelo preditivo melhor

de como a mente funciona. Na prática, nossos propósitos e desejos ainda atuam como fantasmas flutuando livremente, como um magistério separado do restante de nosso conhecimento científico. Podemos falar sobre "racionalidade", "viés" e "como mudar nossas mentes", mas se essas ideias continuarem imprecisas e não forem limitadas por uma teoria abrangente, nossa linguagem que soa científica não nos protegerá de cometer os mesmos tipos de erros presentes nas teorias que incluem espíritos e essências.

Curiosamente, o mistério e a mistificação que envolvem as mentes não apenas obscurecem nossa visão dos seres humanos, mas também se aplicam a sistemas que parecem mentais ou propositais na biologia evolutiva e na inteligência artificial (IA). Talvez, se não pudermos compreender prontamente o que somos olhando para nós mesmos, possamos aprender mais usando processos claramente não humanos como um espelho.

Existem muitos fantasmas para aprender aqui — fantasmas do passado, do presente e ainda por vir. E essas ilusões são eventos cognitivos reais, fenômenos genuínos que podemos estudar e explicar. Se parece haver um fantasma na máquina, essa aparência é, em si, o trabalho oculto de uma máquina.

A primeira sequência de "A Máquina no Fantasma", intitulada "A Matemática Simples da Evolução", visa transmitir a dissonância e a discrepância entre nossa história hereditária, nossa biologia atual e nossas aspirações finais. Isso requer uma investigação mais profunda do que é comum nas introduções à evolução destinadas a não biólogos, que frequentemente se limitam às características superficiais da seleção natural.

A terceira sequência, intitulada "Um Guia Humano para Palavras", discute a relação fundamental entre cognição e formação de conceitos. Em seguida, há um ensaio mais extenso apresentando a inferência bayesiana.

Preenchendo a lacuna entre esses tópicos, "Propósitos Frágeis" abstrai da cognição e da evolução humana para a ideia de mentes e sistemas direcionados a objetivos em sua forma mais geral. Esses ensaios também servem ao propósito secundário de explicar a abordagem geral do autor em relação à filosofia e à ciência da racionalidade, a qual é amplamente informada por seu trabalho em Inteligência Artificial (IA).

Reconstruindo Inteligência

Eliezer Yudkowsky é um teórico de decisão e matemático que trabalha em questões fundamentais de Inteligência Artificial Geral (AGI). A AGI é o estudo teórico de sistemas que podem resolver problemas gerais em diversos domínios. O trabalho de Yudkowsky em IA tem sido uma força motriz importante por trás de sua exploração da psicologia da racionalidade humana, como ele destacou em sua primeira postagem no blog sobre Superar o Viés: "A Arte Marcial da Racionalidade":

"A compreensão que adquiri sobre a racionalidade foi desenvolvida durante a luta com o desafio da Inteligência Artificial Geral (uma empreitada que, para ser verdadeiramente bem-sucedida, requer um domínio suficiente da racionalidade para construir um racionalista completo e funcional com palitos de dente e elásticos). Em muitos aspectos, o problema da IA é muito mais exigente do que a arte pessoal da racionalidade, mas, de certa forma, também é mais fácil. Na arte marcial da mente, precisamos adquirir a habilidade de processar informações em tempo real, de acionar as alavancas certas no momento adequado em uma grande máquina pensante pré-existente, cujos mecanismos internos não podem ser modificados pelo usuário final. Parte desse aparato está otimizado para pressões de seleção evolutiva que vão diretamente contra nossos objetivos declarados ao utilizá-lo. Decidimos conscientemente buscar apenas a verdade. No entanto, nossos cérebros possuem tendências programadas para racionalizar falsidades. […]"

Tentar sintetizar uma arte pessoal da racionalidade, utilizando a ciência da racionalidade, pode se mostrar incômodo: imagine tentar inventar uma arte marcial usando uma teoria abstrata da física, da teoria dos jogos e da anatomia humana. No entanto, os seres humanos não são meros reflexos cegos; possuímos um instinto inato de introspecção. O olho interior não é cego, mas enxerga embaçado, com distorções sistemáticas. Portanto, precisamos aplicar a ciência às nossas intuições, utilizando o conhecimento abstrato para corrigir nossos movimentos mentais e aprimorar nossas habilidades metacognitivas. Não estamos escrevendo um software para fazer um fantoche executar movimentos de arte marcial; são nossos próprios membros mentais que devemos movimentar. Portanto, devemos conectar a teoria à prática. Devemos enxergar o que

a ciência realmente significa para nós mesmos, para nossa vida interior diária.

Da perspectiva de Yudkowsky, falar sobre a racionalidade humana sem dizer nada de interessante sobre a IA é tão difícil quanto falar sobre a IA sem dizer nada de interessante sobre racionalidade.

A longo prazo, Yudkowsky prevê que a IA ultrapassará os humanos em uma "explosão de inteligência", um cenário no qual a IA automodificadora aprimora sua própria capacidade de se reprojetar produtivamente, desencadeando uma rápida sucessão de melhorias adicionais. O termo "singularidade tecnológica" é às vezes usado como sinônimo de "explosão de inteligência". Até janeiro de 2013, o MIRI era conhecido por outro nome. Ele se chamava "Instituto da Singularidade para Inteligência Artificial" e sediava a Cúpula anual da Singularidade. Desde então, Yudkowsky passou a preferir o termo mais antigo de I.J. Good, "explosão de inteligência", para diferenciar seus pontos de vista de outras previsões futuristas, como a tese do progresso tecnológico exponencial de Ray Kurzweil.[2]

Tecnologias como a IA mais inteligente do que o ser humano parecem resultar em grandes mudanças sociais, para melhor ou para pior. Yudkowsky cunhou o termo "teoria da IA amigável" para se referir à pesquisa de técnicas que visam alinhar as preferências de uma IA Geral com as preferências humanas. Até o momento, sabemos muito pouco sobre quando um software geralmente inteligente poderá ser inventado e quais abordagens de segurança seriam eficazes nesses casos. A IA autônoma atual já pode ser bastante desafiadora para ser verificada e validada com muita confiança, e muitas técnicas atuais provavelmente não serão generalizadas para sistemas mais inteligentes e adaptáveis. Portanto, a "IA amigável" está mais próxima de um conjunto de questões matemáticas e filosóficas básicas do que de um conjunto bem definido de objetivos de programação.

Desde 2015, as opiniões de Yudkowsky sobre o futuro da IA continuam a ser debatidas por analistas de tecnologia e pesquisadores de IA na indústria e no meio acadêmico, sem que tenha havido uma convergência para uma posição consensual. O livro de Nick Bostrom, "Superinteligência" oferece uma visão geral das muitas questões morais e estratégicas levantadas pela IA mais inteligente do que os seres humanos.[3]

Para uma introdução geral ao campo da IA, o livro didático mais amplamente utilizado é "Inteligência Artificial: Uma Abordagem Moderna" [4], de Russell e Norvig. Em um capítulo que discute as questões morais e filosóficas levantadas pela IA, Russell e Norvig observam a dificuldade técnica de especificar um bom comportamento em uma IA altamente adaptável:

> [Yudkowsky] afirma que a amizade (o desejo de não prejudicar os humanos) deve ser incorporada desde o início, mas os projetistas devem reconhecer que seus próprios projetos podem ter falhas e que o robô aprenderá e evoluirá com o tempo. Portanto, o desafio é um projeto de mecanismo — definir um mecanismo para a evolução dos sistemas de IA sob um sistema de freios e contrapesos e fornecer aos sistemas funções de utilidade que permaneçam amigáveis diante dessas mudanças. Não podemos simplesmente atribuir a um programa uma função de utilidade estática, pois as circunstâncias e nossas respostas desejadas a essas circunstâncias mudam com o tempo.[1]

Incomodados com a possibilidade de que avanços futuros em IA, nanotecnologia, biotecnologia e outros campos possam representar riscos à civilização humana, Bostrom e Ćirković compilaram a primeira antologia acadêmica sobre o tema, intitulada *Global Catastrophic Risks* (Riscos Globais Catastróficos) [5]. Os riscos mais extremos são chamados de riscos existenciais, riscos que podem resultar na estagnação permanente ou até mesmo na extinção da humanidade.[6]

As pessoas, incluindo especialistas, tendem a ser extremamente ruins para prever grandes eventos futuros, incluindo novas tecnologias. Uma parte do objetivo de Yudkowsky ao discutir a racionalidade é iden-

1 NT: Texto original em inglês. *[Yudkowsky] asserts that friendliness (a desire not to harm humans) should be designed in from the start, but that the designers should recognize both that their own designs may be flawed, and that the robot will learn and evolve over time. Thus the challenge is one of mechanism design—to define a mechanism for evolving AI systems under a system of checks and balances, and to give the systems utility functions that will remain friendly in the face of such changes. We can't just give a program a static utility function, because circumstances, and our desired responses to circumstances, change over time.*

tificar nossos preconceitos. Esses preconceitos interferem em nossa capacidade de prever e nos preparar para grandes mudanças com antecedência. Suas contribuições para o livro *Global Catastrophic Risks*, intituladas "Vieses cognitivos que podem afetar o julgamento de riscos globais" e "Inteligência artificial como um fator positivo e negativo no risco global", combinam sua pesquisa em ciência cognitiva e IA. Yudkowsky e Bostrom resumem as preocupações tanto de curto quanto de longo prazo, em um capítulo do "Manual de Inteligência Artificial de Cambridge" intitulado "A ética da inteligência artificial". [7]

Embora este livro trate da racionalidade humana, o tema da IA é relevante como uma fonte de ilustrações simples dos aspectos da cognição humana. A previsão de tecnologia de longo prazo também é uma aplicação importante da racionalidade Bayesiana, que pode modelar o raciocínio correto, mesmo em domínios onde os dados são escassos ou ambíguos.

Conhecer o projeto pode revelar muito sobre o projetista, e conhecer o projetista pode revelar muito sobre o projeto.

Portanto, começaremos explorando o que nosso próprio projetista pode nos ensinar sobre nós mesmos.

Referências

[1] Gilbert Ryle, The Concept of Mind (University of Chicago Press, 1949).

[2] Irving John Good, "Speculations Concerning the First Ultraintelligent Machine," in Advances in Computers, ed. Franz L. Alt and Morris Rubinoff, vol. 6 (New York: Academic Press, 1965), 31–88, doi:10.1016/S0065-2458(08)60418-0.

[3] Nick Bostrom, Superintelligence: Paths, Dangers, Strategies (Oxford University Press, 2014).

[4] Stuart J. Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, 3rd ed. (Upper Saddle River, NJ: Prentice-Hall, 2010).

[5] Bostrom and Ćirković, Global Catastrophic Risks.

[6] Um exemplo de risco existencial é o cenário conhecido como "grey goo" (lama cinzenta), no qual robôs moleculares projetados para uma autorreplicação eficiente desempenham sua função de maneira tão eficaz que acabam superando rapidamente os organismos vivos ao consumirem toda a matéria disponível na Terra.

[7] Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence," in The Cambridge Handbook of Artificial Intelligence, ed. Keith Frankish and William Ramsey (New York: Cambridge University Press, 2014).

Interlúdio: o poder da inteligência

Em nossos crânios, carregamos aproximadamente três quilos de tecido viscoso, úmido e cinzento, enrugado como papel higiênico amassado.

À primeira vista, olhando para esse nódulo pouco atraente, você não pensaria que ele é uma das coisas mais poderosas do universo conhecido. Se você nunca tivesse visto um livro de anatomia e encontrasse um cérebro caído na rua, provavelmente diria "Eca!" e tentaria não sujar seus sapatos. Até mesmo Aristóteles acreditava que o cérebro era um órgão responsável por resfriar o sangue. Não parece perigoso.

Há cinco milhões de anos, os ancestrais dos leões governavam o dia, enquanto os ancestrais dos lobos vagavam durante a noite. Os predadores dominantes estavam equipados com dentes e garras — bordas cortantes afiadas e duras, apoiadas por músculos poderosos. Em autodefesa, suas presas desenvolveram conchas blindadas, chifres afiados, venenos tóxicos e camuflagem. A guerra durou centenas de eras e incontáveis corridas armamentistas. Muitos perdedores foram eliminados do jogo, mas não havia indícios de um vencedor. Onde uma espécie possuía conchas, outra evoluía para quebrá-las; onde uma espécie se tornava venenosa, outra evoluía para tolerar o veneno. Cada espécie ocupava seu nicho particular — afinal, quem poderia viver nos mares, nos céus e na terra ao mesmo tempo? Não havia arma ou defesa definitiva, e não havia razão para acreditar que isso seria possível.

Então chegou o "*Dia das Coisas Macias*".

Eles não possuíam armaduras, não tinham garras e nem veneno.

Se você assistisse a um filme de uma explosão nuclear e te dissessem que uma forma de vida terrestre era a responsável, você nunca imaginaria, nem em seus sonhos mais loucos, que as Coisas Macias poderiam ser as culpadas. Afinal, "*Coisas Macias*" não são radioativas.

No início, as *Coisas Macias* não tinham caças, metralhadoras, rifles ou espadas. Sem bronze, sem ferro. Sem martelos, sem bigornas, sem alicates, sem forjas, sem minas. Todas as *Coisas Macias* tinham dedos moles — muito fracos para quebrar uma árvore, muito menos uma montanha. Claramente, não eram perigosas. Para cortar pedra, seria necessário aço, e as *Coisas Macias* não poderiam produzir aço. Não havia lâminas de aço no ambiente para os dedos *Macios* pegarem. Seus corpos não podiam gerar temperaturas nem mesmo próximas o suficiente para derreter o metal. O cenário era obviamente absurdo.

E quanto às *Coisas Macias* manipulando DNA — isso seria além do ridículo. Dedos macios não são tão pequenos. Não havia acesso ao nível do DNA para as *Coisas Macias*; seria como tentar agarrar um átomo de hidrogênio. Ah, tecnicamente, tudo faz parte do mesmo universo, tecnicamente as *Coisas Macias* e o DNA estão inseridos no mesmo mundo, nas mesmas leis unificadas da física, na mesma grande teia de causalidade. Mas sejamos realistas: você não pode chegar lá a partir daqui.

Mesmo que as *Coisas Macias* evoluíssem para realizar essas façanhas algum dia, levaria milhares de milênios. Observamos o fluxo e refluxo da Vida através dos éons, e deixe-me dizer a você, um ano não é nem mesmo um único tique-taque do tempo evolutivo. Ah, claro, tecnicamente um ano é composto por seiscentos trilhões de trilhões, de trilhões de trilhões de intervalos de Planck. Mas nada acontece em menos de seiscentos milhões de trilhões de trilhões de trilhões de trilhões de intervalos de Planck, então é um ponto discutível. As *Coisas Macias*, enquanto correm pela savana agora, não estarão voando pelos continentes por pelo menos mais dez milhões de anos; ninguém poderia fazer tanto sexo.

Agora, explique-me mais uma vez porque uma Inteligência Artificial não pode realizar nada interessante na Internet, a menos que um programador humano construa um corpo de robô para ela.

Percebo que a reação inicial de alguém em relação à "inteligência" — o pensamento que passa pela mente nos primeiros instantes após ouvir a palavra "inteligência" — determina geralmente a sua reação à ideia de uma explosão de inteligência. Frequentemente, eles buscam a palavra-chave "inteligência" e associam-na à concepção convencional de "esperteza" — uma imagem mental do grande mestre de xadrez que não consegue sair com uma garota ou do professor universitário que não consegue sobreviver fora do meio acadêmico.

"As pessoas costumam dizer: 'Inteligência não é suficiente para ter sucesso profissional', como se o carisma estivesse nos rins e não no cérebro. Elas afirmam: 'A inteligência não é páreo para uma arma', como se as armas crescessem em árvores. E perguntam: 'Onde uma Inteligência Artificial conseguiria dinheiro?', como se o primeiro *Homo sapiens* tivesse encontrado notas de dinheiro flutuando do céu e as usasse em lojas de conveniência nas florestas. Nossa espécie não nasceu em uma economia de mercado. As abelhas não venderiam mel se lhes oferecessem uma transferência eletrônica de fundos. A espécie humana, imaginou a existência do dinheiro, e ele existe — para nós, não para ratos ou vespas — porque continuamos acreditando nele.

Continuo tentando explicar às pessoas que o arquétipo da inteligência não é o Dustin Hoffman[2] em "*Rain Man*". É um ser humano, ponto. São seres moles que explodem no vácuo, deixando pegadas em sua lua. Dentro dessa massa cinzenta e úmida reside o poder de traçar caminhos através da complexa teia de causalidade e encontrar soluções para o aparentemente impossível — poder esse que chamamos de criatividade.

Às vezes, as pessoas — especialmente os capitalistas de risco — questionam como os resultados de uma verdadeira IA construída pelo MIRI (Instituto de Pesquisa em Inteligência de Máquina) [3] seriam comercializados. Isso é o que chamamos de problema de enquadramento.

Ou talvez seja algo mais profundo do que um simples choque de suposições. Com um pouco de pensamento criativo, as pessoas podem imaginar como viajar para a Lua, curar a varíola ou fabricar computadores. No entanto, imaginar um truque que pudesse realizar todas essas coisas ao mesmo tempo, parece totalmente impossível — mesmo que tal poder esteja a poucos centímetros atrás de seus próprios olhos. A coisa cinza e úmida ainda parece misteriosa para a própria coisa cinza e úmida.

E assim, porque as pessoas não conseguem ver como tudo isso funcionaria, o poder da inteligência parece menos real; mais difícil de imaginar do que uma torre de fogo lançando uma nave para Marte. A perspectiva de visitar Marte cativa a imaginação. No entanto, se alguém prometesse uma visita a Marte, além de uma grande teoria unificada da física, uma prova da hipótese de Riemann, uma cura para a obesidade, para o câncer e para o envelhecimento, e ainda uma cura para a estupidez… bem, isso parece bem errado, só isso.

E, bem, deveria parecer estranho mesmo. É uma falta de imaginação enorme pensar que a inteligência serve para tão pouco. Quem poderia ter imaginado, há tanto tempo, o que as mentes conseguiriam fazer um dia? Talvez nem saibamos quais são nossos verdadeiros problemas.

Mas, enquanto isso, por ser difícil de entender como um processo pode ter poderes tão diversos, também é difícil imaginar que com um único truque poderíamos resolver de uma só vez até mesmo problemas prosaicos como a obesidade, o câncer e o envelhecimento.

No entanto, um truque curou a varíola, construiu aviões, cultivou trigo e domou o fogo. Nossa ciên-

---

2 NT. Dustin Hoffman interpreta Raymond Babbitt, um autista savant com habilidades matemáticas extraordinárias, em "*Rain Man*" (1988), dirigido por Barry Levinson. Sua atuação retrata de forma sensível e realista as nuances do personagem, destacando sua relação complexa com o irmão, Charlie, e conquistando reconhecimento crítico, incluindo o Oscar de Melhor Ator.

3 NT: O MIRI (Machine Intelligence Research Institute) é uma organização de pesquisa sem fins lucrativos focada em garantir que o desenvolvimento da inteligência artificial avançada beneficie a humanidade, com ênfase em segurança e alinhamento de sistemas de IA. Fundado em 2000, o instituto busca soluções teóricas e práticas para os desafios associados à criação de inteligência artificial geral (AGI).

cia atual ainda pode não concordar completamente sobre como exatamente esse truque funciona, mas ele funciona mesmo assim. Se você está temporariamente ignorante sobre um fenômeno, isso é um fato sobre o seu estado mental atual, não um fato sobre o fenômeno em si. Um mapa em branco não corresponde a um território em branco. Mesmo que alguém não compreenda completamente esse poder responsável por deixar pegadas na Lua, ainda assim, as pegadas continuam lá — pegadas reais, em uma Lua real, deixadas por um poder real. Se alguém compreendesse com profundidade suficiente, poderia criar e moldar esse poder. A inteligência é tão real quanto a eletricidade. Ela apenas é muito mais poderosa, mais perigosa e tem implicações muito mais profundas para o desenvolvimento da vida no universo — e é um pouquinho mais difícil descobrir como construir um gerador.

L — A matemática simples da evolução

131 — Um Deus alienígena

"Um aspecto curioso da teoria da evolução", disse Jacques Monod," é que todos pensam que a entendem".

Um ser humano, ao observar o mundo natural, enxerga um propósito mil vezes maior. Pernas de coelho, construídas e articuladas para correr; mandíbulas de raposa, construídas e articuladas para rasgar. Mas o que se vê não é exatamente o que está lá…

Nos tempos anteriores a Darwin, a causa por trás de todo esse propósito aparente era um grande enigma para a ciência. Os monoteístas diziam "Deus fez isso", porque se ganhava 50 pontos de bônus cada vez que se usava a palavra "Deus" em uma frase. Talvez eu esteja sendo injusto. Nos tempos anteriores a Darwin, parecia uma hipótese muito mais razoável. Encontre um relógio no deserto, dizia William Paley, e você poderá inferir a existência de um relojoeiro.

No entanto, ao olhar para todo o propósito aparente da Natureza, em vez de escolher e selecionar exemplos, você começa a perceber coisas que não se encaixam no conceito judaico-cristão de um Deus benevolente. As raposas parecem bem projetadas para capturar coelhos. Os coelhos parecem bem projetados para fugir das raposas. Será que o Criador estava com dificuldades para decidir?

Quando projeto uma torradeira, não incluo uma parte que tenta levar eletricidade às bobinas e outra parte que tenta evitar que a eletricidade chegue às bobinas. Seria um desperdício de esforço. Quem projetou o ecossistema, com seus predadores e presas, vírus e bactérias? Até mesmo o cacto, que pode ser considerado bem projetado para fornecer água e frutas aos animais do deserto, é coberto por espinhos inconvenientes.

O ecossistema faria muito mais sentido se não tivesse sido projetado por um único "Quem", mas sim criado por uma horda de divindades — digamos, das religiões hindu ou xintoísta. Isso explicaria facilmente tanto os propósitos onipresentes quanto os conflitos onipresentes: múltiplas divindades agindo, frequentemente com propósitos opostos. A raposa e o coelho foram projetados, mas por divindades concorrentes distintas. Pergunto-me se alguém já comentou sobre a evidência aparentemente excelente fornecida dessa forma para o hinduísmo em detrimento do cristianismo. Provavelmente não.

Da mesma forma, o Deus judaico-cristão é considerado benevolente — bem, mais ou menos. No entanto, grande parte do propósito da natureza parece absolutamente cruel. Darwin desconfiou de um Criador fora do padrão ao estudar as vespas Icnêumon, cujas picadas paralisantes preservam suas presas para serem devoradas vivas por suas larvas: "Não consigo me convencer", escreveu Darwin, "de que um Deus benevolente e onipotente teria criado intencionalmente os *Ichneumonoidea*[4] com a intenção expressa de se alimentarem dos corpos vivos das Lagartas, ou que um gato devesse brincar com ratos." [1] Eu me pergunto se algum pensador anterior já comentou sobre a excelente evidência assim fornecida para as religiões maniqueístas em detrimento das monoteístas.

Neste ponto, todos conhecemos a piada: basta dizer "evolução".

Preocupo-me com o fato de que é assim que algumas pessoas estão absorvendo a explicação "científi-

4 NT. *Ichneumonoidea* é uma superfamília de vespas parasitóides, conhecidas por depositar seus ovos em outros insetos (ou aracnídeos). As larvas da vespa se desenvolvem dentro do hospedeiro, consumindo-o e eventualmente o matando.

fica", como uma fábrica de propósitos mágicos na natureza. Já discuti anteriormente o exemplo da personagem Tempestade no filme *X-Men*[5], que, por meio de uma mutação, adquire a habilidade de lançar raios. Por quê? Bem, existe algo chamado "evolução" que, de alguma forma, injeta propósito na natureza e as mudanças ocorrem por meio de "mutações". Portanto, se Tempestade sofrer uma mutação significativa, ela poderá ser reprojetada para lançar raios. A radioatividade é frequentemente citada como uma causa super poderosa: a radiação causa mutações, então uma radiação mais intensa causaria mutações ainda mais poderosas. Isso parece lógico.

Mas a evolução não permite que qualquer forma de propósito se infiltre na natureza. É isso que torna a evolução um sucesso como uma hipótese empírica. Se a biologia evolutiva pudesse explicar não apenas uma árvore, mas também uma torradeira, seria inútil como teoria. Há muito mais na teoria da evolução do que simplesmente apontar para a natureza e dizer: "Agora o propósito é permitido" ou "A evolução fez isso!". A força de uma teoria não reside no que ela permite, mas no que ela proíbe. Se você puder inventar uma explicação igualmente persuasiva para qualquer resultado, então você não tem conhecimento algum.

George Williams observou que muitos não-biólogos têm a falsa noção de que os chocalhos nas caudas das cascavéis são uma característica benéfica para a própria cascavel. [2] No entanto, esse tipo de propósito não é permitido pela evolução. A evolução não funciona permitindo que lampejos de propósito surjam ao acaso, remodelando uma espécie para o benefício de um destinatário aleatório.

A evolução é impulsionada pela correlação sistemática entre os diferentes genes e como eles constroem os organismos, e quantas cópias desses genes são transmitidas para a próxima geração. Para os chocalhos crescerem nas caudas das cascavéis, os genes relacionados aos chocalhos devem se tornar cada vez mais frequentes em cada geração sucessiva. (Na verdade, os genes para chocalhos se tornam cada vez mais complexos. No entanto, descrever todos os detalhes e ressalvas da biologia evolutiva seria uma tarefa demorada.)

Não existe uma "Fada da Evolução", que examine o estado atual da natureza, decide o que seria uma "boa ideia" e opta por aumentar a frequência dos genes construtores de chocalhos.

Suspeito que muitas pessoas fiquem presas nesse entendimento da biologia evolutiva. Elas compreendem que genes "úteis" se tornam mais comuns, mas "úteis" permitem que qualquer tipo de propósito se infiltre. Elas não acreditam que haja uma Fada da Evolução, mas ainda assim perguntam quais genes serão "úteis", como se um gene de cascavel pudesse "ajudar" outros seres.

A principal percepção é não haver uma Fada da Evolução. Não existe uma força externa que decida quais genes devem ser promovidos. O que acontece, acontece devido aos próprios genes.

Os genes responsáveis pela formação dos chocalhos (cada vez mais aprimorados) devem ter se tornado mais comuns no *pool* genético das cascavéis devido aos chocalhos. Isso ocorre provavelmente porque as cascavéis com chocalhos mais eficientes têm uma maior taxa de sobrevivência, em vez de acasalarem ou terem irmãos que se reproduzem com mais sucesso, entre outros fatores.

É possível que os predadores desconfiem dos chocalhos e evitem pisar nas cobras. Ou talvez os chocalhos desviem a atenção da cabeça da cobra. (como sugeriu George Williams: "O desfecho de uma luta entre um cachorro e uma víbora dependeria muito se o cachorro agarrasse inicialmente o réptil pela cabeça ou pelo rabo".)

Mas isso é apenas um exemplo de chocalho de cobra. Existem maneiras muito mais complexas pelas quais um gene pode aumentar a frequência de suas cópias na próxima geração. Seu irmão ou irmã compartilha metade dos seus genes. Um gene que sacrifica uma unidade de recursos para beneficiar três unidades de recursos em um irmão pode promover suas cópias ao sacrificar um organismo construído. (Se você deseja conhecer todos os detalhes e nuances, recomendo a leitura de um livro sobre biologia evolutiva; não há atalhos.)

O ponto principal é que o efeito do gene deve levar ao aumento da frequência das suas cópias na

5 NT: "X-Men" é um filme de super-heróis baseado nos quadrinhos da Marvel, que segue um grupo de mutantes com habilidades especiais, liderados pelo Professor Xavier, enquanto enfrentam conflitos internos e a ameaça de Magneto, que busca dominar a humanidade. A narrativa explora temas como preconceito, coexistência e aceitação.

próxima geração. Não há nenhuma Fada da Evolução que intervenha. Não há nada que decida quais genes são "úteis" e que devem, portanto, se tornar mais frequentes. É apenas uma questão de causa e efeito, a partir dos próprios genes.

Isso explica o estranho propósito conflitante da Natureza e sua crueldade frequente. Explica melhor do que uma horda de divindades xintoístas.

Por que existe tanta guerra na Natureza? Porque não há uma única Evolução conduzindo todo o processo. Existem tantas "evoluções" diferentes quanto populações reprodutivas. Os genes dos coelhos estão se tornando mais ou menos frequentes nas populações de coelhos. Os genes das raposas estão se tornando mais ou menos frequentes nas populações de raposas. Os genes das raposas que constroem raposas que capturam coelhos inserem mais cópias de si mesmos na próxima geração. Os genes dos coelhos que constroem coelhos que conseguem escapar das raposas são naturalmente mais comuns na próxima geração de coelhos. Daí surge o termo "seleção natural".

Por que a Natureza é cruel? Você, como ser humano, pode olhar para uma vespa *Icnêumon* e considerar cruel ela se alimentar de sua presa viva. Você pode pensar que, se ela vai se alimentar da presa viva, pelo menos poderia impedi-la de sofrer. Certamente, não seria difícil para a vespa anestesiar sua presa e paralisá-la. E o que dizer dos elefantes idosos que morrem de fome quando perdem os últimos dentes? Esses elefantes não vão se reproduzir, de qualquer forma. Quanto custaria à evolução — ou melhor, à evolução dos elefantes — garantir que o elefante morresse imediatamente, em vez de lenta e agonizantemente? Quanto custaria à evolução anestesiar o elefante ou proporcionar-lhe sonhos agradáveis antes da morte? Nada; aquele elefante não vai se reproduzir mais ou menos de qualquer forma.

Se você estivesse em uma conversa com um ser humano, tentando resolver um conflito de interesses, estaria em uma posição favorável para negociar — teria um trabalho fácil de persuasão. Custaria tão pouco anestesiar a presa, permitir que o elefante morresse sem sofrimento! Ah, por favor, você não faria isso, por gentileza… hum…

Mas não há com quem argumentar.

Os seres humanos falsificam suas justificativas, encontram o que querem usando um método e depois justificam usando outro método. Não existe uma Fada da Evolução dos Elefantes que: (a) determina o melhor para eles e depois (b) encontra maneiras de justificar suas decisões para o Supervisor Evolucionário. Esse Supervisor (c) não tem interesse em diminuir a capacidade reprodutiva dos elefantes, mas (d) concorda com a ideia de uma morte sem dor, desde que isso não afete nenhum gene.

Não há defensor dos elefantes em lugar algum no sistema.

Os seres humanos, que frequentemente se preocupam profundamente com o bem-estar dos animais, podem ser persuasivos ao argumentar que várias atitudes benevolentes não afetariam a aptidão reprodutiva. Infelizmente, a evolução dos elefantes não segue um algoritmo semelhante; ela não seleciona bons genes que podem ser argumentados de forma plausível para melhorar a aptidão reprodutiva. Simplesmente: os genes que se replicam com mais frequência se tornam mais comuns na próxima geração. É como a água fluindo morro abaixo, igualmente benevolente.

Um ser humano, ao observar a Natureza, começa a pensar em todas as maneiras como projetaríamos organismos. E então, tendemos a justificar por que nossas melhorias de projeto aumentariam a aptidão reprodutiva — é um instinto político, tentando vender nossa própria opção preferida como compatível com a justificativa preferida do chefe.

E assim, os biólogos evolutivos amadores fazem previsões maravilhosas e completamente equivocadas. Isso ocorre porque os biólogos amadores estão traçando os seus resultados, e o mais importante, localizando suas previsões no espaço de hipóteses, usando um algoritmo diferente daquele usado pelas evoluções para traçar seus resultados.

Um engenheiro humano teria projetado papilas gustativas humanas para medir a quantidade de cada nutriente que temos e a quantidade que precisamos. Quando a gordura era escassa, amêndoas ou cheeseburgers eram deliciosos. Mas se você começasse a ficar obeso ou se faltassem vitaminas, a alface seria

saborosa. No entanto, não existe uma Fada da Evolução dos Humanos que tenha sabiamente planejado e projetado um sistema abrangente para cada eventualidade. Era uma característica confiável do ambiente ancestral dos humanos que as calorias fossem escassas. Portanto, os genes cujos organismos adoravam calorias tornaram-se mais comuns. Assim como a água fluindo ladeira abaixo.

Somos simplesmente a história incorporada de quais organismos realmente sobreviveram e se reproduziram, não quais organismos deveriam ter prudentemente sobrevivido e se reproduzido.

A retina humana é construída ao contrário: as células sensíveis à luz estão na parte de trás e os nervos emergem da frente e voltam pela retina até o cérebro. Daí o ponto cego. Para um engenheiro humano, isso parece simplesmente estúpido — outros organismos desenvolveram retinas corretas independentemente. Por que não reprojetar a retina?

O problema é que nenhuma mutação única redirecionaria toda a retina simultaneamente. Um engenheiro humano pode reprojetar várias partes simultaneamente ou planejar mudanças futuras. No entanto, se uma única mutação comprometer alguma parte vital do organismo, não importa quantas coisas maravilhosas uma Fada possa construir sobre ela — o organismo morrerá e a frequência do gene diminuirá.

Se você mexer nas células da retina de alguém sem reprogramar também os nervos e o cabo óptico, o sistema na totalidade não funcionará. Não importa que, para uma Fada ou um engenheiro humano, isso represente um passo à frente no re-projeto da retina. O organismo ficará cego. A evolução não tem previsão, é simplesmente a história congelada da qual os organismos de fato se reproduziram. A evolução é tão cega quanto uma retina parcialmente reprojetada.

Encontrar um relógio no deserto, disse William Paley, e você pode inferir a existência de um relojoeiro. Houve quem negasse isso, achando que a vida "simplesmente acontecia" sem a necessidade de um processo de otimização, com ratos sendo gerados espontaneamente a partir de palha e camisas sujas.

Se avaliarmos quem estava mais correto nessa discussão, os teólogos que defendiam um Deus-Criador ou os ateus intelectualmente insatisfeitos que argumentavam que os camundongos eram gerados espontaneamente — então os teólogos devem ser considerados vencedores: a evolução não é Deus, mas está mais próxima de Deus do que da pura entropia aleatória.

As mutações são aleatórias, mas a seleção não é aleatória. Isso não significa que uma Fada inteligente esteja buscando e selecionando ativamente. Significa haver uma correlação estatística diferente de zero entre um gene e a frequência com que um organismo se reproduz. Ao longo de milhões de anos, essa correlação estatística diferente de zero resulta em um fenômeno poderoso. Não é um deus, mas é algo que se assemelha mais a um deus do que à estática na tela de uma televisão.

De muitas maneiras, a evolução compartilha semelhanças com a teologia. Como afirmou Damien Broderick, "os deuses são ontologicamente distintos das criaturas, ou então não valem o papel em que foram escritos" E de fato, o Modelador da Vida não é ele próprio uma criatura. A evolução é incorpórea, assim como a divindade judaico-cristã. Onipresente na natureza, imanente na queda de cada folha. Vasta como a superfície de um planeta. Tem bilhões de anos. Ela mesma não foi criada, surgindo naturalmente da estrutura da física. Tudo isso não soa como algo que poderia ser dito sobre Deus?

No entanto, o Criador não possui uma mente ou um corpo. De certa forma, seu trabalho manual é incrivelmente pobre quando comparado aos padrões humanos. É internamente dividido e, acima de tudo, não é benevolente.

Em certo sentido, Darwin descobriu Deus — um Deus que não correspondia aos preconceitos teológicos e, portanto, passou despercebido. Se Darwin tivesse descoberto que a vida foi criada por um agente inteligente — uma mente incorpórea que nos ama e nos puniria com um raio se disséssemos o oposto — as pessoas teriam exclamado: "Meu Deus! Isso é Deus!"

Mas, em vez disso, Darwin descobriu um estranho Deus alienígena — não confortavelmente "inefável", mas real e genuinamente diferente de nós. A evolução não é um Deus, mas se o fosse, não seria Jeová. Seria Azathoth, de H. P. Lovecraft, o Deus cego e estúpido que borbulha caoticamente no centro de tudo, cercado pelo som suave e monótono das flautas.

Imagine o que poderíamos ter previsto se tivéssemos realmente observado a Natureza.

Isso coloca em dúvida a alegação de alguns religiosos de que acreditam em uma divindade vaga com uma alta probabilidade correspondente. Qualquer um que realmente acreditasse em uma divindade vaga teria reconhecido seu estranho criador desumano quando Darwin exclamou “Aha!”

Da mesma forma, a afirmação de alguns religiosos de que estão inocentemente curiosos aguardando a descoberta de Deus pela ciência também não se sustenta. A ciência já descobriu o tipo de criador dos seres humanos — mas isso não era o que os religiosos queriam ouvir. Eles estavam esperando pela descoberta de seu Deus específico, o Deus altamente específico que eles desejam que exista. No entanto, devem esperar indefinidamente, pois a grande descoberta já ocorreu, e o vencedor é Azathoth.

Bem, mais poder para nós, humanos. Gosto de ter um Criador que posso enganar. É melhor do que ser um animal de estimação. Fico feliz que tenha sido Azathoth e não Odin.

Referências

[1] Francis Darwin, ed., The Life and Letters of Charles Darwin, vol. 2 (John Murray, 1887).

[2] George C. Williams, Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought, Princeton Science Library (Princeton, NJ: Princeton University Press, 1966).

132 — A maravilha da evolução

A maravilha da evolução é que ela funciona.

Digo isso literalmente: se você deseja se maravilhar com a evolução, é por isso que vale a pena.

Como a otimização surge pela primeira vez no universo? Se um agente inteligente projetou a Natureza, quem projetou o agente inteligente? Onde está o primeiro projeto que não tem um projetista? O enigma não é como o estágio inicial do processo pode ser superinteligente e supereficiente; o enigma é como isso pode ocorrer em primeiro lugar.

A evolução resolve a regressão infinita não sendo superinteligente e super eficiente, mas sendo estúpida, ineficiente e funcionando de qualquer maneira. É isso que é maravilhoso.

Por motivos profissionais, muitas vezes sou obrigado a discutir a lentidão, a aleatoriedade e a falta de planejamento da evolução. Então alguém diz: "Você acabou de dizer que a evolução não pode planejar mudanças simultâneas e que ela é muito ineficiente porque as mutações são aleatórias. Não é isso que os criacionistas dizem? Que você não poderia montar um relógio sacudindo aleatoriamente as peças em uma caixa?"

Porém, a resposta aos criacionistas não é que você pode montar um relógio sacudindo as peças em uma caixa. A resposta é que a evolução não ocorre dessa forma. Se você acredita que a evolução ocorre como redemoinhos montando aviões 747, então os criacionistas conseguiram distorcer com sucesso os princípios da biologia para você; eles criaram um espantalho.

A verdadeira resposta é que a complexidade do maquinário evolui incrementalmente, adaptando o maquinário complexo anterior para um novo propósito. Os esquilos pulam de uma copa de árvore para outra usando apenas seus músculos, mas a distância que podem percorrer depende, em certa medida, da aerodinâmica de seus corpos. Então, agora temos esquilos voadores, tão aerodinâmicos que podem planar por curtas distâncias. Se as aves fossem extintas, os descendentes dos esquilos voadores poderiam ocupar aquele nicho ecológico novamente em dez milhões de anos, desenvolvendo membranas transformadas em asas. E os criacionistas diriam: "Para que serve meia asa? Você simplesmente cairia e se espatifaria. Como os esquilos pássaros poderiam ter evoluído incrementalmente?"

É assim que funciona: uma adaptação complexa pode dar origem a uma nova adaptação complexa. A complexidade também pode aumentar incrementalmente a partir de uma única mutação.

Primeiro surge o gene A, que é simples, mas pelo menos um pouco útil por si só, aumentando sua prevalência no pool genético. Agora surge o gene B, que só é útil na presença de A, mas, A está consistentemente presente no *pool* genético, de modo que há uma pressão de seleção a favor de B. Posteriormente, surge uma versão modificada de A, chamada A*, que depende de B para sua função, mas não elimina a dependência de B por A. Portanto, a presença de A* no pool genético

não afeta a pressão de seleção a favor de B. Em seguida, surge o gene C, que depende de A* e B, e o gene B*, que depende de A* e C. Logo você terá um maquinário "irredutivelmente complexo", que quebra se você tirar qualquer peça.

E mesmo assim, você pode visualizar o caminho de volta até aquela única peça: você pode, sem quebrar toda a máquina, tornar uma peça menos dependente da outra, e fazer isso algumas vezes, até que seja possível remover uma peça inteira sem quebrar a máquina, e assim por diante, até transformar um relógio de pulso em um relógio solar simples.

Aqui está um exemplo: o DNA armazena informações eficientemente, em um formato durável que permite a duplicação exata. Um ribossomo converte essa informação armazenada em uma sequência de aminoácidos, ou seja, uma proteína, que se dobra em várias formas quimicamente ativas. O sistema combinado, DNA e ribossomo[6] , consegue construir diversos tipos de maquinaria proteica. Mas, para que serve o DNA sem um ribossomo que traduza as informações do DNA em proteínas? E de que adianta um ribossomo sem o DNA para indicar quais proteínas produzir?

Organismos nem sempre deixam fósseis, e a biologia evolutiva nem sempre consegue descobrir o caminho incremental. No entanto, neste caso, sabemos como isso ocorreu. O RNA compartilha com o DNA a propriedade de poder transportar informações e se autorreplicar, embora seja menos durável e apresente menor precisão na replicação. Além disso, o RNA compartilha a capacidade das proteínas de se dobrarem em formas quimicamente ativas, embora não seja tão versátil quanto as cadeias de aminoácidos das proteínas. É quase certo que o RNA é o gene A original que precede os genes A∗ e B, mutuamente dependentes.

É igualmente importante observar que o RNA não realiza tão bem o trabalho combinado do DNA e das proteínas, quanto cada um individualmente. É incrível o suficiente que uma única molécula possa armazenar informações e manipular a química. Realizar esse trabalho com perfeição seria um milagre totalmente desnecessário.

Qual foi o primeiro replicador a existir? Pode muito bem ter sido uma cadeia de RNA, pois, por uma estranha coincidência, os compostos químicos necessários para a formação do RNA são produtos químicos que provavelmente estavam presentes na Terra pré-biótica, há cerca de 4 bilhões de anos. Observe: a evolução não explica a origem da vida; a biologia evolutiva não precisa explicar o primeiro replicador, porque o primeiro replicador não surgiu de outro replicador. A evolução descreve as tendências estatísticas da replicação. O primeiro replicador não foi uma tendência estatística, foi um puro acidente. A noção de que a evolução deve explicar a origem da vida é um mero espantalho - mais uma distorção criacionista.

Se você estivesse observando a sopa primordial no dia em que ocorreu o primeiro replicador, o dia que transformou a Terra, não ficaria impressionado com a qualidade da replicação desse primeiro replicador. Provavelmente, o primeiro replicador se copiou como um macaco bêbado sob efeito de LSD. Não teria exibido nenhum dos sinais de ajuste fino cuidadosamente incorporados nos replicadores modernos, porque o primeiro replicador foi um acidente. Não era necessário que aquele único filamento de RNA, ou hiperciclo químico, ou padrão de argila se replicasse perfeitamente. Simplesmente precisava acontecer. Ainda assim, provavelmente era muito improvável se considerarmos como um evento isolado - mas só precisava ocorrer uma vez, e havia muitas poças de maré. Alguns bilhões de anos depois, os replicadores estão caminhando na Lua.

O primeiro replicador acidental foi, de fato, a molécula mais importante da história. No en-

6 NT. O ribossomo é uma estrutura celular responsável pela síntese de proteínas, traduzindo o código genético do RNA mensageiro (RNAm) em cadeias polipeptídicas. Presente em todos os seres vivos, ele pode estar livre no citoplasma ou associado ao retículo endoplasmático em células eucariotas.

tanto, se exaltarmos demais suas capacidades e atribuirmos a ele todos os tipos de recursos maravilhosos para auxiliar na replicação, estaremos perdendo o ponto essencial.

Na batalha política entre evolucionistas e criacionistas, não se deve assumir que elogiar a evolução significa estar do lado da ciência. A ciência tem uma compreensão precisa das capacidades da evolução. Se elogiarmos a evolução, além disso, estaremos sendo cientificamente imprecisos, ponto final. Cairíamos na armadilha criacionista ao insistir que um redemoinho consegue montar um Boeing 747. Não é incrível? Como a evolução é maravilhosamente inteligente, como é louvável! Olhe para mim, estou jurando minha lealdade à ciência! Quanto mais coisas boas eu disser sobre a evolução, mais eu devo estar do lado da evolução contra os criacionistas!No entanto, exaltar demasiadamente a evolução mina a verdadeira maravilha, que não está na habilidade da evolução em projetar coisas com maestria, mas sim na capacidade de um processo que ocorre naturalmente de projetar qualquer coisa.

Portanto, devemos abandonar a ideia de que a evolução é um projetista maravilhoso ou um condutor dos destinos das espécies, que nós, como seres humanos, devemos imitar. Para a inteligência humana, imitar a evolução como projetista seria comparável a uma bactéria moderna sofisticada tentando imitar o primeiro replicador como bioquímico. Como T. H. Huxley, conhecido como "O Buldogue de Darwin", afirmou: [1]

> Compreendamos de uma vez por todas que o progresso ético da sociedade não depende de imitar o processo cósmico, muito menos de fugir dele, mas sim de combatê-lo. [7]

Huxley não disse isso porque não acreditava na evolução, mas sim porque a compreendia muito bem.

Referências

[1] Thomas Henry Huxley, Evolution and Ethics and Other Essays (Macmillan, 1894).

7 NT. Texto original em inglês: *Let us understand, once and for all, that the ethical progress of society depends, not on imitating the cosmic process, still less in running away from it, but in combating it.*

133 — Evoluções são estúpidas (mas ainda assim, elas funcionam)

No ensaio anterior, mencionei:

> A ciência tem uma compreensão precisa das capacidades da evolução. Se elogiamos a evolução além desse ponto, não estamos "defendendo a evolução" contra o criacionismo. Estamos sendo cientificamente imprecisos, ponto final.

Neste ensaio, abordo algumas ineficiências e limitações amplamente conhecidas das evoluções. Ao falar sobre "evoluções", no plural, refiro-me ao fato de que a evolução das raposas ocorre de forma oposta à evolução dos coelhos, e nenhum dos dois pode aprender com a evolução das cobras como desenvolver presas venenosas.

Portanto, estou discutindo as limitações da evolução aqui, mas isso não significa que estou tentando introduzir o criacionismo sorrateiramente. Isso é Biologia Evolutiva 201 (se você quiser deduzir as equações). As evoluções, com todas as suas limitações, ainda conseguem explicar a biologia observada; na verdade, essas limitações são essenciais para compreendê-la. Lembre-se de que a maravilha das evoluções não reside em quão bem elas funcionam, mas no fato de que elas simplesmente funcionam.

A inteligência humana é tão complicada que ninguém tem uma maneira precisa de calcular sua eficiência. A seleção natural, embora não seja simples, é mais simples do que um cérebro humano; portanto, é mais lenta e menos eficiente, como convém ao primeiro processo de otimização que já existiu. Na verdade, as evoluções são suficientemente simples para podermos calcular exatamente o quão ineficientes elas são.

As evoluções são lentas. Quão lentas? Suponhamos que ocorra uma mutação benéfica que confira uma vantagem de 3% em aptidão: em média, os portadores desse gene terão 1,03 vezes mais filhos do que os não portadores. Supondo que essa mutação se espalhe, quanto tempo levará para se tornar predominante em toda a população? Isso depende do tamanho da população. Um gene que confere uma vantagem de 3% em aptidão, espalhando-se por uma população de 100.000 indivíduos, levaria, em média, 768 gerações para se tornar universal no pool genético. Para uma população de 500.000, seriam necessárias 875 gerações. A fórmula geral é:

$$Gerações\ para\ Fixação\ =\ 2\ \cdot \frac{ln(N)}{s}$$

onde N é o tamanho da população e (1 + s) é a aptidão. (Se cada portador do gene tem 1,03 vezes mais filhos do que um não portador, s = 0,03.)

Assim, se o tamanho da população fosse de 1.000.000 — estimado durante a era dos caçadores-coletores — seriam necessárias 2.763 gerações para que um gene transmitindo uma vantagem de 1% se propagasse pelo pool genético.[1]

Isso não deveria ser surpreendente; os genes têm que fazer todo o trabalho de propagação por si pró-

prios. Não há uma Fada da Evolução que observe o pool genético e diga: “Mm, este gene parece estar se espalhando rapidamente — eu deveria distribuí-lo para todos“. Em uma economia de mercado humana, alguém que obtém legitimamente um retorno de investimento de 20% — especialmente se houver um mecanismo óbvio e claro por trás disso — pode rapidamente obter mais capital de outros investidores, e outros iniciarão empresas similares. Os genes precisam se espalhar sem mercados de ações, bancos ou imitadores — é como se Henry Ford tivesse que fabricar um carro, vendê-lo, comprar as peças para mais 1,01 carros (em média), vender esses carros e continuar fazendo isso até chegar a um milhão de carros.

Tudo isso pressupõe que o gene se espalhe inicialmente. Nesse caso, a equação é mais simples e não depende do tamanho da população:

$$Probabilidade\ de\ Fixação\ =\ 2\cdot s$$

Uma mutação que confere uma vantagem de 3% (o que é bastante significativo em relação às mutações) tem 6% de chance de se espalhar, pelo menos nessa ocasião. [2] Mutações podem ocorrer mais de uma vez, mas em uma população de um milhão, com uma fidelidade de cópia de 10^{-8} erros por base por geração, pode ser necessário esperar cem gerações por outra oportunidade, e mesmo assim, ainda temos apenas 6% de chance de fixação.

No entanto, a longo prazo, uma evolução tem uma boa chance de eventualmente ser bem-sucedida. (Esse será um tema recorrente.)

Adaptações complexas levam muito tempo para evoluir. Primeiro, surge o alelo A, vantajoso por si só e leva mil gerações para se fixar no *poo*l genético. Somente então outro alelo B, dependente de A, pode começar a se estabelecer. Um casaco de pele não oferece grande vantagem, a menos que o ambiente tenha uma tendência estatisticamente confiável de ser frio. Bem, os genes fazem parte do ambiente de outros genes, e se B depende de A, então B não terá uma grande vantagem a menos que A esteja consistentemente presente no ambiente genético.

Digamos que B confira uma vantagem de 5% na presença de A e nenhuma vantagem caso contrário. Enquanto A ainda esteja presente em 1% da população, B só confere sua vantagem uma vez a cada 100 vezes, resultando em uma média de vantagem de aptidão de B de 0,05% e uma probabilidade de fixação de B de 0,1%. Com uma adaptação complexa, primeiro A$*$ precisa evoluir ao longo de mil gerações, depois B precisa evoluir ao longo de mais mil gerações, e assim por diante. Vários milhões de anos depois, uma nova adaptação complexa surge.

Portanto, outras evoluções não o imitam. Se a evolução da cobra desenvolve um novo veneno incrível, isso não ajuda na evolução da raposa ou do leão.

Compare tudo isso com um programador humano, que pode projetar um novo mecanismo complexo com cem partes interdependentes em uma única tarde. Como isso é possível? Não tenho todas as respostas, e minha suposição é que a ciência também não tem; os cérebros humanos são muito mais complexos do que as evoluções. Eu poderia mencionar conceitos como “encadeamento reverso direcionado a objetivos usando representações modulares combinatórias“, mas isso não ajudaria você a projetar seu próprio ser humano. Ainda assim, os humanos podem projetar novas peças prevendo o projeto posterior de outras novas peças; realizar mudanças simultâneas coordenadas em um maquinário interdependente; aprender observando casos de teste únicos; focar nos pontos problemáticos e pensar abstratamente em como resolvê-los; e priorizar quais ajustes valem a pena tentar, em vez de esperar que um acaso cósmico produza um bom resultado. Em comparação com os padrões da seleção natural, isso é simplesmente mágico.

Os humanos podem fazer coisas que as evoluções provavelmente não conseguiriam realizar durante a expectativa de vida do universo. Como disse certa vez a renomada bióloga Cynthia Kenyon em um jantar ao qual tive a honra de comparecer: “Um estudante de pós-graduação pode realizar em uma hora o que a evolução não conseguiria fazer em um bilhão de anos“. Conforme o melhor conhecimento atual dos biólogos, as evoluções inventaram uma roda completamente funcional em apenas três ocasiões.

E não devemos esquecer a parte na qual o programador publica o trecho de código na Internet.

É verdade que algumas obras evolutivas são impressionantes, mesmo em comparação com a melhor tecnologia desenvolvida pelos *Homo sapiens*. No entanto, nossa explosão cambriana apenas começou, só começamos a acumular conhecimento por volta de… o que, quatrocentos anos atrás? Em certos aspectos, a biologia ainda supera a melhor tecnologia humana. Ainda não conseguimos construir um sistema autorreplicante do tamanho de uma borboleta. Por outro lado, a tecnologia humana avança rapidamente, deixando a biologia para trás. Temos rodas, aço, armas, facas, foguetes, transistores, usinas nucleares. A cada década, essa diferença se amplia ainda mais.

Portanto, mais uma vez, é importante destacar que para um ser humano, olhar para a seleção natural como inspiração para o projeto é como uma bactéria moderna sofisticada tentando imitar a bioquímica do primeiro replicador desajeitado. O primeiro replicador seria instantaneamente devorado instantaneamente na ecologia competitiva atual. O mesmo destino recairia sobre qualquer planejador humano que tentasse introduzir mutações aleatórias em suas estratégias e esperasse 768 iterações de teste para adotar uma melhoria de apenas 3%.

Não devemos elogiar as evoluções além do que elas merecem.

A seguir: mais limites matemáticos emocionantes da evolução!

Referências

[1] Dan Graur and Wen-Hsiung Li, Fundamentals of Molecular Evolution, 2nd ed. (Sunderland, MA: Sinauer Associates, 2000).

[2] John B. S. Haldane, “A Mathematical Theory of Natural and Artificial Selection,“ Mathematical Proceedings of the Cambridge Philosophical Society 23 (5 1927): 607–615, doi:10.1017/S0305004100011750.

134 — Sem evoluções para corporações ou nanodispositivos

> As leis da física e as regras da matemática continuam a se aplicar. Isso me leva a acreditar que a evolução não cessa. Isso também me leva a crer que a natureza — com suas presas e garras ensanguentadas, como alguns a chamaram — será levada a um novo patamar…
>
> "Se livrar da evolução darwiniana" é como tentar se livrar da gravidade. Enquanto houver recursos limitados e múltiplos atores competindo, capazes de transmitir características, haverá pressão seletiva.[8]

— Perry Metzger, prevendo que o reinado da seleção natural continuaria indefinidamente no futuro indefinido

Na biologia evolutiva, assim como em muitos outros campos, é crucial adotar uma abordagem quantitativa, em vez de qualitativa. Uma mutação benéfica "se espalha às vezes, mas nem sempre"? Bem, poderes psíquicos seriam uma mutação benéfica, então poderíamos esperar que se espalhassem, certo? No entanto, esse é um raciocínio qualitativo, não quantitativo — se X é verdadeiro, então Y também é; se poderes psíquicos são benéficos, eles podem se espalhar. Em "Evoluções são Estúpidas", descrevi as equações que calculam a probabilidade de fixação de uma mutação benéfica, aproximadamente o dobro da vantagem adaptativa (6% para uma vantagem de 3%). Somente esse tipo de pensamento numérico nos permite compreender que mutações raramente úteis são extremamente improváveis de se espalharem, sendo praticamente impossível que adaptações complexas surjam sem um uso constante. Se poderes psíquicos realmente existissem, esperaríamos vê-los sendo utilizados o tempo todo — não apenas por serem incrivelmente úteis, mas também porque, caso contrário, eles não poderiam ter evoluído em primeiro lugar.

"Enquanto houver recursos limitados e vários atores competindo, capazes de transmitir características, haverá pressão seletiva." Esse é um raciocínio qualitativo. Mas qual é a magnitude dessa pressão seletiva?

Embora existam vários candidatos para a equação mais importante da biologia evolutiva, eu escolheria a Equação de Price, que, em sua formulação mais simples, diz o seguinte:

$$\Delta z = cov(v_i, z_i)$$

(Mudança na característica média é a covariância entre a aptidão relativa e a característica)

Essa é uma fórmula poderosa e abrangente. Por exemplo, um determinado gene relacionado à altura pode ser representado por Z, a característica que se altera, nesse caso, a Equação de Price afirma que a mudança na probabilidade de possuir esse gene é igual à covariância entre o gene e a aptidão reprodutiva. Ou podemos considerar a altura em geral como a característica Z, independentemente de quaisquer genes

8 NT. Texto original em inglês: *The laws of physics and the rules of math don't cease to apply. That leads me to believe that evolution doesn't stop. That further leads me to believe that nature—bloody in tooth and claw, as some have termed it—will simply be taken to the next level . . . [Getting rid of Darwinian evolution is] like trying to get rid of gravitation. So long as there are limited resources and multiple competing actors capable of passing on characteristics, you have selection pressure.*

específicos, e a Equação de Price nos diz que a mudança na altura na próxima geração será determinada pela covariância entre a altura e a aptidão reprodutiva relativa.

(No entanto, isso é válido apenas se a altura for hereditária direta. Se a nutrição melhorar, levando a um aumento na altura devido a um genótipo fixo, será necessário adicionar um termo de correção à Equação de Price. Se houver interações complexas não lineares entre múltiplos genes, será necessário adicionar um termo de correção ou calcular a equação de maneira tão complexa que não deixa de ser esclarecedora.)

Podemos obter muitos esclarecimentos ao estudar as diferentes formas e derivações da Equação de Price. Por exemplo, a equação final afirma que a mudança na característica média está relacionada à sua covariância com a aptidão relativa, e não com a aptidão absoluta. Isso significa que, se um gene de Frodo salvasse toda a espécie da extinção, a característica média de Frodo não aumentaria, já que a ação de Frodo beneficiou igualmente todos os genótipos e não co-variou com a aptidão relativa."

Diz-se que Price ficou tão perturbado com as implicações de sua equação para o altruísmo que cometeu suicídio, embora possa ter enfrentado outros problemas pessoais. (*Overcoming Bias* não endossa o suicídio após o estudo da Equação de Price.)

Uma das revelações que podem surgir ao refletir sobre a Equação de Price é que "recursos limitados" e "múltiplos atores concorrentes capazes de transmitir características" não são suficientes para impulsionar a evolução. "Coisas que se replicam" não são uma condição suficiente. Mesmo a "competição entre coisas replicantes" não é suficiente.

As corporações evoluem? Certamente, elas competem. Às vezes, elas têm descendentes. Seus recursos são limitados. Às vezes, elas falham.

Mas até que ponto a descendência de uma corporação se assemelha aos seus pais? Grande parte da identidade de uma corporação deriva dos executivos-chave, e os CEOs não podem se dividir por fissão. A Equação de Price só se aplica enquanto as características são herdadas por meio das gerações. Se os tataranetos não se parecem muito com as tataravós, não haverá mais do que quatro gerações de pressão de seleção cumulativa — qualquer coisa que tenha ocorrido há mais de quatro gerações desaparecerá. Sim, a personalidade de uma corporação pode influenciar suas ramificações - mas isso não se compara à hereditariedade do DNA, que é digital em vez de analógica e pode se auto-transmitir com 10^{-8} erros por base por geração.

Com o DNA, temos hereditariedade que se estende por milhões de gerações. É assim que adaptações complexas podem surgir puramente por meio da evolução — o DNA digital persiste o suficiente para que um gene que confere uma vantagem de 3% se espalhe ao longo de 768 gerações, permitindo assim o surgimento de genes dependentes dele. Mesmo que as corporações se replicassem com fidelidade digital, elas estariam atualmente limitadas a, no máximo, dez gerações no mundo do RNA.

Agora, as corporações são certamente selecionadas, no sentido de que as corporações incompetentes vão à falência. Logicamente, isso torna mais provável observar corporações com características que contribuem para a competência. Da mesma forma, uma estrela que se transforma em uma supernova logo após sua formação, tem menos probabilidade de ser visível quando olhamos para o céu noturno. Mas se um acidente na dinâmica estelar faz com que uma estrela queime por mais tempo do que outra, isso não torna mais provável que as estrelas futuras também queimem por mais tempo — essa característica não será transmitida para outras estrelas. Não devemos esperar que futuros astrofísicos descubram características internas complexas em estrelas projetadas para queimar por mais tempo. Esse tipo de adaptação mecânica requer pressões de seleção cumulativas muito maiores do que uma única seleção.

Pense no princípio discutido em "Arrogância de Einstein" - a grande maioria das evidências necessárias para pensar na Relatividade Geral teve que ser usada para elevar essa equação específica ao nível da atenção pessoal de Einstein; a quantidade de evidências necessária para aumentar sua certeza de considerável para 99,9% foi trivial em comparação. Da mesma forma, as características complexas das corporações, que requerem centenas de bits para serem especificadas, são predominantemente produzidas pela inteligência humana, em vez de um número limitado de gerações de evolução de baixa fidelidade. Em biologia, as mutações são aleatórias, e a evolução fornece milhares de bits de pressão de seleção cumulativa. Nas corporações, os humanos contribuem com "mutações" complexas de mil bits projetadas inteligentemente, e a pressão de seleção adicional de "Foi à falência ou não?" acrescenta apenas alguns bits para explicar o que se observa.

A nanotecnologia molecular avançada - do tipo artificial e não biológico - deve ser capaz de se replicar com precisão digital por milhares de gerações. Será que a Equação de Price poderia encontrar um suporte sólido? A correlação é calculada dividindo-se a covariância pela variância, então se A for altamente preditivo de B, pode haver uma forte "correlação" entre eles, mesmo que A varie de 0 a 9 e B varie apenas de 50,0001 a 50,0009. A Equação de Price se baseia na covariância das características com a reprodução, não na correlação! Se for possível reduzir a variação nas características numa faixa estreita, a covariância diminuirá muito, assim como a mudança cumulativa na característica.

O Foresight Institute sugere, entre outras propostas sensíveis, que as instruções de replicação para qualquer nanodispositivo devam ser criptografadas. Além disso, criptografadas de forma que a inversão de um único bit das instruções codificadas embaralhará totalmente a saída descriptografada. Se todos os nanodispositivos produzidos são cópias moleculares precisas e, além disso, quaisquer erros na linha de montagem não forem hereditários porque os descendentes obtiveram uma cópia digital das instruções criptografadas originais para uso na criação de netos, então seus nanodispositivos não estarão evoluindo muito.

Ainda assim, seria necessário considerar os príons — erros de montagem autorreplicantes além das instruções criptografadas, onde um braço robótico pode falhar em pegar um átomo de carbono usado na montagem de um homólogo de si mesmo, resultando em falha do braço robótico dos descendentes para pegar um átomo de carbono, mesmo com todas as instruções criptografadas permanecendo constantes. Mas qual é a probabilidade de haver uma correlação entre esse tipo de erro transmissível e uma taxa reprodutiva mais alta? Digamos que um nanodispositivo produza uma cópia de si mesmo a cada 1.000 segundos, e a nova cópia for magicamente mais eficiente (não apenas possui um príon, mas também um príon benéfico) e se copie a cada 999,99999 segundos. Ele precisa de um átomo de carbono a menos, veja bem. Essa não é uma variação muito grande na reprodução, portanto, também não é uma covariância muito grande.

E com que frequência esses nanodispositivos precisarão se replicar? A menos que eles tenham mais átomos disponíveis do que os existentes no sistema solar, ou no universo visível, apenas um pequeno número de gerações se passará antes que eles atinjam o limite de recursos. "Recursos limitados" não é condição suficiente para a evolução; é necessário que uma fração substancial da população morra repetidamente para liberar recursos. Na verdade, o conceito de "gerações" não é tanto um número inteiro, mas uma integral sobre a fração da população composta por indivíduos recém-criados.

Para mim, a coisa mais assustadora sobre a gosma cinza ou as armas nanotecnológicas é a possibilidade de que elas possam consumir toda a Terra sem deixar nada interessante para acontecer depois. O diamante é mais estável do que as proteínas mantidas juntas pelas forças de van der Waals, então a gosma só precisaria reconstruir alguns fragmentos de si mesma quando um asteroide colidisse. Mesmo que os príons fossem uma linguagem suficientemente poderosa para sustentar a evolução — lembrando que a evolução já é lenta o suficiente com o DNA digital! —, menos de uma geração poderia transcorrer entre o momento em que a gosma devorasse a Terra e a morte do Sol.

Resumindo, se você tiver todas as seguintes características:

- Entidades que se replicam;
- Variação substancial em suas características;
- Variação substancial em sua reprodução;
- Correlação persistente entre as características e a reprodução;
- Hereditariedade de longo alcance e alta fidelidade nas características;
- Nascimento frequente de uma fração significativa da população reprodutora;
- E se tudo isso permanecer verdadeiro ao longo de muitas iterações…

Então, você terá pressões de seleção cumulativas significativas, suficientes para produzir adaptações complexas por meio do poder da evolução.

135 — Evoluindo para a extinção

É um equívoco muito comum pensar que a evolução opera em benefício de uma espécie específica. Você já ouviu alguém falar sobre dois coelhos que se reproduzem e geram oito coelhos, "contribuindo para a sobrevivência de sua espécie"? Um biólogo evolutivo moderno jamais diria nada assim; eles prefeririam dizer que os coelhos estão se reproduzindo com sucesso.

Este é mais um caso em que é necessário considerar simultaneamente vários conceitos abstratos e mantê-los distintos. A evolução não atua em indivíduos específicos; os indivíduos mantêm os genes com os quais nasceram. A evolução ocorre em uma população reprodutiva, em uma espécie, ao longo do tempo. Existe uma tendência natural de pensar que, se a "Fada da Evolução" está operando em uma espécie, ela deve estar otimizando a favor da espécie na totalidade. No entanto, o que realmente muda são as frequências dos genes, e essas frequências não aumentam ou diminuem dependendo do quanto um gene ajuda a espécie na totalidade. Como veremos mais adiante, é bastante possível que uma espécie evolua até a extinção.

Por que meninos e meninas nascem em números aproximadamente iguais? (Deixando de lado os países que usam tecnologias artificiais de seleção de gênero.) Para entender por que isso é surpreendente, considere que 1 macho pode engravidar 2, 10 ou 100 fêmeas; não parece necessário ter o mesmo número de machos e fêmeas para garantir a sobrevivência da espécie. Isso é ainda mais surpreendente na grande maioria das espécies animais, nas quais o macho contribui muito pouco para criar os filhotes — os humanos são excepcionais, mesmo entre os primatas, em termos do nível de investimento paterno. Proporções de gênero equilibradas são encontradas até mesmo em espécies nas quais o macho engravida a fêmea e depois desaparece.

Considere dois grupos em lados opostos de uma montanha: no grupo A, cada mãe dá à luz 2 machos e 2 fêmeas; no grupo B, cada mãe dá à luz 3 fêmeas e 1 macho. Ambos os grupos terão o mesmo número de filhos, mas o grupo B terá 50% mais netos e 125% mais bisnetos. Você pode pensar que isso seria uma vantagem evolutiva significativa.

Mas vamos considerar o seguinte: quanto mais raros os machos se tornam, mais valiosos eles se tornam em termos de reprodução — não para o grupo, mas para o indivíduo progenitor. Cada criança tem um pai e uma mãe. Portanto, em cada geração, a contribuição genética total de todos os machos é igual à contribuição genética total de todas as fêmeas. Quanto menos machos houver, maior será a contribuição genética individual de cada macho. Se todas as fêmeas ao seu redor priorizam o coletivo e as espécies, gerando dez fêmeas para cada macho, você pode fazer uma matança genética, gerando exclusivamente machos, cada macho gerado por esses indivíduos terá (em média), dez vezes mais netos do que suas primas fêmeas.

Assim, enquanto a seleção de grupo deveria favorecer mais fêmeas, a seleção individual favorece um investimento igual nos descendentes masculinos e femininos. Ao observar as estatísticas de uma maternidade, é fácil perceber que o equilíbrio quantitativo entre as forças de seleção de grupo e as forças de seleção individual está fortemente inclinado a favor da seleção individual na espécie *Homo sapiens*.

(Tecnicamente, isso não é apenas um vislumbre. A seleção individual favorece investimentos parentais iguais em descendentes masculinos e femininos. Se os machos custam metade do preço para nascer e/ou criar, nascerão duas vezes mais machos do que fêmeas no equilíbrio evolutivamente estável. Se o mesmo número de machos e fêmeas nascesse na população em geral, mas os machos forem duas vezes mais baratos para dar à luz, então você poderia novamente fazer uma matança genética ao dar à luz mais machos. Por-

tanto, a proporção de nascimentos deveria refletir o equilíbrio dos custos parentais em uma sociedade de caçadores-coletores, entre criar meninos e criar meninas; e isso teria que ser avaliado de alguma forma. Mas, sabe, não parece ser muito mais caro para uma família de caçadores-coletores criar uma menina, então é um tanto suspeito que nasçam o mesmo número de meninos e meninas.)

A seleção natural não se refere a grupos, espécies ou mesmo indivíduos. Em uma espécie sexuada, um organismo individual não evolui; ele mantém os genes com os quais nasceu. Um indivíduo é uma combinação única de genes que nunca se repetirá; como você pode selecionar isso? Quando você considera que quase todos os seus ancestrais estão mortos, fica claro que a ideia de "sobrevivência do mais apto" é um tremendo equívoco. "Replicação do bem-ajustado" seria uma descrição mais precisa, embora tecnicamente a aptidão física seja definida apenas em termos de replicação.

A seleção natural está realmente relacionada às frequências gênicas. Para obter uma adaptação complexa, uma máquina com múltiplas partes dependentes, cada novo gene que evolui depende da presença confiável de outros genes em seu ambiente genético. Eles devem ter altas frequências. Quanto mais complexa a máquina, maior deve ser a frequência. A assinatura da seleção natural é um gene aumentando de 0,00001% do pool genético para 99% do pool genético. Isso é a informação no sentido teórico; é o que deve acontecer para que grandes adaptações complexas evoluam.

A verdadeira luta na seleção natural não é a competição dos organismos por recursos; isso é efêmero, pois todos os participantes desaparecerão em uma próxima geração. A verdadeira luta é a competição dos alelos pela frequência no pool genético. Essa é a consequência duradoura que cria informações duradouras. Os dois carneiros batendo cabeça e travando os chifres são apenas momentos passageiros.

É perfeitamente possível que um alelo se espalhe e se fixe superando um alelo alternativo que era "melhor para a espécie". Se o Monstro do Espaguete Voador, criasse magicamente uma espécie cuja mistura de gêneros fosse perfeitamente otimizada para garantir a sobrevivência da espécie - a mistura de gêneros ideal para se recuperar de eventos de quase extinção, adaptar-se a novos nichos e assim por diante - então a evolução degradaria rapidamente o ideal dessa espécie de volta ao ponto ideal da seleção individual, com investimento parental igual em machos e fêmeas.

Imagine um "gene Frodo"[9] que sacrifica seu portador para salvar toda a sua espécie de um evento de extinção. O que aconteceria com a frequência do alelo como resultado? Ela diminuiria. Tchau, tchau!

Se as ameaças de extinção ao nível de espécie ocorressem regularmente (chamemos isso de "ambiente Buffy"[10]), o gene Frodo diminuiria sistematicamente em frequência, desapareceria e, consequentemente, a espécie também.

Um exemplo hipotético? Talvez. Se a espécie humana fosse permanecer biológica por mais um século, seria uma boa ideia começar a clonar Gandhi. Nos vírus, existe uma tensão entre os vírus individuais que se replicam o mais rápido possível e o benefício de manter o hospedeiro vivo tempo suficiente para transmitir a doença. Isso é um exemplo real de seleção de grupo e, se o vírus evoluir a um ponto no qual as pressões de seleção de grupo não superem as pressões individuais, o vírus pode desaparecer rapidamente. Não sei se alguma vez uma doença foi observada evoluindo para a extinção, mas provavelmente isso ocorreu inúmeras vezes.

Os distorcedores da segregação subvertem os mecanismos que normalmente garantem a justiça da reprodução sexual. Por exemplo, há um distorcedor da segregação no cromossomo sexual masculino de alguns camundongos, resultando no nascimento apenas de filhotes machos, todos portando o distorcedor da segregação. Então esses machos se reproduzem com as fêmeas, que dão à luz apenas filhotes machos, e

9 NT. **Frodo Bolseiro** é um personagem central da trilogia "O Senhor dos Anéis", de J.R.R. Tolkien, um hobbit encarregado da missão de destruir o Um Anel para salvar a Terra-média das forças do mal. Sua jornada destaca temas como coragem, sacrifício e resistência diante da tentação e do poder.

10 NT. A série **Buffy, a Caçadora de Vampiros** (*Buffy the Vampire Slayer*), criada por Joss Whedon, estreou em 1997 e durou até 2003. A história segue Buffy Summers, uma jovem escolhida para combater forças do mal, como vampiros e demônios, enquanto equilibra sua vida pessoal e os desafios da adolescência e da vida adulta. A série é conhecida por sua mistura única de drama, humor e temas sobrenaturais.

assim por diante. Você pode argumentar “Isso é trapaça!“, mas essa é uma perspectiva humana; a aptidão reprodutiva desse alelo é extremamente alta, ao produzir duas vezes mais cópias de si mesmo na geração seguinte do que sua alternativa não mutante. Mesmo quando as fêmeas se tornam cada vez mais raras, os machos que carregam esse gene não têm menos probabilidade de acasalar do que qualquer outro macho, e assim o distorcedor da segregação continua sendo duas vezes mais vantajoso do que seu alelo alternativo. Especula-se que a seleção de grupo no mundo real possa ter desempenhado um papel em manter a frequência desse gene tão baixa quanto parece. Nesse caso, se os camundongos desenvolvessem a capacidade de voar e migrar no inverno, provavelmente formariam uma única população reprodutiva e evoluiriam para a extinção à medida que o distorcedor da segregação se fixasse.

Aproximadamente 50% do genoma total do milho consiste em transpósons (genes saltadores), elementos de DNA cuja principal função é copiar-se para outros locais do DNA. Uma classe de transpósons chamadas “elementos P“ parece ter surgido pela primeira vez na Drosófila por volta do meio do século XX e se espalhado para todas as populações da espécie em 50 anos. A “sequência Alu“ em humanos, um transpóson de 300 bases, é repetida entre 300.000 e um milhão de vezes no genoma humano. Isso pode não levar à extinção de uma espécie, mas certamente não ajuda; os transpósons causam mais mutações, as quais são predominantemente prejudiciais, diminuindo a fidelidade da cópia efetiva do DNA. No entanto, esses trapaceiros são altamente adaptados.

Suponha que em algumas espécies que se reproduzem sexualmente, um mecanismo de cópia perfeito do DNA seja inventado. Como a maioria das mutações é prejudicial, esse complexo de genes seria uma vantagem para seus portadores. Agora você pode se perguntar sobre as mutações benéficas — elas ocorrem ocasionalmente, então os indivíduos não mutantes não estariam em desvantagem? No entanto, em uma espécie sexual, uma mutação benéfica que surja em um indivíduo mutante pode se espalhar para os descendentes de indivíduos não mutantes também. Os indivíduos mutantes sofrem mutações degeneradas a cada geração, enquanto os indivíduos não mutantes podem adquirir benefícios sexualmente e, assim, se beneficiar de quaisquer mutações benéficas que ocorram nos indivíduos mutantes. Portanto, os indivíduos mutantes possuem uma desvantagem pura. O mecanismo de cópia perfeito do DNA aumenta em frequência até se fixar. Dez mil anos depois, ocorre uma era glacial e a espécie desaparece. Evoluiu para a extinção.

O “efeito espectador“ ocorre quando alguém está enfrentando problemas e é mais provável que indivíduos solitários intervenham do que grupos. Um estudante universitário aparentemente tendo um ataque epiléptico foi ajudado em 85% das vezes por um único espectador e em 31% das vezes por cinco espectadores. Especulo que, mesmo que a relação de parentesco em uma tribo de caçadores-coletores fosse forte o suficiente para criar uma pressão de seleção para ajudar indivíduos não diretamente relacionados, quando vários potenciais ajudantes estivessem presentes, uma corrida armamentista genética poderia ocorrer para ser o último a intervir. Todos hesitam, esperando que alguém o faça. Atualmente, a humanidade enfrenta várias ameaças de extinção ao nível de espécie, e devo dizer que não há muitas pessoas se apresentando. Se perdermos essa batalha porque quase ninguém apareceu no campo de batalha, então — como provavelmente inúmeras espécies que não vemos hoje — teremos evoluído para a extinção.

As células cancerígenas são altamente bem-sucedidas no corpo, prosperando e acumulando mais recursos, superando em muito as suas contrapartes mais obedientes. Por um tempo.

Organismos multicelulares só podem existir porque desenvolveram poderosos mecanismos internos para impedir a evolução. Se as células começarem a evoluir, elas rapidamente evoluirão para a extinção: o organismo morre.

Portanto, não elogie a evolução por sua preocupação com o indivíduo; quase todos os seus ancestrais estão mortos. Não elogie a evolução por sua preocupação com uma espécie; nunca foi encontrada uma adaptação complexa que só possa ser interpretada como operando para preservar uma espécie, e a matemática sugere que isso é virtualmente impossível.

De fato, é perfeitamente possível que uma espécie evolua para a extinção. A humanidade pode estar chegando a esse ponto agora. Você não pode nem elogiar a evolução por sua preocupação com os genes; a batalha entre dois alelos alternativos no mesmo local é um jogo de soma zero em termos de frequência.

A aptidão física nem sempre está a seu favor.

136 — A tragédia do selecionismo de grupo

Antes de 1966, era comum ver biólogos sérios defendendo hipóteses evolutivas que hoje consideraríamos como pensamento mágico. Essas ideias confusas tiveram um papel importante na história do desenvolvimento da teoria evolutiva, o erro levando à correção; assim como a loucura dos reis ingleses levou à criação da Magna Carta e da democracia constitucional.

Como um exemplo de romance, Vero Wynne-Edwards, Warder Allee e J. L. Brereton, entre outros, acreditavam que os predadores controlariam voluntariamente sua reprodução para evitar a superpopulação em seu habitat e o esgotamento da população de presas.

Mas a evolução não permite comportamentos arbitrários. Não podemos explicar o chocalho de uma cascavel dizendo que ele existe para beneficiar outros animais que, de outra forma, seriam mordidos. Não há uma Fada da Evolução externa decidindo quando um gene deve ser favorecido; o efeito do gene deve, de alguma forma, diretamente aumentar sua prevalência na próxima geração. É compreensível por que nosso senso de estética humana, ao testemunhar a queda populacional das raposas que devoraram todos os coelhos, clama "Alguma providência deveria ter sido tomada!" Mas como um complexo de genes que restringe a reprodução - entre todas as coisas! – poderia torna-se mais frequente na próxima geração?

Um ser humano que cria uma pequena ecologia de brinquedo - por diversão, como um modelo de ferrovia - pode ficar frustrado se suas populações cuidadosamente construídas de raposas e coelhos se autodestruírem, com as raposas devorando todos os coelhos e depois morrendo de fome. Assim, o ser humano interfere na ecologia de brinquedo - uma restrição à reprodução das raposas é a solução óbvia que vem à mente humana - até que a ecologia pareça boa e organizada. A natureza não tem seres humanos, é claro, mas isso não deve nos impedir - agora que sabemos o que queremos esteticamente, só precisamos apresentar um argumento plausível que convença a natureza a querer o mesmo do ponto de vista evolutivo.

Obviamente, a seleção no nível do indivíduo não resultará em restrição individual na reprodução. Indivíduos que se reproduzem sem controle, naturalmente, produzirão mais descendentes do que aqueles que se restringem.

(A seleção individual não levará a um sacrifício individual de oportunidades de reprodução. Certamente, a seleção individual pode resultar em indivíduos que, após adquirir todos os recursos disponíveis, usam esses recursos para produzir quatro ovos grandes em vez de oito ovos pequenos - não para conservar recursos sociais, mas porque é o ponto ideal individual em termos de (número de ovos) × (probabilidade de sobrevivência do ovo). Isso não resolve o problema dos comuns.

Mas vamos supor que a população da espécie fosse dividida em subpopulações, que em sua maioria estivessem isoladas e cruzassem ocasionalmente. Nesse caso, certamente as subpopulações que restringissem sua reprodução teriam menos probabilidade de se extinguir e enviariam mais mensageiros, criando novas colônias para repovoar os territórios das populações que diminuíram.

O problema com esse cenário não era que fosse matematicamente impossível. O problema

residia no fato de que era possível, porém extremamente difícil.

O problema fundamental é que não são apenas os criadores controlados que se beneficiam da restrição reprodutiva. Se algumas raposas se abstêm de reproduzir filhotes que se alimentam de coelhos, os coelhos não comidos não são exclusivamente destinados aos filhotes que possuem a adaptação de reprodução restrita. As raposas desenfreadas e seus muitos filhotes alegremente devorarão qualquer coelho que não seja caçado. A única maneira do gene restritivo sobreviver a essa pressão é se os benefícios da restrição forem preferencialmente concedidos aos indivíduos que restringem sua reprodução.

Especificamente, o requisito é $C/B < F_{ST}$, onde C representa o custo do altruísmo para o doador, B é o benefício do altruísmo para o receptor e F_{ST} é a estrutura espacial da população: a média de parentesco entre um organismo selecionado aleatoriamente e seu vizinho selecionado aleatoriamente, sendo um "vizinho" qualquer outra raposa que se beneficie da contenção de uma raposa altruísta. [1]

Então, o custo da restrição reprodutiva é suficientemente baixo e o benefício empírico de evitar a fome é suficientemente alto em comparação com a estrutura espacial empírica das populações de raposas e coelhos para que o argumento da seleção de grupo possa funcionar?

A matemática sugere que isso é bastante improvável. Nesta simulação, por exemplo, o custo para os altruístas é de 3% de sua aptidão, os grupos puramente altruístas têm uma aptidão duas vezes maior que os grupos puramente egoístas, o tamanho da subpopulação é de 25, e 20% de todas as mortes são substituídas por mensageiros de outro grupo: o resultado é um polimorfismo entre egoísmo e altruísmo. Se o tamanho da subpopulação for dobrado para 50, o egoísmo é corrigido; se o custo para os altruístas for aumentado para 6%, o egoísmo é corrigido; se o benefício altruísta for reduzido pela metade, o egoísmo é fixado ou se torna predominante. Os grupos locais devem ser muito pequenos, com cerca de 5 membros, para que a seleção de grupo funcione quando o custo do altruísmo excede 10%. Isso não parece plausível para as raposas que restringem sua reprodução.

Imagino que você possa compreender, creio eu, que os defensores da seleção de grupo acabaram perdendo o argumento científico. O ponto crucial não foi o argumento matemático, mas sim a observação empírica: as raposas não restringiram sua reprodução (esqueci a espécie exata da disputa; não eram raposas e coelhos) e, na realidade, os sistemas predador-presa estão sempre em conflito. Mais tarde, o selecionismo de grupo ressurgiu de forma notavelmente diferente — falando matematicamente, existe uma estrutura de vizinhança que implica em uma pressão de seleção de grupo diferente de zero, embora não necessariamente capaz de superar a pressão de seleção individual compensatória. Se não levarmos isso em conta, nossos cálculos estarão errados, ponto final. Além disso, os mecanismos de aplicação evoluídos (que não foram originalmente postulados) transformam completamente o jogo. Então, por que essa disputa científica, agora histórica, é relevante para o *Overcoming Bias*?

Uma década após a controvérsia, um biólogo teve uma ideia fascinante. As condições matemáticas necessárias para a seleção de grupo superar a seleção individual eram tão extremas que não poderiam ser encontradas naturalmente na natureza. Por que não as criar artificialmente em laboratório? Michael J. Wade fez exatamente isso, selecionando repetidamente populações de insetos com um número reduzido de adultos por subpopulação. [2] E qual foi o resultado? Os insetos restringiram sua reprodução e viveram em harmonia com comida suficiente para todos?

Não; os adultos adaptaram-se para canibalizar ovos e larvas, principalmente as larvas femininas.

É evidente que selecionar por subpopulações pequenas não resultaria em indivíduos que restringissem sua própria reprodução, mas sim em indivíduos que devorariam os filhotes de outros indivíduos, especialmente as fêmeas.

Ao obter esses resultados experimentais — os quais são extremamente óbvios em retrospecto — torna-se claro como os defensores originais da seleção de grupo permitiram que o romantismo, uma noção humana de estética, obscurecesse suas previsões sobre a natureza.

Isso é um exemplo arquetípico de uma Terceira Alternativa perdida, resultado de uma racionalização de um resultado final pré definido que gerou uma justificativa falsa e foi motivadamente encerrada. Os defensores da seleção de grupo não abordaram o conceito de seleção de grupo com mentes claras e imparciais, não extrapolaram de forma neutra o resultado provável. Eles começaram com a bela ideia de populações de raposas que restringiriam voluntariamente sua reprodução conforme a capacidade da população de coelhos, criando uma natureza em perfeita harmonia. Em seguida, procuraram uma razão para isso ocorrer e tiveram a ideia da seleção de grupo. Como eles já sabiam qual resultado desejavam obter com a seleção de grupo, não consideraram adaptações menos belas e esteticamente agradáveis que a seleção de grupo provavelmente promoveria. Se realmente estivessem tentando prever com calma e imparcialidade o resultado da seleção de pequenas subpopulações resistentes à fome, teriam considerado a possibilidade de canibalismo dos filhotes de outros organismos ou algum resultado igualmente "feio" — muito antes de imaginarem algo tão surpreendente em termos evolutivos, como a restrição individual na reprodução!

Isso também ilustra o ponto que eu estava enfatizando em "A arrogância de Einstein": quando se trata de amplas áreas de resposta, a maioria do trabalho real é direcionada para promover uma resposta possível que chame a atenção. Se uma hipótese for promovida de forma inadequada para chamar sua atenção - sua percepção estética sugere uma maneira bonita de como a natureza deveria ser, mas a seleção natural não é guiada por uma "Fada da Evolução" que compartilha sua apreciação pela beleza - isso, por si só, pode selar sua ruína, a menos que você consiga limpar sua mente completamente e começar de novo.m princípio, até a pessoa mais estúpida do mundo pode afirmar que o Sol está brilhando, mas isso não significa que está escuro. Mesmo que uma resposta seja sugerida por um lunático sob efeito de LSD, você deve ser capaz de avaliar imparcialmente as evidências a favor e contra, e se necessário, desconsiderar essa resposta.

Na prática, os selecionistas de grupo estavam condenados desde o início, pois o resultado final que buscavam foi originalmente sugerido pelo seu senso estético, enquanto o resultado final da natureza foi produzido pela seleção natural. Esses dois processos não tinham nenhuma razão fundamental para que suas saídas se correlacionassem e, de fato, não se correlacionavam. Todas as discussões acaloradas que se seguiram não alteraram esse fato.

Se você parte de seus próprios desejos sobre o que a natureza deveria fazer, considera as razões pelas quais a própria natureza realiza suas ações e, em seguida, racionaliza um argumento extremamente persuasivo de por que a natureza deveria produzir o resultado que você prefere, usando as próprias razões da natureza, infelizmente, ela ainda não ouvirá. O universo não tem mente e não está sujeito a persuasão política inteligente. Você pode argumentar o dia todo sobre porque a gravidade deveria fazer a água fluir para cima, mas a água acaba no mesmo lugar, independentemente disso. É como se o universo estivesse indiferente. Como disse J. R. Molloy: "A natureza é a suprema fanática, pois é obstinada e intolerantemente devota aos seus próprios preconceitos e recusa-se categoricamente a ceder às racionalizações mais persuasivas dos seres humanos."

Costumo recomendar a biologia evolutiva a amigos porque o campo moderno tenta treinar seus alunos a não racionalizarem, evitando assim o erro que exige correção. Físicos e engenheiros elétricos não precisam ser cuidadosamente treinados para evitar a antropomorfização dos elétrons, pois os elétrons não exibem comportamentos mentais. A seleção natural cria propósitos que são estranhos aos seres humanos, e os estudantes da teoria evolutiva são alertados sobre isso. É um bom treinamento para qualquer pessoa que pense, mas é especialmente importante se você quiser ter clareza ao refletir sobre outros processos mentais estranhos que não funcionam como os nossos.

Referências

[1] David Sloan Wilson, “A Theory of Group Selection,“ Proceedings of the National Academy of Sciences of the United States of America 72, no. 1 (1975): 143–146.

[2] Michael J. Wade, “Group selections among laboratory populations of Tribolium,“ Proceedings of the National Academy of Sciences of the United States of America 73, no. 12 (1976): 4604–4607, doi:10.1073/pnas.73.12.4604.

137 — Critérios falsos de otimização

Já me estendi bastante sobre as formas de racionalização pelas quais nossas crenças parecem se alinhar com as evidências muito mais fortemente do que realmente o fazem. E não estou exagerando nesse ponto, também. Se pudéssemos superar essa tendência fundamental e compreender o que cada hipótese realmente previu, conseguiríamos corrigir quase qualquer erro factual.

O desafio do "espelho" na teoria da tomada de decisões é identificar qual opção um critério de escolha realmente endossa.Se seus princípios morais declarados defendem fornecer notebooks para todos, isso realmente justifica comprar para si mesmo um notebook de um milhão de dólares cravejado de joias ou gastar o mesmo dinheiro enviando 5.000 notebooks de baixo custo para um projeto como o OLPC?

Parece que desenvolvemos uma habilidade especial para argumentar que praticamente qualquer objetivo pode justificar praticamente qualquer ação. Um teórico do flogisto explicando por que o magnésio ganha peso quando queimado não tem nada a ver com um inquisidor explicando por que o amor infinito de Deus por todos os Seus filhos exige que alguns deles sejam queimados na fogueira.

Não há mistério nisso. A política já fazia parte do nosso ambiente ancestral. Somos descendentes daqueles que argumentaram de forma mais persuasiva que o bem da tribo exigia a execução de seu odiado rival, Uglak. (Com certeza não somos descendentes de Uglak.)

E ainda assim… é possível provar que se Robert Mugabe estivesse realmente preocupado apenas com o bem do Zimbábue, ele renunciaria à presidência? Podemos argumentar que a política deriva dos objetivos, mas não acabamos de ver que os seres humanos podem combinar qualquer objetivo com qualquer política? Como saber que estamos certos e Mugabe está errado? (Há várias razões pelas quais essa é uma suposição válida, mas peço que tenha paciência comigo aqui.)

As motivações humanas são múltiplas e obscuras, nossos processos de tomada de decisões são tão complexos quanto nossos cérebros. E o próprio mundo é extremamente complicado em cada escolha política da vida real. Será que podemos sequer provar que os seres humanos estão racionalizando — distorcendo sistematicamente a conexão entre princípios e política — quando não temos um ponto de apoio firme para nos sustentarmos? Quando não há como descobrir exatamente o que um único critério de otimização implica? (Na verdade, podemos simplesmente observar que as pessoas discordam sobre a política do escritório de maneiras que estranhamente se correlacionam com seus próprios interesses, ao mesmo tempo, em que negam que tais interesses estejam em jogo. Mas, novamente, peço que tenha paciência comigo aqui.)

Onde está o processo de otimização consequencialista padronizado, de código aberto, geralmente inteligente, no qual podemos inserir uma moralidade completa como um arquivo XML, para descobrir o que essa moralidade realmente recomenda quando aplicada ao nosso mundo? Existe pelo menos um único caso no mundo real em que podemos saber exatamente o que um critério de escolha recomenda? Onde está o pensador moral puro - com uma função de utilidade conhecida,

livre de todos os outros desejos errantes que possam distorcer sua otimização - cujo resultado confiável podemos contrastar com as racionalizações humanas da mesma função de utilidade?

Ora, é o nosso velho conhecido, o deus alienígena, é claro! A seleção natural é garantidamente livre de toda misericórdia, todo amor, toda compaixão, todas as sensibilidades estéticas, todo partidarismo político, todas as lealdades ideológicas, todas as ambições acadêmicas, todo libertarianismo, todo socialismo, todo azul e todo verde. A seleção natural não maximiza seu critério de aptidão genética inclusiva - ela não é tão inteligente assim. No entanto, quando observamos o resultado da seleção natural, estamos de fato olhando para um resultado otimizado apenas para a aptidão genética inclusiva, e não para os interesses específicos da indústria agrícola dos Estados Unidos.

Nos estudos de caso da ciência evolutiva - como por exemplo, na Tragédia do Selecionismo de Grupos, podemos comparar diretamente as racionalizações humanas com o resultado da otimização pura de um critério conhecido. O que Wynne-Edwards acreditava que seria o resultado da seleção de grupos para subpopulações pequenas? Restrição individual voluntária na reprodução e comida suficiente para todos. No entanto, qual foi o resultado real dos experimentos de laboratório? Canibalismo.

Agora, você pode questionar: esses casos históricos da ciência evolutiva são realmente relevantes para a moralidade humana, que não se importa muito com a aptidão genética inclusiva quando isso vai contra o amor, a compaixão, a estética, a cura, a liberdade, a justiça, etc.? As sociedades humanas nem sequer possuíam o conceito de “aptidão genética inclusiva“ até o século XX.

Mas, em contrapartida, eu pergunto: se não conseguimos ver claramente o resultado de um único critério de otimização monótona — se não conseguimos nem mesmo treinar nossos ouvidos para ouvir uma única nota pura - como poderemos apreciar uma orquestra? Como reconheceremos que “Seja sempre egoísta“ ou “Sempre obedeça ao governo“ são princípios orientadores ruins para os seres humanos adotarem — se acreditarmos que até mesmo a otimização dos genes para a aptidão inclusiva resultará em organismos que sacrificam oportunidades reprodutivas em prol da conservação de recursos sociais?

Para treinarmos nossa capacidade de enxergar com clareza, precisamos de casos práticos simples.

138 — Executores de adaptação, não maximizadores de aptidão

> "Os organismos individuais são mais bem compreendidos como executores de adaptação do que como maximizadores de aptidão."[11]

— John Tooby e Leda Cosmides,

The Psychological Foundations of Culture (As Bases Psicológicas da Cultura) [1]

Há 50 mil anos, as papilas gustativas do *Homo sapiens* direcionavam seus portadores aos recursos alimentares mais escassos e cruciais: açúcar e gordura. Em uma palavra, calorias. Hoje, o contexto da função das papilas gustativas mudou, mas as próprias papilas permanecem as mesmas. As calorias, longe de serem escassas (pelo menos nos países do Primeiro Mundo), são ativamente prejudiciais. Os micronutrientes que eram abundantes em folhas e nozes estão ausentes no pão, mas nossas papilas gustativas não reclamam. Uma bola de sorvete é um superestímulo, contendo mais açúcar, gordura e sal do que qualquer coisa encontrada em nosso ambiente ancestral.

Nenhum ser humano com o objetivo deliberado de maximizar a aptidão genética inclusiva de seus alelos comeria um biscoito, a menos que estivesse morrendo de fome. No entanto, é melhor pensar nos organismos individuais como executores de adaptação, não como maximizadores de aptidão.

Uma chave de fenda Phillips, embora tenha sido projetada para girar parafusos, não se transformará em uma chave de fenda de cabeça chata para cumprir sua função. Criamos essas ferramentas, mas elas existem independentemente de nós e continuam existindo independentemente de nós.

Os átomos de uma chave de fenda não possuem pequenas *tags* XML descrevendo seu propósito "objetivo". O projetista tinha algo em mente, é verdade, mas isso não é o mesmo que ocorre no mundo real. Se esquecermos que o projetista é uma entidade separada da coisa projetada, poderíamos pensar: "O propósito da chave de fenda é girar parafusos" — como se isso fosse uma propriedade explícita da própria chave de fenda, em vez de uma propriedade do estado mental do projetista. Poderíamos ficar surpresos ao descobrir que a chave de fenda não se reconfigura para se ajustar a um parafuso de cabeça chata, já que, afinal, o propósito da chave de fenda é girar parafusos.

A causa da existência da chave de fenda é a mente do projetista, que imaginou um parafuso imaginário e uma manivela imaginária girando. O funcionamento real da chave de fenda, seu ajuste real a uma cabeça de parafuso real, não pode ser a causa objetiva de sua existência: o futuro não pode causar o passado. No entanto, o cérebro do projetista, como algo que realmente existiu no passado, pode ser a causa da chave de fenda.

As consequências da existência da chave de fenda podem não corresponder às consequências imaginadas na mente do projetista. A lâmina da chave de fenda pode escorregar e cortar a mão do usuário.

E o significado da chave de fenda — bem, isso é algo que existe na mente do usuário, não em pequenas etiquetas nos átomos da chave de fenda. O projetista pode ter a intenção de que ela gire parafusos, mas um assassino pode comprá-la para usá-la como arma. E então, por acidente, a chave de fenda pode ser deixada cair e apanhada por uma criança, que a utiliza como cinzel.

---

11 NT. Texto original em inglês. *Individual organisms are best thought of as adaptation-executers rather than as fitness-maximizers.*

Portanto, a causa da chave de fenda, sua forma, suas consequências e seus diversos significados são coisas distintas; e apenas uma dessas coisas é encontrada na própria chave de fenda.

De onde vêm as papilas gustativas? Elas não são obra de um projetista inteligente que previu suas consequências, mas de uma história de ancestralidade congelada: Adão gostou de açúcar, comeu uma maçã e se reproduziu, Bárbara, gostou do açúcar e comeu uma maçã ese reproduziu. Carlos gostou do açúcar e comeu uma maçã e se reproduziu, e 2763 gerações mais tarde, o alelo se tornou fixo na população. Por simplicidade, muitas vezes resumimos toda essa história dizendo: "A evolução fez isso". Mas a evolução não é um evento rápido ou localizado, como um projetista humano criando uma ferramenta. Essa é a causa objetiva das nossas papilas gustativas.

Qual é a forma objetiva de uma papila gustativa? Em termos técnicos, é um sensor molecular conectado a um sistema de recompensa no cérebro. Isso adiciona outro tipo de indireção, pois a papila gustativa não busca comida diretamente. Ela influencia a mente do organismo, fazendo com que o organismo queira comer alimentos que são similares aos que acabamos de experimentar.

Qual é a consequência objetiva de uma papila gustativa? Para um ser humano moderno em um país desenvolvido, o resultado pode ser uma série de eventos em cadeia: do desejo de comer mais chocolate, ao plano de comer mais chocolate, ao consumo de chocolate, ao engordar, ao ter menos encontros, à reprodução com menos sucesso.

Essa consequência é diretamente oposta à principal regularidade na longa cadeia de sucessos ancestrais que causou o formato do paladar. Mas como comer demais só recentemente se tornou um problema , nenhuma evolução significativa (regularidade comprimida de ancestralidade) influenciou ainda mais o formato das papilas gustativas..

Qual é o significado de comer chocolate? Isso fica entre você e sua filosofia moral. Pessoalmente, acho que chocolate tem um sabor delicioso, mas gostaria que fosse menos prejudicial; soluções aceitáveis incluiriam reprojetar o chocolate ou reprojetar a minha bioquímica.

Agrupando vários desses conceitos, poderíamos dizer, de certa forma: "Os seres humanos modernos fazem hoje o que teria propagado nossos genes em uma sociedade de caçadores-coletores, independentemente de ajudar ou não nossos genes em uma sociedade moderna". Mas isso ainda não está totalmente correto, pois não estamos realmente nos perguntando quais comportamentos maximizariam a aptidão inclusiva de nossos ancestrais. Além disso, muitas de nossas atividades atuais não possuem um equivalente ancestral. Na sociedade de caçadores-coletores, não existia chocolate.

Portanto, é mais adequado considerar nossas papilas gustativas como uma adaptação ajustada às condições ancestrais, que incluíam quase morrer de fome, maçãs e coelhos assados, e que os seres humanos modernos executam em um novo contexto que envolve chocolate barato e constante bombardeio de propagandas.

Assim, afirma-se: os organismos individuais são melhores compreendidos como executores de adaptação, não como maximizadores de aptidão.

Referências

[1] John Tooby and Leda Cosmides, "The Psychological Foundations of Culture," in The Adapted Mind: Evolutionary Psychology and the Generation of Culture, ed. Jerome H. Barkow, Leda Cosmides, and John Tooby (New York: Oxford University Press, 1992), 19–136.

139 — Psicologia Evolutiva

Assim como “chat IRC[12] ” ou “protocolo TCP/IP[13]”, a expressão “órgão reprodutivo“ é redundante. Todos os órgãos são órgãos reprodutivos. De onde surgem as asas de um pássaro? Seria obra de uma Fada da Evolução dos Pássaros, que achava o voo muito legal? As asas do pássaro estão presentes porque contribuíram para a reprodução dos ancestrais do pássaro, assim como o coração, os pulmões e os órgãos genitais da ave. No máximo, podemos distinguir entre órgãos reprodutivos diretos e indiretos.

Essa observação também se aplica ao cérebro, o sistema orgânico mais complexo conhecido pela biologia. Alguns órgãos cerebrais são diretamente reprodutivos, como o desejo sexual, enquanto outros são indiretamente reprodutivos, como a raiva.

De onde vem a emoção humana da raiva? Uma Fada da Evolução dos Humanos que pensou que a raiva era uma característica interessante? O circuito neural da raiva é tão certamente um órgão reprodutivo quanto o fígado. A raiva existe no *Homo sapiens* porque os ancestrais raivosos tiveram mais filhos. Não há outra maneira de ela ter chegado a nós.

Esse fato histórico sobre a origem da raiva confunde muitas pessoas. Elas questionam: “Espere, você está dizendo que, quando estou com raiva, estou subconscientemente tentando ter filhos? Não é isso que penso quando alguém me dá um soco no nariz“.

Não, não, não, NÃO!

Organismos individuais são mais bem compreendidos como executores de adaptação, não como maximizadores de aptidão. A causa de uma adaptação, a forma de uma adaptação e a consequência de uma adaptação são coisas distintas. Se você construísse uma torradeira, não esperaria que ela mudasse de forma ao tentar inserir uma fatia de pão inteira. Sim, você pretendia usá-la para fazer torradas, mas essa intenção é um fato sobre você, não sobre a torradeira. A torradeira não tem consciência de seu próprio propósito.

No entanto, uma torradeira não é um objeto intencional. Ela não possui mente, e, portanto, não somos tentados a atribuir objetivos a ela. Quando vemos uma torradeira em relação ao seu propósito, não supomos que ela tenha consciência disso, pois não acreditamos que a torradeira saiba de nada.

Isso é semelhante ao antigo teste de ser solicitado a dizer a cor das letras em “azul“. As pessoas levam mais tempo para nomear a cor, devido à necessidade de processar o significado das letras e a cor em que estão escritas. Não teríamos problemas semelhantes em nomear a cor das letras em “vento“.

Mas um cérebro humano, além de ser um artefato produzido historicamente pela evolução, também é uma mente capaz de ter suas próprias intenções, propósitos, desejos, metas e planos. Tanto uma abelha

12 NT. O **chat IRC** (Internet Relay Chat) é um protocolo de comunicação em tempo real criado em 1988, que permite a interação em salas de bate-papo (canais) ou conversas privadas entre usuários. Amplamente utilizado nas décadas de 1990 e 2000, o IRC foi uma das primeiras formas de comunicação online em grupo, sendo popular para discussões técnicas, comunidades de interesse e suporte em tempo real.

13 NT. O **protocolo TCP/IP** (Transmission Control Protocol/Internet Protocol) é o conjunto de regras e padrões que permite a comunicação e a transferência de dados entre dispositivos em redes, incluindo a internet. Enquanto o IP é responsável pelo endereçamento e roteamento dos pacotes de dados, o TCP garante a entrega confiável e ordenada desses pacotes, estabelecendo uma conexão entre os dispositivos. Juntos, formam a base da comunicação digital moderna.

quanto um humano são produtos de um projeto, mas apenas um humano é um projetista. A abelha é como o "vento", enquanto o humano é como o "azul".

As causas cognitivas são ontologicamente distintas das causas evolutivas. Elas são feitas de um tipo diferente de material. As causas cognitivas são feitas de neurônios, enquanto as causas evolutivas são feitas de ancestrais.

O tipo mais óbvio de causa cognitiva é intencional, como o desejo de ir ao supermercado ou o planejamento de fazer uma torrada. No entanto, as emoções também existem fisicamente no cérebro, como um tremor de impulsos neurais ou uma nuvem de hormônios em expansão. Da mesma forma, um instinto, um lampejo de visualização ou um pensamento momentaneamente reprimido — se pudéssemos escanear o cérebro em três dimensões e compreender seu código, seríamos capazes de observá-los.

Mesmo as cognições subconscientes existem fisicamente no cérebro. Lord Acton observou que "o poder tende a corromper". Stalin pode ou não ter acreditado ser um altruísta, trabalhando para o bem maior do maior número de pessoas. No entanto, é provável que em algum lugar do cérebro de Stalin existissem circuitos neurais que reforçavam o prazer do exercício do poder e circuitos neurais que detectavam antecipações de aumentos e diminuições de poder. Se o cérebro de Stalin não tivesse nada relacionado ao poder — nenhuma luz que se acendesse para o comando político e se apagasse para a fraqueza política — como ele poderia saber que tinha sido corrompido pelo poder?

As pressões evolutivas de seleção são ontologicamente distintas dos artefatos biológicos que elas criam. A causa evolutiva das asas de um pássaro são os milhões de pássaros ancestrais que se reproduziram com maior frequência do que seus concorrentes, devido às melhorias progressivas em suas asas em comparação com as de outros pássaros. Resumimos esse gigantesco fato histórico-estatístico dizendo que "a evolução fez isso".

A seleção natural é ontologicamente distinta das criaturas; a evolução não é uma pequena criatura peluda à espreita em uma floresta desconhecida. A evolução é uma regularidade estatística causal na história reprodutiva dos ancestrais.

Essa lógica também se aplica ao cérebro. A evolução criou asas que batem, mas que não entendem o movimento de bater. Ela criou pernas que andam, mas que não entendem o ato de caminhar. A evolução moldou ossos de íons de cálcio, mas os próprios ossos não possuem um conceito explícito de força, muito menos de aptidão genética inclusiva. E a evolução projetou os próprios cérebros capazes de projetar; ainda assim, esses cérebros não possuíam mais conceito de evolução do que um pássaro tem de aerodinâmica. Até o século XX, nem mesmo um único cérebro humano representava explicitamente o conceito complexo e abstrato de aptidão genética inclusiva.

Quando nos dizem que "o propósito evolutivo da raiva é aumentar a aptidão genética inclusiva", há uma tendência de pensar que "o propósito da raiva é a reprodução" ou "o propósito cognitivo da raiva é a reprodução". Não! A regularidade estatística da história ancestral não está presente no cérebro, nem mesmo de forma subconsciente, assim como as intenções do projetista de uma torradeira não estão presentes na torradeira em si!

Acreditar que o seu circuito interno de raiva incorpora um desejo explícito de reprodução é o mesmo que acreditar que a sua mão é um desejo mental incorporado de agarrar coisas.

Sua mão não está completamente separada dos seus desejos mentais. Em circunstâncias específicas, você pode controlar o movimento dos seus dedos por um ato de vontade. Se você se abaixar e pegar uma moeda, isso pode representar um ato de vontade; no entanto, não foi um ato de vontade que fez a sua mão se desenvolver em primeiro lugar.

É necessário distinguir um evento específico de raiva (raiva-1, raiva-2, raiva-3) do circuito neural subjacente à raiva. Um evento de raiva é uma causa cognitiva, e um evento de raiva pode ter causas cognitivas, mas você não desejou que o circuito da raiva estivesse conectado ao seu cérebro.

Portanto, é preciso distinguir o evento de raiva, do circuito da raiva, do complexo genético que estabeleceu o modelo neural e do macrofato ancestral que explica a presença desse complexo genético.

Se alguma vez houve uma disciplina que genuinamente exigiu uma atenção extrema aos detalhes, é a psicologia evolutiva.

Considerem, ó meus leitores, esta história sórdida e alegre: Um homem e uma mulher se conhecem em um bar. O homem sente-se atraído pela pele clara e pelos seios firmes dela, que seriam sinais de fertilidade no ambiente ancestral, mas que, nesse caso, resultam de maquiagem e um sutiã. Isso não incomoda o homem; ele simplesmente gosta da aparência dela. Seu circuito neural de detecção de pele clara não sabe que seu objetivo é detectar a fertilidade, assim como os átomos em sua mão não contêm pequenas tags XML[14] que dizem "*<purpose>pick things up</purpose>"* (<propósito>pegar as coisas</propósito). A mulher se sente atraída por seu sorriso confiante e suas maneiras firmes, sinais de status elevado, o que, no ambiente ancestral, significaria a capacidade de fornecer recursos para os filhos. Ela planeja usar controle de natalidade, mas seus detectores de sorrisos confiantes não sabem disso, assim como uma torradeira não sabe que seu projetista pretendia que ela fizesse torradas. Ela não está preocupada filosoficamente com o significado dessa rebelião, porque seu cérebro é um criacionista e nega veementemente que a evolução exista. Ele não está preocupado filosoficamente com o significado dessa rebelião, pois só quer transar. Eles vão para um hotel e se despem. Ele coloca uma camisinha, porque não quer ter filhos, apenas a descarga de dopamina-noradrenalina do sexo, que produziu descendentes de forma confiável há 50.000 anos, quando era uma característica invariável do ambiente ancestral a inexistência de preservativos. Eles fazem sexo, tomam banho e seguem caminhos separados. A principal consequência objetiva é manter o bar, o hotel e o fabricante de preservativos em atividade, o que não era o objetivo cognitivo em suas mentes e não tem praticamente nada a ver com as principais regularidades estatísticas de reprodução de 50.000 anos atrás, que explicam como eles obtiveram os genes que construíram seus cérebros que executaram todo esse comportamento.

Para raciocinar corretamente sobre a psicologia evolucionária, é preciso considerar simultaneamente muitos fatos abstratos complicados que estão fortemente relacionados, mas que são significativamente distintos, sem uma única confusão ou fusão.

---

14 NT. Uma **tag XML** é um elemento básico da linguagem XML (eXtensible Markup Language) usado para estruturar e organizar dados. Ela é composta por um nome entre colchetes angulares (< >), como <nome>, e pode conter atributos, conteúdo ou outras tags aninhadas.

140 — Um experimento de psicologia evolutiva especialmente elegante

> Em um estudo canadense de 1989, solicitou-se a adultos que imaginassem a perda de crianças de diferentes idades e estimassem qual perda causaria o maior sentimento de dor em um dos pais. Os resultados, representados em um gráfico, mostram o luto aumentando até pouco antes da adolescência e, em seguida, diminuindo. Quando essa curva foi comparada a uma curva que mostrava as mudanças no potencial reprodutivo ao longo do ciclo de vida (um padrão calculado com base em dados demográficos canadenses), a correlação foi bastante significativa. Mas muito mais forte - quase perfeita, na verdade - foi a correlação entre as curvas de luto desses canadenses modernos e a curva de potencial reprodutivo de um povo caçador-coletor da África, os !Kung. Em outras palavras, o padrão de mudança do luto era praticamente o que um darwinista poderia prever, considerando as realidades demográficas do ambiente ancestral.[15]
>
> — Robert Wright, em *The Moral Animal* (O animal moral),
>
> resumindo Crawford et al. [1]

A primeira correlação foi de 0,64, enquanto a segunda, extremamente alta, foi de 0,92 (N = 221). Conforme descrito, a principal falha deste estudo é que ele solicitou a adultos humanos que imaginassem a dor dos pais em vez de perguntar a pais reais com filhos de idades específicas. (Presumivelmente, isso teria sido mais custoso/teria permitido menos participantes.) No entanto, entendo que os resultados se alinham bem com os dados de estudos mais aprofundados sobre o luto dos pais, buscando outras correlações (ou seja, a correlação direta entre o luto dos pais e a idade dos filhos).

Dito isso, considere alguns aspectos notáveis deste experimento:

1. Uma correlação de 0,92 (!) pode parecer suspeitosamente alta. A evolução realmente poderia ter gerado um ajuste tão preciso? Sim, poderia! A pressão seletiva não foi apenas suficientemente intensa para ajustar a dor dos pais, mas forte o suficiente para moldá-la completamente desde o início.
2. Aqueles que afirmam que a psicologia evolutiva não fez previsões avançadas são (ironicamente) vítimas do "ninguém sabe o que a ciência não sabe". Ninguém teria sequer cogitado realizar esse experimento se não fosse pela psicologia evolutiva.
3. Este experimento ilustra, da maneira mais bela e clara que já vi, a distinção entre um motivo oculto consciente ou subconsciente e uma adaptação em execução sem sensibilidade em tempo real à pressão seletiva original que a criou.

A dor dos pais não está nem mesmo subconscientemente relacionada ao valor reprodutivo; caso contrário, estaria sintonizada com o valor reprodutivo canadense, e não com o valor reprodutivo !Kung. O luto é uma adaptação que simplesmente existe agora, real na mente, e continua a seguir sua própria inércia.

---

15 NT. Texto original em inglês. I*n a 1989 Canadian study, adults were asked to imagine the death of children of various ages and estimate which deaths would create the greatest sense of loss in a parent. The results, plotted on a graph, show grief growing until just before adolescence and then beginning to drop. When this curve was compared with a curve showing changes in reproductive potential over the life cycle (a pattern calculated from Canadian demographic data), the correlation was fairly strong. But much stronger—nearly perfect, in fact—was the correlation between the grief curves of these modern Canadians and the reproductive-potential curve of a hunter-gatherer people, the !Kung of Africa. In other words, the pattern of changing grief was almost exactly what a Darwinian would predict, given demographic realities in the ancestral environment.*

Os pais não se importam com os filhos por causa de sua contribuição reprodutiva. Os pais se importam com os filhos pelo bem deles próprios; e a razão não cognitiva e histórico-evolutiva pela qual tais mentes existem no universo em primeiro lugar é porque as crianças carregam os genes de seus pais.

De fato, a evolução é a razão pela qual existem mentes no universo. Portanto, você pode entender por que eu gostaria de traçar uma linha nítida no meu cinismo em relação a segundas intenções no limite evolutivo-cognitivo; caso contrário, eu poderia muito bem ficar na fila do caixa de um supermercado e dizer: "Ei! Você só está processando corretamente as informações visuais enquanto ensaca minhas compras para maximizar sua aptidão genética inclusiva!"

(1) Acredito que a correlação de 0,92 seja a mais alta que já vi em qualquer experimento de psicologia evolutiva e, de fato, uma das correlações mais altas que já encontrei em qualquer experimento psicológico. (Embora eu tenha visto, por exemplo, uma correlação de 0,98 relatada ao perguntar a um grupo de participantes "Quão similar é A em relação a B?" e a outro grupo "Qual é a probabilidade de A dado B?" em questões como "Qual é a probabilidade de você retirar 60 bolas vermelhas e 40 bolas brancas deste barril contendo 800 bolas vermelhas e 200 bolas brancas?" — Em outras palavras, os participantes simplesmente as processaram como a mesma pergunta).

Dado que todos somos bayesianos aqui, podemos considerar nossas prioridades e questionar se pelo menos parte dessa correlação inesperadamente alta se deve à sorte. Podemos dar como certo o ajuste fino evolutivo, considerando a imensa pressão de seleção envolvida. As fontes restantes de variação suspeitosamente baixas são: (a) se um grande grupo de adultos poderia corretamente visualizar, em média, os níveis relativos de luto parental (aparentemente, eles podem) e (b) se os sobreviventes !Kung são representativos dos caçadores-coletores ancestrais nessa dimensão, ou se a variação entre os diferentes grupos tribais de caçadores-coletores deveria ter sido muito alta para permitir uma correlação de 0,92.

No entanto, mesmo considerando quaisquer céticos anteriores, uma correlação de 0,92 com N = 221 é uma evidência bastante forte, e nossas conclusões devem ser menos céticas em relação a todos esses aspectos.

(2) Alguém pode considerar como falta de elegância do experimento, ter sido conduzido prospectivamente com luto imaginário, em vez de retrospectivamente com luto real. No entanto, é o luto prospectivamente imaginado que realmente funciona para desviar o comportamento dos pais diante da perda de um filho! Do ponto de vista evolutivo, uma criança falecida é um custo irrecuperável; a evolução "deseja" que os pais aprendam com a dor, evitem repetir o mesmo erro, se ajustem de volta ao seu estado hedônico anterior e continuem criando outros filhos.

(3) Da mesma forma, o gráfico que se correlaciona com a dor dos pais diz respeito ao potencial reprodutivo futuro de uma criança que sobreviveu até uma determinada idade, e não ao custo irrecuperável de criar a criança até essa idade. (Poderíamos obter uma correlação ainda maior se tentássemos considerar o custo de oportunidade reprodutiva de criar uma criança da idade X até a maturidade independente, enquanto descartamos todos os custos irrecuperáveis de criar uma criança até a idade X?)

Em geral, os seres humanos percebem os custos irrecuperáveis — presumivelmente, isso é uma adaptação para nos impedir de mudar de estratégia com muita frequência (compensando um observador de oportunidades ansioso demais?) ou um reflexo infeliz da dor sentida com o desperdício de recursos.

Por outro lado, a evolução — não é que a evolução "se preocupe com os custos irrecuperáveis", mas sim que a evolução não "pensa" remotamente dessa forma; a "evolução" é apenas um fato macro sobre as reais consequências históricas reprodutivas.

Portanto, é claro que a adaptação do luto dos pais é ajustada de uma forma que não tem nada a ver com o investimento passado em uma criança e tem tudo a ver com as futuras consequências reprodutivas da perda dessa criança. A seleção natural não se importa com custos irrecuperáveis como nós.

No entanto, é claro que a adaptação ao luto dos pais continua funcionando como se os pais estivessem vivendo em uma tribo !Kung em vez do Canadá. A maioria dos seres humanos notaria a diferença.

Os seres humanos e a seleção natural são insanos de maneiras complicadas e estáveis, mas diferentes.

Referências

[1] Robert Wright, The Moral Animal: Why We Are the Way We Are: The New Science of Evolutionary Psychology (Pantheon Books, 1994); Charles B. Crawford, Brenda E. Salter, and Kerry L. Jang, “Human Grief: Is Its Intensity Related to the Reproductive Value of the Deceased?“ Ethology and Sociobiology 10, no. 4 (1989): 297–307.

141 — Superestímulos e o colapso da civilização ocidental

Pelo menos três pessoas morreram após passarem dias jogando jogos online sem descanso. Elas perderam seus cônjuges, empregos e filhos para o World of Warcraft. Se as pessoas têm o direito de jogar videogames — e é difícil imaginar um direito mais fundamental — então o mercado responderá oferecendo os videogames mais cativantes possíveis, ao ponto de consumidores excessivamente engajados serem excluídos do pool genético.

Como um produto de consumo se torna tão envolvente que, após 57 horas de uso, o consumidor prefere continuar usando o produto por mais uma hora em vez de comer ou dormir? (Alguém poderia argumentar que o consumidor decide racionalmente de que prefere jogar Starcraft por mais uma hora do que viver o resto de sua vida, mas evitaremos essa discussão. Por favor.)

Uma barra de chocolate é um superestímulo: contém mais açúcar concentrado, sal e gordura do que qualquer coisa que existia no ambiente ancestral. Ela ativa as papilas gustativas que evoluíram em um ambiente caçador-coletor, porém de maneira muito mais intensa do que qualquer coisa que realmente existisse nesse ambiente. O sinal que antes se correlacionava de forma confiável com alimentos saudáveis foi sequestrado, representado com um ponto no espaço do sabor que não estava presente nos dados de treinamento — um valor discrepante e impossivelmente distante nos gráficos ancestrais antigos. O sabor, que anteriormente representava os correlatos evolutivamente identificados da saúde, foi revertido por engenharia e perfeitamente combinado com uma substância artificial. Infelizmente, não há um incentivo de mercado igualmente poderoso para tornar o alimento resultante tão saudável quanto saboroso. Afinal, não podemos provar a saúde.

O famoso vídeo "Dove Evolution" mostra a meticulosa construção de outro superestímulo: uma mulher comum transformada por meio de maquiagem, fotografia cuidadosa e, por fim, extensas edições de Photoshop, em uma modelo de outdoor — uma beleza impossível, inalcançável por mulheres reais no mundo intocado. Mulheres reais estão se prejudicando (por exemplo, supermodelos usando cocaína para manter o peso baixo) para competir com concorrentes que literalmente não existem.

Da mesma forma, um videogame pode ser muito mais envolvente do que a mera realidade, mesmo por meio de um simples monitor de computador, levando alguém a jogar sem comer ou dormir até a morte literal. Não conheço todos os truques utilizados nos videogames, mas posso presumir alguns deles — desafios posicionados no ponto crítico entre o fácil e o impossível, reforço intermitente, feedback mostrando uma pontuação cada vez mais alta, envolvimento social em jogos multijogador massivos.

Existe um limite para o incentivo do mercado em tornar os videogames cada vez mais envolventes? Podemos esperar que não haja incentivo além do ponto no qual os jogadores perdem seus empregos; afinal, eles devem conseguir pagar suas mensalidades. Isso implica em um "ponto ideal" para o vício em jogos, em que a moda da curva do sino é a diversão, e apenas algumas almas infelizes na

cauda se tornam viciadas a ponto de perderem seus empregos. Em 2007, jogar *World of Warcraft*[16] por 58 horas consecutivas até a morte ainda era a exceção, e não a regra. Mas os fabricantes de videogames competem entre si, e se você puder tornar seu jogo 5% mais viciante, poderá roubar 50% dos clientes de seu concorrente. É fácil perceber como esse problema pode piorar consideravelmente.

Se as pessoas têm o direito de serem tentadas — e é disso que se trata o livre-arbítrio — o mercado responderá fornecendo o máximo de tentações que puderem ser vendidas. O incentivo é tornar seus estímulos 5% mais atraentes do que os de seus principais concorrentes atuais. Isso vai muito além do ponto no qual os estímulos se tornam superestímulos ancestralmente anômalos. Pense em como nossos padrões de venda de produtos de beleza feminina mudaram desde os anúncios da década de 1950. E, como as barras de chocolate demonstram, o incentivo de mercado também vai além do ponto no qual o superestímulo começa a causar danos colaterais ao consumidor.

Então, por que simplesmente não dizemos "não"? Uma suposição fundamental da economia de livre mercado é que, na ausência de coerção e fraude, as pessoas sempre podem se recusar a participar de uma transação prejudicial. (Se essa suposição for verdadeira, um mercado livre não apenas seria a melhor política, em geral, mas também teria poucos ou nenhum efeito negativo.)

Um organismo que regularmente deixa de comer morrerá, como alguns jogadores de videogame descobriram da maneira mais difícil. Mas, em algumas ocasiões no ambiente ancestral, um comportamento tipicamente benéfico (e, portanto, tentador) pode, de fato, ser prejudicial.Os seres humanos, como organismos, possuem uma capacidade excepcionalmente forte de perceber essas situações especiais por meio do pensamento abstrato. Por outro lado, também tendemos a imaginar muitas consequências de situações especiais que não existem, como espíritos ancestrais nos ordenando a não comer coelhos perfeitamente saudáveis.

A evolução parece ter chegado a um meio-termo, ou talvez tenha simplesmente agregado novos sistemas aos antigos. O *Homo sapiens* ainda é tentado pela comida, mas nossos córtices pré-frontais superdimensionados nos conferem uma capacidade limitada de resistir à tentação. Essa habilidade não é ilimitada — nossos antepassados com grande força de vontade provavelmente passaram fome em sacrifício aos deuses ou fracassaram em resistir ao adultério em diversas ocasiões. Os jogadores de videogame que morreram provavelmente exerceram alguma forma de força de vontade para continuar jogando por tanto tempo sem comer ou dormir; o perigo evolutivo do autocontrole.

Resistir a qualquer tentação requer um consumo consciente de uma reserva limitada de energia mental. Na realidade, não podemos simplesmente dizer "não" sem incorrer em algum custo pessoal. Mesmo aqueles que têm a força de vontade ou previsão para vencer na loteria ainda pagam um preço ao resistir à tentação. O preço é apenas mais facilmente pago.

Nossa força de vontade limitada evoluiu para lidar com as tentações ancestrais; pode não ser eficaz diante de tentações que vão além das conhecidas pelos caçadores-coletores. Mesmo quando conseguimos resistir a superestímulos, é plausível que o esforço necessário esgote nossa força de vontade muito mais rapidamente do que resistir às tentações ancestrais.

A exposição pública a superestímulos acaba gerando uma externalidade negativa, inclusive para aqueles que dizem não ceder. Será que deveríamos proibir anúncios de biscoitos de chocolate ou vitrines que exibem abertamente sorvetes?

A existência de um problema não implica necessariamente (sem uma justificativa adicional e

16 NT. **World of Warcraft** (WoW) é um MMORPG (Massively Multiplayer Online Role-Playing Game) lançado em 2004 pela Blizzard Entertainment, ambientado no universo de Warcraft. Os jogadores exploram o mundo de Azeroth, completam missões, participam de batalhas e interagem com outros jogadores em um ambiente persistente e em constante evolução. O jogo é conhecido por sua narrativa rica, gráficos impressionantes e impacto duradouro na cultura dos games.

um ônus substancial de prova) que o governo possa resolvê-lo. O incentivo da carreira do regulador não se concentra em produtos que combinam danos de baixo grau aos consumidores com superestímulos viciantes; ele se concentra em produtos com modos de falhas espetaculares o suficiente para terem cobertura jornalística. Por outro lado, o fato de o governo não conseguir solucionar algo não significa que não esteja acontecendo algo errado.

Deixo vocês com um argumento final baseado em evidências fictícias: o romance online de Simon Funk, *After Life* (Depois da vida), retrata (entre outros elementos da trama) a extinção planejada do *Homo sapiens* biológico - não por meio de exércitos de robôs em marcha, mas por meio de crianças artificiais muito mais adoráveis, doces e divertidas de criar do que crianças reais. Talvez o colapso demográfico das sociedades avançadas ocorra porque o mercado oferece alternativas cada vez mais tentadoras para ter filhos, enquanto a tarefa de trocar fraldas permanece inalterada ao longo do tempo. Onde estão os outdoors que promovem a "raça"? Quem pagará por consultores de imagem profissionais para tornar a discussão com adolescentes mal-humorados, mais atraente do que uma viagem ao Taiti?

"No final", escreveu Simon Funk, "a espécie humana foi simplesmente comercializada até desaparecer".

142 — Tu és *Estilhaço Divino (Godshatter)*[17]

Antes do século XX, nenhum ser humano tinha um conceito explícito de "aptidão genética inclusiva", a única e absoluta obsessão do deus cego e insensato. Não possuímos repulsa instintiva por preservativos ou sexo oral. Nossos cérebros, esses órgãos reprodutivos supremos, não verificam a eficácia reprodutiva antes de nos proporcionar prazer sexual.

Por quê não? Por que não somos conscientemente obcecados com a aptidão genética inclusiva? Por que a Fada da Evolução dos Humanos criou cérebros que inventariam preservativos? "Teria sido tão fácil", pensa o humano, que pode projetar novos sistemas complexos em uma tarde.

A Fada da Evolução, como todos sabemos, é obcecada pela aptidão genética inclusiva. Quando ela decide quais genes promover à universalidade, parece não considerar nada além do número de cópias que um gene produz. (Que estranho!)

Mas, uma vez que o criador da inteligência é obcecado dessa forma, por que não criar agentes inteligentes - você não pode chamá-los de humanos - que também se importariam puramente com a aptidão genética inclusiva? Tais agentes teriam relações sexuais apenas como meio de reprodução e não se importariam com sexo que envolvesse controle de natalidade. Eles poderiam comer alimentos com base em uma crença explicitamente fundamentada de que o alimento era necessário para se reproduzir, não porque gostassem do sabor, e, portanto, não consumiriam doces se isso prejudicasse a sobrevivência ou a reprodução. As mulheres na pós-menopausa cuidariam dos netos até que ficassem doentes o suficiente para drenar os recursos e, então, cometeriam suicídio.

Parece uma melhoria de projeto tão óbvia - do ponto de vista da Fada da Evolução.

Agora está claro que é difícil construir um agente consequencialista poderoso o suficiente. A seleção natural meio que raciocina consequentemente, mas apenas com base nas consequências reais. Os teóricos da evolução humana precisam realizar um raciocínio abstrato realmente complexo para imaginar as ligações entre as adaptações e o sucesso reprodutivo.

Mas os cérebros humanos podem claramente imaginar essas ligações das proteínas. Então, quando a Fada da Evolução criou os humanos, por que ela se preocupou com qualquer outra motivação, exceto a aptidão genética inclusiva?

Faz menos de dois séculos desde que um cérebro de proteína representou pela primeira vez o conceito de seleção natural. A noção moderna de "aptidão genética inclusiva" é ainda mais sutil, um conceito altamente abstrato. O que importa não é o número de genes compartilhados. Os chimpanzés compartilham 95% de seus genes. O que importa é a variância genética compartilhada em uma população reprodutiva - sua irmã tem metade de seu parentesco, porque qualquer variação em seu genoma, na espécie humana, tem 50% de probabilidade de ser compartilhada por sua irmã.

Somente no último século - sem dúvida, nos últimos cinquenta anos - os biólogos evolucionistas re-

17 NT. **Estilhaço Divino** (ou ***Godshatter,*** no original) é um conceito presente na série de quadrinhos "*Planetary*", escrita por Warren Ellis e ilustrada por John Cassaday. Refere-se a fragmentos de energia cósmica ou divina que concedem poderes extraordinários a quem os possui, muitas vezes ligados a eventos traumáticos ou transformadores. Esses fragmentos são explorados como símbolos de mudança, poder e conexão com forças maiores no universo da história.

almente começaram a entender toda a gama de causas do sucesso reprodutivo, como o altruísmo recíproco e a sinalização custosa. Sem todo esse conhecimento altamente detalhado, um agente inteligente que se propusesse a "maximizar a aptidão inclusiva" falharia miseravelmente.

Então, por que não pré-programar os cérebros de proteína com esse conhecimento? Por que o conceito de "aptidão genética inclusiva" não foi programado em nós, juntamente com uma biblioteca de estratégias explícitas? Dessa forma, poderíamos dispensar todos os mecanismos de recompensa. O organismo nasceria sabendo que, com grande probabilidade, alimentos gordurosos levariam ao condicionamento físico. Se o organismo descobrisse posteriormente que não era mais assim, pararia de comer alimentos gordurosos. Seria possível reconfigurar todo o sistema. E não seriam necessários inventos como camisinhas ou biscoitos.

Isso parece perfeitamente possível em princípio. Às vezes, encontro pessoas que não entendem muito bem o consequencialismo e dizem: "Mas se o organismo não tiver um impulso separado para comer, ele morrerá de fome e, portanto, deixará de se reproduzir". Contudo, se o organismo estiver ciente desse fato e tiver uma função de utilidade que valorize a reprodução, ele automaticamente se alimentará. Na verdade, esse é exatamente o raciocínio consequencialista que a própria seleção natural usou para criar seres que comem automaticamente.

E quanto à curiosidade? Um consequencialista não ficaria curioso apenas quando visse um motivo específico para isso? E isso não faria com que ele perdesse muitos conhecimentos importantes que surgiram sem nenhuma razão específica para investigação? Novamente, um consequencialista investigará apenas com base no conhecimento desse mesmo fato. Se considerarmos o impulso de curiosidade em seres humanos - que não é indiscriminado, mas responde a características específicas dos problemas - então essa adaptação complexa é simplesmente o resultado do raciocínio consequencialista do DNA, uma representação implícita do conhecimento: os ancestrais que se engajaram nesse tipo de investigação deixaram mais descendentes.

Portanto, em princípio, é possível ter um consequencialista reprodutivo puro. Em princípio, toda a história ancestral representada implicitamente em adaptações cognitivas pode ser convertida em conhecimento explicitamente representado, operando em um núcleo consequencialista.

No entanto, o deus cego e insensato não é tão inteligente. A evolução não é um programador humano capaz de refatorar arquiteturas inteiras de código simultaneamente. A evolução não é um programador humano que pode sentar e digitar instruções a sessenta palavras por minuto.

Por milhões de anos antes do consequencialismo hominídeo, existia a aprendizagem por reforço. Os sinais de recompensa eram eventos que se correlacionavam de forma confiável com a reprodução. Não se pode pedir a um cérebro não hominídeo que preveja que uma criança que coma alimentos gordurosos agora, sobreviverá ao inverno. Portanto, o DNA constrói um cérebro de proteína que gera um sinal de recompensa ao comer alimentos gordurosos. Cabe ao organismo, então, descobrir quais presas são mais saborosas.

O DNA constrói cérebros proteicos com sinais de recompensa que possuem uma correlação distante com a aptidão reprodutiva, mas uma correlação próxima com o comportamento do organismo. Você não precisa descobrir que comer alimentos açucarados no outono levará à digestão de calorias que podem ser armazenadas em gordura para sobreviver no inverno, se acasalar na primavera e produzir descendentes no verão. Uma maçã simplesmente tem um sabor bom, e seu cérebro só precisa planejar como obter mais maçãs da árvore.

Consequentemente, os organismos desenvolvem recompensas por comer, construir ninhos, assustar concorrentes, ajudar irmãos, descobrir verdades importantes, formar alianças fortes, discutir persuasivamente e, é claro, fazer sexo… Quando os cérebros dos hominídeos capazes de raciocínio consequencial de vários domínios começaram a aparecer, eles raciocinaram consequentemente sobre como obter os reforçadores existentes.

Foi um "truque" relativamente simples, muito mais simples do que reconstruir um "maximizador de condicionamento físico inclusivo" do zero. Os cérebros de proteína planejaram como adquirir calorias e sexo, sem qualquer representação cognitiva explícita de "aptidão inclusiva".

Um engenheiro humano teria dito: "Uau, acabei de inventar um consequencialista! Agora posso pegar todo o conhecimento anterior adquirido com muito esforço sobre quais comportamentos melhoram o

condicionamento físico e declará-lo explicitamente! Posso converter todo esse mecanismo complicado de aprendizado por reforço em uma declaração de conhecimento declarativa simples, como 'alimentos gordurosos e sexo geralmente melhoram seu condicionamento físico inclusivo'. O raciocínio consequente cuidará automaticamente do resto. Além disso, não terá o modo de falha óbvio de inventar preservativos!"

Mas um engenheiro humano também não teria construído a retina ao contrário.

O deus cego e insensato não é um propósito único, mas uma atenção multifragmentada. As raposas evoluem para pegar coelhos, os coelhos evoluem para fugir das raposas; há tantas evoluções quanto espécies. Mas dentro de cada espécie, o deus cego e insensato é obcecado puramente com a aptidão genética inclusiva. Nenhuma qualidade é valorizada, nem mesmo a sobrevivência, exceto enquanto aumenta a aptidão reprodutiva. De nada adianta um organismo com pele de aço se isso resultar em 1% a menos de capacidade reprodutiva.

Ainda assim, quando o deus cego e insensato criou os computadores de proteína, seu foco monomaníaco na aptidão genética inclusiva não foi transmitido fielmente. Seu critério de otimização não funcionou com sucesso. Nós, obra da evolução, somos tão alheios à evolução quanto nosso Criador é alheio a nós. Uma função de utilidade pura dividida em milhares de fragmentos de desejos.

Por quê? Acima de tudo, porque a evolução é estúpida em um sentido absoluto. Mas também porque os primeiros computadores de proteína não eram nem de longe tão gerais quanto o deus cego e insensato, e só podiam utilizar desejos de curto prazo.

Em última análise, questionar por que a evolução não nos projetou para maximizar a aptidão genética inclusiva é como questionar por que ela não deu aos seres humanos um ribossomo e os ordenou a projetar sua própria bioquímica. É por isso que a evolução não consegue refatorar o código genético com tanta rapidez. Contudo, talvez em um bilhão de anos de seleção natural contínua, seja exatamente isso que aconteceria, caso a inteligência fosse suficientemente tola para permitir que o deus insensato continuasse reinando.

Em *The Mote in God's Eye* (O Cisco nos Olhos de Deus), de Niven e Pournelle, é retratada uma espécie inteligente que permaneceu biológica por tempo demais, gradualmente se tornando lentamente escravizada pela evolução, transformando-se gradualmente em verdadeiros maximizadores do condicionamento físico, obcecados em se reproduzir mais do que os outros. Porém, felizmente, não foi isso que ocorreu. Não aqui na Terra. Pelo menos, ainda não.

Assim sendo, os seres humanos adoram o sabor do açúcar e da gordura, e amamos nossos filhos e filhas. Buscamos posição social e sexo. Cantamos, dançamos e tocamos instrumentos. Aprendemos por amor ao conhecimento. Milhares de sabores deliciosos, combinados com reforçadores antigos que antes se correlacionavam com a aptidão reprodutiva - agora procuramos saber se eles melhoram ou não a reprodução. Sexo com controle de natalidade, chocolate, a música de Bach, falecido há muito tempo, gravada em um CD.

E quando finalmente aprendemos sobre a evolução, pensamos: 'Ficar obcecado o dia todo com a aptidão genética inclusiva? Onde está a diversão nisso?"

O único objetivo monomaníaco do deus insensato e cego fragmentou-se em mil pedaços de desejo. E isso é bom, eu acho, embora eu seja um humano dizendo isso. Do contrário, o que faríamos com o futuro? O que faríamos com os bilhões de galáxias no céu noturno? Preenchê-las com replicadores de eficiência máxima? Nossos descendentes deveriam ficar deliberadamente obcecados em maximizar sua aptidão genética inclusiva, considerando todo o resto apenas como um meio para esse fim?

Ser fragmentado em mil pedaços de desejo nem sempre é divertido, mas pelo menos não é entediante. Em algum ponto ao longo do caminho, desenvolvemos apreço pela novidade, complexidade, elegância e desafio — preferências que avaliam o enfoque monomaníaco do deus insensato e cego como esteticamente insatisfatório.

E sim, compartilhamos essas mesmas preferências com o estilhaço do deus insensato e cego. E daí?

Parte M — Propósitos frágeis

143 — Crença na inteligência

Não sei que jogadas Garry Kasparov faria em um jogo de xadrez. Então, qual é o conteúdo empírico da minha crença de que "Kasparov[18] é um jogador de xadrez altamente inteligente"? Que experiência do mundo real essa crença me leva a antecipar? Seria uma forma habilmente disfarçada de total ignorância?

Para aprofundar o dilema, suponha que Kasparov esteja jogando contra um mero grande mestre de xadrez, Sr. G, que não está competindo pelo título de campeão mundial. Minha própria habilidade é muito baixa para distinguir entre esses níveis de jogo. Ao tentar adivinhar a jogada de Kasparov ou a próxima jogada do Sr. G, tudo que posso fazer é tentar adivinhar "a melhor jogada de xadrez" usando meu conhecimento limitado do jogo. Portanto, eu faria a mesma previsão para a jogada de Kasparov ou para a jogada do Sr. G em qualquer posição específica do xadrez. Então, qual é o conteúdo empírico da minha crença de que Kasparov é um jogador de xadrez melhor do que o Sr. G"?

O conteúdo empírico da minha crença é a previsão testável e falsificável de que a posição final do xadrez estará na categoria de posições em que Kasparov vence, em vez de jogos empatados ou vitórias para o Sr. G. (Considerando a renúncia como um movimento legal que leva a uma posição de xadrez classificada como perdida). O grau em que acredito que Kasparov é um "jogador melhor" reflete-se na probabilidade que atribuo à categoria de resultados "Kasparov vence" em relação às categorias de "jogo empatado" e "Sr. G vence". Essas categorias são extremamente abrangentes no sentido de que se referem a vastos espaços de possíveis posições de xadrez — mas "Kasparov vence" é mais específico do que a máxima entropia, ao poder ser definitivamente refutado por um vasto conjunto de posições de xadrez.

O resultado do jogo de Kasparov é previsível porque conheço e entendo seus objetivos. Nos limites do tabuleiro de xadrez, conheço as motivações de Kasparov — conheço seu critério de sucesso, sua função de utilidade, seu objetivo como um processo de otimização. Sei para onde Kasparov está tentando direcionar o futuro e prevejo que ele é capaz o suficiente para alcançar esse objetivo, embora não preveja muito sobre como Kasparov o fará.

Imaginem que estou visitando uma cidade distante e um amigo local se oferece para me levar ao aeroporto. Não conheço bem o bairro. Cada vez que meu amigo se aproxima de uma interseção, não sei se ele virará à esquerda, à direita ou seguir em frente. Não consigo prever o movimento do meu amigo mesmo quando nos aproximamos de cada interseção individual — muito menos prever toda a sequência de movimentos com antecedência.

Ainda assim, posso prever o resultado das ações imprevisíveis do meu amigo: chegaremos ao aeroporto. Mesmo que a casa do meu amigo esteja em outro lugar da cidade, fazendo com que ele faça uma sequência de curvas completamente diferente, prevejo com a mesma confiança que chegaremos ao aeroporto. Posso prever isso com bastante antecedência, antes mesmo de entrar no carro. Meu voo parte em breve e não há tempo a perder; eu nem entraria no carro se não pudesse prever com segurança que ele me levaria ao aeroporto por um caminho imprevisível.

---

18 NT. **Garry Kasparov** é um grande mestre de xadrez russo, considerado um dos maiores enxadristas da história. Foi campeão mundial de xadrez de 1985 a 2000 e ficou famoso por suas partidas contra o supercomputador **Deep Blue** da IBM, em 1996 e 1997, que marcaram um marco na relação entre humanos e inteligência artificial. Além de sua carreira no xadrez, Kasparov é ativista político e escritor, defendendo a democracia e a liberdade.

Não é uma situação notável do ponto de vista científico? Posso prever o resultado de um processo sem conseguir prever nenhuma das etapas intermediárias desse processo.

Como isso é possível? Normalmente, alguém prevê imaginando o presente e, em seguida, avançando a visualização no tempo. Se você deseja um modelo preciso do Sistema Solar, considerando as perturbações planetárias, deve começar com um modelo de todos os objetos principais e executar esse modelo ao longo do tempo, passo a passo.

Às vezes, problemas mais simples têm uma solução fechada, onde calcular o futuro no tempo T requer a mesma quantidade de trabalho, independentemente de T. Uma moeda repousa sobre uma mesa e, a cada minuto, a moeda é virada. A moeda começa mostrando a face. Qual face ela mostrará cem minutos depois? Obviamente, você não respondeu a essa pergunta visualizando uma centena de etapas intermediárias.

Mas quando meu amigo me leva ao aeroporto, posso prever o resultado com sucesso usando um modelo estranho que não funcionaria para prever nenhuma das etapas intermediárias. Meu modelo nem exige que eu insira as condições iniciais — não preciso saber de onde começamos na cidade!

É essencial que eu saiba algo sobre o meu amigo. Preciso estar ciente de que ele deseja que eu pegue meu voo e reconhecer que ele é um planejador competente o suficiente para me levar com êxito ao aeroporto (caso assim o queira). Essas são características do estado inicial do meu amigo, características que me permitem prever o destino final, embora não as curvas intermediárias.

Também devo atribuir ao meu amigo o conhecimento suficiente sobre a cidade para dirigir com sucesso. Isso pode ser considerado uma relação entre meu amigo e a cidade; portanto, é uma propriedade compartilhada por ambos. No entanto, trata-se de uma propriedade extremamente abstrata, que não requer conhecimento específico nem sobre a cidade, nem sobre o conhecimento do meu amigo sobre a cidade.

Essa é uma perspectiva na qual dediquei minha vida — essas situações notáveis que nos colocam em posições epistêmicas tão peculiares. E, de certa forma, meu trabalho pode ser visto como revelar a forma precisa desse conhecimento abstrato estranho que podemos possuir; por meio do qual, mesmo sem conhecer as ações, podemos conhecer justificadamente as consequências.

“Inteligência“ é um termo bastante restrito para descrever essas situações notáveis em sua generalidade. Eu diria, antes, “processo de otimização“. Uma situação semelhante acompanha o estudo da seleção natural na biologia, por exemplo; não podemos prever a forma exata do próximo organismo observado.

No entanto, minha especialidade está centrada no tipo de processo de otimização chamado “inteligência“; e ainda mais específico, um tipo particular de inteligência denominada “Inteligência Artificial Amigável“ — da qual, espero, conseguirei obter um conhecimento abstrato especialmente preciso.

144 — Humanos em trajes engraçados

Muitas vezes, a humanidade viajou para o espaço apenas para descobrir que as estrelas eram habitadas por alienígenas que se assemelhavam muito a seres humanos vestidos em trajes engraçados — ou até mesmo humanos com uma pitada de maquiagem e látex — ou simplesmente caucasianos beges em trajes simples.

Star Trek: A Série Original[19], "Arena", © CBS Corporation

É impressionante como a forma humana se tornou a linha de base natural do universo, a partir da qual todas as outras espécies exóticas foram criadas com algumas modificações.

O que poderia explicar esse fenômeno fascinante? Convergência evolutiva, é claro! Mesmo que essas formas de vida alienígenas tenham evoluído em mil planetas diferentes, completamente independentes da vida na Terra, todas elas acabaram se tornando semelhantes.

Não se engane pelo fato de que um canguru (um mamífero) se parece menos conosco do que um chimpanzé (um primata), ou que um sapo (um anfíbio, assim como nós, tetrápodes) se parece menos conosco do que um canguru. Não se engane pela desconcertante variedade de insetos, que se separaram de nós há mais tempo do que os sapos; não se engane pensando que os insetos possuem seis patas, exoesqueletos e um sistema óptico e práticas sexuais bastante diferentes.

Você pode pensar que uma espécie verdadeiramente alienígena seria mais diferente de nós do que somos dos insetos. Mas não se engane. Para uma espécie alienígena desenvolver inteligência, ela precisaria ter duas pernas com um joelho cada, ligadas a um torso ereto e caminhar de maneira semelhante a nós. Veja, qualquer forma de inteligência precisa de mãos, então seria necessário adaptar um par de pernas para isso — e se a espécie alienígena não começar com uma estrutura de quatro patas, ela não poderia desenvolver uma marcha de corrida e caminhar ereto, liberando as mãos.

19 NT. **Star Trek** é uma franquia de ficção científica criada por Gene Roddenberry em 1966, que acompanha as aventuras da tripulação da nave estelar **USS Enterprise** e outras embarcações em sua missão de explorar o espaço, buscar novas civilizações e promover a paz. Conhecida por abordar temas sociais, filosóficos e científicos, a série original inspirou diversas spin-offs, filmes, livros e uma legião de fãs, tornando-se um marco cultural e influente na história da televisão e do cinema.

… Ou talvez devêssemos considerar, como uma teoria alternativa, que seja mais fácil usar seres humanos em trajes engraçados.

No entanto, o verdadeiro problema não está na forma; está na mente. “Humanos em trajes engraçados“ é um termo amplamente conhecido entre os fãs da literatura de ficção científica e não se refere a uma criatura com quatro membros que caminha ereta. Uma criatura angular feita de cristal é considerada um “humano em trajes engraçados“ se ela pensar notavelmente como um humano — especialmente um humano de uma cultura de língua inglesa do final do século XX ou início do século XXI.

Não assisto a muitos filmes antigos. Quando assisti ao filme “Psicose[20]” (1960) alguns anos atrás, fiquei surpreso com a diferença cultural entre os americanos retratados na tela e a minha América. Os personagens de “Psicose“, vestidos com camisas abotoadas, são consideravelmente mais alienígenas do que a grande maioria dos chamados “alienígenas” que encontro na TV ou no cinema.

Para escrever sobre uma cultura que não seja exatamente como a sua, é preciso conseguir ver sua própria cultura como um caso especial — não como uma norma que todas as outras culturas devem adotar como ponto de partida. Estudar história pode ajudar, mas no final são apenas palavras impressas em páginas brancas, não uma experiência viva. Às vezes, me pergunto que coisas eu possa estar perdendo (não lá, mas aqui).

Ver a própria humanidade como um caso especial é muito mais difícil do que isso.

Em todas as culturas conhecidas, os humanos parecem sentir alegria, tristeza, medo, repulsa, raiva e surpresa. Em todas as culturas conhecidas, essas emoções são expressas pelas mesmas expressões faciais. Na próxima vez que você encontrar um “alienígena“ - ou uma “IA“, aliás - aposto que quando ele ficar com raiva (e ele vai ficar com raiva), ele mostrará a expressão facial humana universal para a raiva.

Nós, humanos, somos muito semelhantes por baixo da nossa pele - afinal, somos uma espécie que se reproduz sexualmente; não seria viável ter diferentes adaptações complexas, elas simplesmente não se encaixariam. (Os alienígenas se reproduzem sexualmente, assim como os humanos e muitos insetos? Eles compartilham pequenos trechos de material genético, como as bactérias? Eles formam colônias, como os fungos? Será que a regra da unidade psicológica também se aplica a eles?)

As únicas inteligências que nossos ancestrais tiveram que manipular — de forma complexa, e não apenas domesticar ou capturar em redes — as únicas mentes que eles tiveram que modelar em detalhes, eram mentes que funcionavam semelhantemente às deles. E assim, evoluímos para prever outras mentes nos colocando em seus lugares, imaginando o que faríamos em suas situações; pois o que deveria ser previsto, era semelhante ao preditor.

“O quê?” você pode dizer. “Eu não presumo que outras pessoas sejam como eu! Talvez eu esteja triste e elas estejam com raiva! Elas acreditam em coisas diferentes de mim; suas personalidades são diferentes das minhas!” Veja por outro ângulo: um cérebro humano é um sistema físico extremamente complexo. Você não está modelando neurônio por neurônio ou átomo por átomo. Se você encontrasse um sistema físico tão complexo quanto o cérebro humano que fosse diferente de você, levaria vidas científicas para desvendá-lo. Você não entende como o cérebro humano funciona de maneira abstrata e geral; você não pode construir um e nem mesmo um modelo de computador que preveja outros cérebros tão bem quanto você os prevê.

A única razão pela qual você pode tentar compreender algo tão complexo e mal compreendido como o cérebro de outro ser humano é porque você configura o seu próprio cérebro para imitá-lo. Você se empatiza (mesmo que talvez não simpatize). Você impõe ao seu próprio cérebro uma sombra da raiva da outra mente e a sombra de suas crenças. Você pode nunca pensar nas palavras “O que eu faria nesta situação?“, mas aquela pequena sombra da outra mente que você mantém dentro de si é algo que ganha vida dentro do seu próprio cérebro, invocando o mesmo mecanismo complexo que existe na outra pessoa, sincronizando engrenagens que você não compreende. Mesmo que você não esteja com raiva, você sabe que se estivesse no lugar da outra pessoa e acreditasse que você é uma escória sem valor, tentaria machucá-la…

20 NT. **Psicose** (*Psycho*), dirigido por Alfred Hitchcock e lançado em 1960, é um clássico do suspense psicológico que narra a história de Marion Crane, uma secretária que foge com dinheiro roubado e se hospeda no Bates Motel, administrado por Norman Bates. O filme é famoso por sua reviravolta chocante, cenas icônicas (como o chuveiro) e por redefinir os padrões do gênero de terror e suspense no cinema.

Esse processo de “inferência empática“ (como eu o chamaria) funciona para os humanos, em certo grau.

Mas mentes com emoções diferentes — mentes que sentem emoções que você nunca sentiu, ou que não conseguem sentir as emoções que você sentiria? Isso é algo que você não pode compreender ao colocar seu cérebro no lugar do outro cérebro. Posso pedir a você que imagine um alienígena que tenha crescido em um universo com quatro dimensões espaciais, em vez de três dimensões espaciais, mas você seria incapaz de reconfigurar seu córtex visual para ver como aquele alienígena vê. Posso tentar escrever uma história sobre alienígenas com emoções diferentes, mas você não conseguirá sentir essas emoções, nem eu.

Imagine um alienígena assistindo a um vídeo dos Irmãos Marx sem ter a menor ideia do que está acontecendo, sem compreender por que alguém buscaria ativamente tal experiência sensorial, já que o alienígena jamais concebeu nada remotamente parecido com senso de humor. Não tenha pena deles por terem perdido; você nunca experimentou.

Você poderia questionar: “Talvez os alienígenas tenham senso de humor, mas será que minhas piadas não são engraçadas o suficiente?” A ideia de usar um idioma estrangeiro em um país estrangeiro, falando alto e devagar, é comparável a isso,com base na teoria de que os falantes nativos teriam um fantasma interior que pode ouvir o significado que escorre de suas palavras, inerente em suas palavras, se você simplesmente as falar alto o suficiente para superar qualquer barreira peculiar em relação ao seu inglês perfeitamente sensato.

É importante perceber que o riso pode ser algo bonito e valioso, mesmo que não seja universalizável, mesmo que não seja compartilhado por todas as mentes possíveis. É a nossa contribuição especial para o futuro. Isso também pode ter algum significado.

É bom que tenha, porque a universalização é uma noção metaética que não posso proporcionar. A universalizabilidade entre humanos, talvez, mas não entre todas as mentes possíveis.

E quanto às mentes que não funcionam em arquiteturas emocionais como a sua, que não possuem equivalentes emocionais? Não se preocupe em explicar por que qualquer mente inteligente o suficiente para construir máquinas complexas teria inevitavelmente estados análogos às emoções. A seleção natural constrói máquinas complexas sem possuir emoções. Agora, há um Verdadeiro Alienígena para você: um processo de otimização que não funciona como o seu.

Grande parte do progresso na biologia desde a década de 1960 consistiu em tentar impedir a evolução antropomorfizadora. Essa foi uma grande batalha acadêmica, e não sei se a sanidade teria prevalecido se não fosse pelas evidências experimentais esmagadoras apoiadas por matemática clara. Fazer com que as pessoas deixem de se colocar no lugar de alienígenas é uma tarefa longa, difícil e árdua. Venho travando essa batalha na IA há anos.

Nosso antropomorfismo está profundamente enraizado em nós; não pode ser eliminado com um simples ato de vontade, uma determinação de dizer: “Agora vou parar de pensar como um humano!” A humanidade é o ar que respiramos; é o nosso molde, o papel branco em que começamos nossos esboços. E não nos consideramos humanos quando estamos sendo humanos.

É proverbial na ficção científica literária que o verdadeiro teste de um autor é sua capacidade de criar Verdadeiros Alienígenas. (E não apenas alienígenas convenientemente incompreensíveis que, por suas próprias razões misteriosas, fazem tudo o que a trama exige.) Jack Vance foi um dos grandes mestres dessa arte. Os humanos de Vance, se provenientes de uma cultura diferente, são mais alienígenas do que a maioria dos “alienígenas“. (Nunca leu Vance? Eu recomendaria começar com *City of the Chasch*. (A cidade de Chasch)) *The mote in God’s eye* (O cisco no olho de Deus), de Niven e Pournelle, também merece menção nesta discussão.

Em contraste - bem, certa vez li um autor de ficção científica (possivelmente Orson Scott Card) dizendo que o momento mais baixo da história da ficção científica na televisão foi um episódio de Jornada nas Estrelas. Nesse episódio, houve uma evolução paralela que resultou em alienígenas não apenas fisicamente semelhantes aos humanos, mas também fluentes em inglês. Além disso, esses alienígenas haviam recriado independentemente o preâmbulo da Constituição dos Estados Unidos, palavra por palavra.

Isso é a Grande Falha da Imaginação. Não pense que se restringe apenas à ficção científica ou à IA. A incapacidade de imaginar o alienígena é a incapacidade de nos vermos - a incapacidade de compreender a nossa própria singularidade. Quem pode perceber um humano camuflado entre outros humanos?

145 — Otimização e explosão de inteligência

Dentre os tópicos que não abordei aqui, está a noção de um processo de otimização. Em termos gerais, essa ideia se refere ao seu poder como mente, à sua capacidade de atingir alvos específicos em um amplo espaço de possibilidades, seja no contexto de futuros possíveis (planejamento) ou no contexto de projetos possíveis (invenção).

Suponhamos que você tenha um carro e que já saibamos que suas preferências incluem viagens. Agora, imagine que você pegue todas as partes do carro, ou até mesmo todos os átomos, e os misture aleatoriamente. É altamente improvável que você obtenha um artefato voltado para viagens, mesmo que seja apenas um carrinho de rodas, e muito menos um artefato de viagem que corresponda tão bem às suas preferências quanto o carro original. Portanto, em relação às suas preferências, o carro é um artefato extremamente improvável. O poder de um processo de otimização é que ele pode produzir esse tipo de improbabilidade.

Tanto a inteligência quanto a seleção natural podem ser consideradas casos específicos de otimização: processos que atingem, em um amplo espaço de possibilidades, alvos muito específicos determinados por preferências implícitas. A seleção natural favorece replicadores mais eficientes, enquanto as inteligências humanas possuem preferências mais complexas. Nem a evolução, nem os humanos possuem funções de utilidade consistentes, portanto, considerá-los como “processos de otimização“ é uma aproximação. Estamos tentando compreender a natureza do trabalho realizado, não afirmando que os humanos ou a evolução desempenham esse trabalho perfeitamente.

É assim que vejo a história da vida e da inteligência: como uma narrativa de bons projetos improváveis sendo produzidos por meio de processos de otimização. A “improbabilidade“ aqui se refere à improbabilidade relativa a uma seleção aleatória no espaço de projeto, e não à improbabilidade em termos absolutos - se há um processo de otimização envolvido, bons projetos “improváveis“ se tornam prováveis.

Ao examinarmos a história da otimização na Terra até o momento, o primeiro passo é fazer uma distinção conceitual entre o meta nível e o nível do objeto - separar a estrutura de otimização daquilo que está sendo otimizado.

Se considerarmos a biologia na ausência de hominídeos, então no nível do objeto teremos coisas como dinossauros, borboletas e gatos. No meta nível, teremos coisas como recombinação sexual e seleção natural de populações assexuadas. Você vai observar que o nível do objeto é um tanto mais complexo do que o meta nível. A seleção natural não é um assunto simples e envolve matemática. Contudo, se analisarmos a anatomia de um gato como um todo, veremos que o gato possui uma dinâmica muito mais complexa do que “mutação, recombinação, reprodução“.

Isso não é surpreendente. A seleção natural é um processo de otimização acidental que, basicamente, começou a ocorrer em algum lugar de uma poça de maré, em um dia qualquer. Um gato é o objeto de milhões de anos e bilhões de anos de evolução.

É claro que os gatos possuem cérebros que funcionam para aprender ao longo da vida; mas, ao final da vida de um gato, essa essa informação é descartada e não se acumula. Portanto, os efeitos cumulativos dos cérebros dos gatos como otimizadores são relativamente pequenos.

Considere também o cérebro de uma abelha ou de um castor. Uma abelha constrói colmeias e um castor constrói represas, mas eles não descobriram como construí-los a partir do zero. Um castor não conse-

gue descobrir como construir uma colmeia, assim como uma abelha não consegue descobrir como construir uma represa.

Dessa forma, os cérebros dos animais - até recentemente - não desempenhavam um papel principal no jogo da otimização planetária; eles eram peças, mas não os jogadores principais. Em comparação com a evolução, os cérebros careciam tanto de um poder de otimização generalizado (não podiam produzir a incrível variedade de artefatos gerados pela evolução) quanto de um poder de otimização cumulativo (seus produtos não acumulavam complexidade ao longo do tempo). Para mais informações sobre o tema, consulte "Reforço de proteínas e consequencialismo de DNA".

Muito recentemente, certos cérebros de animais começaram a exibir tanto um poder de otimização generalizado (produzindo uma ampla gama de artefatos em períodos de tempo curtos demais para que a seleção natural desempenhe um papel significativo) quanto um poder de otimização cumulativo (criando artefatos de complexidade crescente graças às habilidades transmitidas por meio da linguagem e da escrita).

A seleção natural leva centenas de gerações para produzir qualquer resultado e milhões de anos para projetos complexos completamente novos. Os programadores humanos podem projetar uma máquina complexa com cem elementos interdependentes em uma única tarde. Isso não é surpreendente, considerando que a seleção natural é um processo de otimização acidental que começou a ocorrer um dia, enquanto os humanos são otimizadores criados pela seleção natural ao longo de milhões de anos.

A maravilha da evolução não reside na sua eficácia, mas no fato de funcionar sem ser otimizada. Foi assim que a otimização se inicializou no universo — começando, como era de se esperar, a partir de um processo de otimização acidental extremamente ineficiente. Que não é o primeiro replicador acidental, entenda bem, mas o primeiro processo acidental de seleção natural. Faça a distinção entre o nível do objeto e o nível meta!

Desde o surgimento da otimização no universo, uma certa semelhança estrutural tem se mantido tanto na seleção natural quanto na inteligência humana…

A seleção natural seleciona os genes, mas, em geral, os genes não mudam e otimizam a seleção natural. A invenção da recombinação sexual é uma exceção a essa regra, assim como a invenção das células e do DNA. E você pode perceber tanto o poder quanto a raridade de tais eventos pelo fato de que os biólogos evolutivos estruturam histórias inteiras sobre a vida na Terra em torno deles.

No entanto, se você adotar uma perspectiva humana - pensando como um programador - verá que a seleção natural ainda não é tão complicada. Vamos tentar agrupar genes diferentes? Vamos tentar separar o armazenamento de informações do maquinário em movimento? Vamos tentar recombinar aleatoriamente grupos de genes? Em uma escala absoluta, essas são as ideias brilhantes que qualquer hacker inteligente teria nos primeiros dez minutos pensando em arquiteturas de sistemas.

Porque a seleção natural começou tão ineficiente (como um processo completamente acidental), este pequeno punhado de melhorias de meta nível que se alimentam dos replicadores — nem de perto tão complicadas quanto a estrutura de um gato — estruturam as épocas evolutivas da vida na Terra.

E após tudo isso, a seleção natural ainda é um deus cego e idiota. Conjuntos de genes podem evoluir para a extinção, apesar de todas as células e do sexo.

Agora, a seleção natural se alimenta a si mesma no sentido de que cada nova adaptação abre caminho para novas adaptações; mas isso ocorre no nível do objeto. O pool genético se beneficia de sua própria complexidade, mas apenas graças ao intérprete protegido da seleção natural, que opera em segundo plano e não é reescrito ou alterado pela evolução das espécies.

Da mesma forma, os seres humanos inventam ciências e tecnologias, mas ainda não começamos a reescrever a estrutura protegida do próprio cérebro humano. Temos um córtex pré-frontal, um córtex temporal e um cerebelo, assim como os primeiros inventores da agricultura. Ainda não começamos a nos modificar geneticamente. No nível do objeto, a ciência se alimenta da ciência, e cada nova descoberta abre caminho para novas descobertas — mas tudo isso acontece com um intérprete protegido, o cérebro humano, funcionando inalterado no fundo.

Temos invenções de meta nível, como a ciência, que tentam instruir os humanos sobre como pensar. No entanto, a pessoa que inventou o Teorema de Bayes não se tornou automaticamente um bayesiano;. ela não pudesse reescrever geneticamente, pois não possui esse conhecimento nem esse poder. Nossas descobertas significativas na arte de pensar, como a escrita e a ciência, são tão poderosas que moldam o curso da história humana, mas elas não rivalizam com o próprio cérebro em complexidade, e seu efeito sobre o cérebro é comparativamente superficial.

O estado atual da arte no treinamento da racionalidade não é suficiente para transformar um mortal arbitrariamente selecionado em Albert Einstein, que mostra o poder das pequenas peculiaridades genéticas do projeto do cérebro em comparação com todos os livros de autoajuda já escritos no século XX.

Como o cérebro trabalha invisivelmente em segundo plano, as pessoas tendem a ignorar sua contribuição e considerá-la como algo garantido; e falam como se a simples instrução de "Testar ideias por meio de experimentos" ou a regra de significância $p < 0,05$, tivessem a mesma ordem de contribuição de um cérebro humano completo. Tente dizer aos chimpanzés para testarem suas ideias por meio de experimentos e veja o até onde você chega.

Agora… alguns de nós aspiramos a projetar, de forma inteligente, uma inteligência capaz de se reprojetar de forma inteligente, até o nível de seu código-fonte.

Inicialmente, o código-fonte e, posteriormente, as leis da física seriam espécies de níveis protegidos. No entanto, esse "nível protegido" não conteria a dinâmica da otimização; os níveis protegidos não estruturariam o trabalho. O cérebro humano realiza um pouco de otimização por conta própria e acaba estragando tudo, independentemente do que se tente ensinar na escola. Mas esse otimizador recursivo totalmente envolvente não teria nenhum nível protegido que estivesse otimizando. Toda a estrutura de otimização estaria sujeita à própria otimização.

E isso representa uma mudança radical que rompe com todo o passado desde o primeiro replicador, porque desafia a noção de um meta nível protegido.

A história da Terra até agora tem sido uma história de otimizadores girando suas rodas a uma taxa constante, gerando uma pressão constante de otimização. E criando produtos otimizados não em uma taxa constante, mas em uma taxa acelerada, devido a como inovações ao nível de objeto abrem o caminho para outras inovações ao nível de objeto. Mas essa aceleração ocorre com um meta nível protegido fazendo a otimização real. Como uma busca que salta de uma ilha para outra no espaço de busca, onde as boas ilhas tendem a estar adjacentes a ilhas ainda melhores, mas o saltador não muda de perna. Ocasionalmente, pequenas mudanças conseguem retornar ao nível meta, como no caso do sexo ou da ciência, e então a história da otimização entra em uma nova era, e tudo avança mais rapidamente a partir daí.

Imagine uma economia sem investimento, ou uma universidade sem linguagem, uma tecnologia sem ferramentas para fabricar ferramentas. Uma vez em cem milhões de anos, ou uma vez em alguns séculos, alguém inventa um martelo.

É assim que tem sido a otimização na Terra até agora.

Quando olho para a história da Terra, não vejo uma história de otimização ao longo do tempo. Vejo uma história de poder de otimização inserido, e produtos otimizados obtidos. Até agora, graças à existência de meta-níveis quase totalmente protegidos, era possível dividir o histórico de otimização em épocas e representar graficamente a otimização cumulativa ao nível de objeto ao longo do tempo, porque o nível protegido estava operando em segundo plano e não mudava durante uma época.

O que acontece quando criamos uma IA totalmente envolvente e recursivamente autoaperfeiçoada? Então pegamos o gráfico de "otimização para dentro, otimizado para fora" e o dobramos sobre si mesmo. Metaforicamente falando.

Se a IA for fraca, ela não fará nada, porque não será suficientemente poderosa para melhorar significativamente a si mesma — seria o equivalente a dizer a um chimpanzé para reescrever seu próprio cérebro.

Se a IA for poderosa o suficiente para se reescrever de forma a aumentar sua capacidade de realizar melhorias adicionais e isso chegar até a compreensão total de seu próprio código-fonte e de seu projeto

como um otimizador… então, mesmo que o gráfico de “poder de otimização“ e “produção de produtos otimizados“ pareça essencialmente o mesmo, o gráfico de otimização ao longo do tempo será completamente diferente da história da Terra até agora.

Pessoas levantam frequentemente a seguinte questão: “Mas e se forem necessárias quantidades exponencialmente maiores de autorreescrita para obter apenas uma melhoria linear?“ Para isso, a resposta óbvia é: “A seleção natural exerceu um poder de otimização quase constante na linhagem dos hominídeos durante a história humana, e isso não parece ter exigido exponencialmente mais tempo para cada incremento linear de melhoria.“

Tudo isso é mero raciocínio analógico. Uma Inteligência Artificial Geral completa, que reflete sobre a natureza da otimização, conduz sua própria pesquisa em IA e reescreve seu próprio código-fonte, não se assemelha verdadeiramente a um gráfico da história da Terra dobrado sobre si mesmo. É uma criatura diferente. Essas analogias são, no máximo, úteis para previsões qualitativas, e ainda tenho muitas outras crenças que ainda não expliquei, que me guiam na escolha das analogias a fazer, entre outras coisas.

Mas se você está curioso sobre minha relutância em expandir o gráfico de crescimento biológico e econômico ao longo do tempo, considerando o horizonte futuro de uma IA com capacidade de pensar na velocidade dos transistores, criar nanofábricas moleculares autorreplicantes e melhorar seu código-fonte, então esse é o meu motivo: você está desenhando o gráfico errado, em vez de relacionar o produto otimizado com o tempo, o gráfico deveria representar o poder de otimização em relação ao produto otimizado.

146 — Fantasmas na máquina

As pessoas ouvem falar sobre IA amigável e geralmente têm uma das três principais reações iniciais:

"Ah, você pode tentar dizer à IA para ser amigável, mas se a IA puder modificar seu próprio código-fonte, ela simplesmente removerá quaisquer restrições que você tentar impor a ela."

Mas de onde vem essa decisão?

Ela surge de fora da causalidade, em vez de ser um efeito de uma cadeia lógica de causas que começa com o código-fonte originalmente escrito? A IA é a fonte final de seu próprio livre arbítrio?

Uma IA amigável não é uma IA egoísta limitada por um módulo especial de consciência extra que anula os impulsos naturais da IA e dita o que ela deve fazer. Você apenas constrói a consciência, e essa é a IA. Se você tem um programa que calcula qual decisão a IA deve tomar, pronto. O jogo acaba.

Neste ponto, gostaria de citar alguns estudos de caso do site *Computer Stupidities* (Burrices computacionais) e do subfórum de programação. (Não estou incluindo um link aqui porque é uma grande armadilha de tempo, mas você pode pesquisar no Google se estiver interessado.)

Já ensinei estudantes universitários que estavam fazendo um curso de programação de computadores. Alguns deles não entendiam que os computadores não são seres conscientes. Mais de uma pessoa usou comentários em seus programas Pascal para dar explicações detalhadas como: "Agora, preciso que você coloque essas letras na tela". Perguntei a um deles qual era o problema com esses comentários. A resposta foi: "De que outra forma o computador entenderia o que quero que ele faça?" Aparentemente, eles presumiam que, como não conseguiam entender Pascal, o computador também não conseguiria.

Quando estava na faculdade, costumava dar aulas particulares no laboratório de matemática da escola. Um aluno veio até mim porque seu programa básico não estava funcionando. Ele estava fazendo um curso para iniciantes e sua tarefa era escrever um programa que calculasse a receita de biscoitos de aveia com base no número de pessoas para as quais você estava cozinhando. Olhei para o programa dele e era mais ou menos assim:

```
10  Pré-aqueça a 350 graus
20 Junte todos os ingredientes em uma tigela grande
30 Misture até ficar homogêneo
```

Uma vez, um estudante de programação introdutória me pediu para analisar seu programa e descobrir por que ele sempre produzia zero como resultado de um cálculo simples. Olhei para o programa e a resposta era bastante óbvia:

```
begin
read("Número de Maçãs", maçãs)
read("Número de Cenouras", cenouras)
read("Preço de 1 Maçã", a_price)
read("Preço de 1 Cenoura", c_price)
write("Total de Maçãs", a_total)
write("Total para Cenouras", c_total)
write("Total", total)
total= a_total + c_total
a_total= maçãs * a_price
c_total= cenouras * c_price
end
```

Eu: "Bem, o seu programa não pode imprimir resultados corretos antes de eles serem calculados."

Ele: "Mas é lógico qual é a solução certa, e o computador deve reordenar as instruções corretamente."

Existe uma maneira instintiva de imaginar o cenário de "programar uma IA". Ele é mapeado em um esforço humano de aparência semelhante. É como se o "programa" estivesse dando instruções a um pequeno fantasma que reside na máquina, que então examina essas instruções e decide se gosta delas ou não.

Não há um fantasma que examine as instruções e decida como segui-las. O próprio programa é a IA.

Isso não significa que o fantasma faça tudo o que você deseja, como se fosse um gênio. Também não significa que o fantasma faça tudo exatamente do jeito que você quer, como se fosse um escravo extremamente obediente. Significa apenas que sua instrução é o único fantasma que está lá, pelo menos no momento da inicialização.

A construção de uma IA é muito mais difícil do que as pessoas imaginam intuitivamente, justamente porque não é possível simplesmente dizer ao fantasma o que fazer. É necessário construir o fantasma do zero, e tudo o que parece óbvio para você, o fantasma não perceberá, a menos que você saiba como fazê-lo entender. Não é suficiente apenas dizer ao fantasma para entender algo, é preciso criar a capacidade de compreensão a partir do zero.

Se você não sabe como construir algo que possui elementos complexos e inexplicáveis, como a "tomada de decisões", não pode simplesmente ignorar essa questão e contar com o livre arbítrio do fantasma para resolver tudo. Você ficará impotente e sem fantasmas.

Construir um programa de xadrez é muito mais do que desenvolver um processador de alta velocidade para a IA ser verdadeiramente inteligente e, em seguida, digitar no prompt de comando "Faça os movimentos de xadrez que achar melhor". Pode-se pensar que, como os próprios programadores não são jogadores de xadrez muito bons, qualquer conselho que tentem dar ao supercérebro eletrônico apenas deixaria o fantasma mais lento. Porém, não existe um fantasma. Você vê o problema.

Não há um feitiço simples que possa ser executado para invocar magicamente um fantasma completo na máquina. Você não pode simplesmente dizer: "Eu invoquei o fantasma e ele apareceu; isso é causa e efeito para você." (Também não funciona se você usar a noção de "emergência" ou "complexidade" como um subs-

tituto para “invocar“). Você não pode dar uma instrução à CPU dizendo: “Seja um bom jogador de xadrez!” É necessário entender o mistério dos pensamentos do jogador de xadrez e estruturar todo o fantasma do zero.

Nenhum resultado ou ação ocorrerá dentro do fantasma, independentemente de quão lógico, óbvio, correto, autoevidente ou inteligente possa parecer para você. Isso porque tudo o que acontece dentro do programa é o resultado de uma sequência de causa e efeito originada nas instruções que você determinou, juntamente com quaisquer dependências causais de dados sensoriais que você tenha incorporado nas instruções iniciais.

Isso não significa que você programe cada decisão explicitamente. O Deep Blue, por exemplo, era um jogador de xadrez muito melhor do que seus programadores. O Deep Blue fazia movimentos de xadrez superiores a qualquer coisa que seus criadores pudessem ter programado explicitamente, mas isso não acontecia porque os programadores simplesmente ignoravam o assunto e deixavam tudo por conta do programa. O Deep Blue jogava melhor do que seus programadores… no final de uma cadeia de causa e efeito que se originava no código dos programadores e seguia um caminho lógico a partir daí. Nada acontecia apenas porque era um movimento obviamente bom e o livre arbítrio fantasmagórico do Deep Blue assumiu, sem o envolvimento do código e suas consequências lógicas.

Se você tentar se esquivar de limitar a IA, não terá um fantasma livre como um escravo emancipado. Você fica com um monte de areia que ninguém purificou em silício, moldou em uma CPU e programou para pensar.

Vá em frente, tente dizer a um chip de computador “Faça o que quiser!“ Quer saber o que acontece? Nada. Basta um único passo que seja tão óbvio, tão lógico, tão autoevidente que sua mente simplesmente o ignore e você se desvie do caminho do programador de IA. É necessário um esforço semelhante ao que descrevo em “Agarrando Coisas Escorregadias“ para evitar que sua mente faça isso.

147 — Adição artificial

Suponhamos que os seres humanos não tivessem absolutamente nenhuma ideia de como executar operações aritméticas. Imaginem se os seres humanos tivessem evoluído, em vez de terem aprendido a habilidade inata de contar e somar ovelhas. As pessoas que usassem essa habilidade inata não teriam ideia de como ela funcionava, da mesma forma que Aristóteles não tinha ideia de como seu córtex visual sustentava sua capacidade de enxergar. A aritmética de Peano, tal como a conhecemos, não teria sido inventada.

Existem filósofos trabalhando para formalizar as intuições numéricas, mas eles usam notações como

```
Soma-De (Sete, Seis) = Treze
```

para formalizar o fato intuitivamente óbvio de que, ao adicionar "sete" e "seis", obtemos "treze".

Nesse mundo, as calculadoras de bolso funcionam armazenando uma enorme tabela de fatos aritméticos, inseridos manualmente por uma equipe de especialistas em aritmética artificial, para valores iniciais que variam de zero a cem. Embora essas calculadoras possam ser úteis em um sentido pragmático, muitos filósofos argumentam que elas estão apenas simulando a adição, ao invés de realmente realizá-la. Nenhuma máquina pode contar de verdade — é por isso que os humanos precisam contar treze ovelhas antes de digitar "treze" na calculadora. As calculadoras podem recitar os fatos armazenados, mas nunca saberão o significado dessas afirmações. Se você digitar "duzentos mais duzentos", a calculadora dirá "Erro: Fora do alcance", quando intuitivamente é óbvio, se você souber o que as palavras significam, que a resposta é "quatrocentos".

Alguns filósofos, é claro, não são tão ingênuos a ponto de se deixarem enganar por essas intuições. Os números são de fato um sistema puramente formal — o rótulo "trinta e sete" tem significado não devido a qualquer propriedade intrínseca das palavras em si, mas porque o rótulo se refere a trinta e sete ovelhas no mundo externo. Um número adquire essa propriedade referencial por meio de sua rede semântica de relações com outros números. É por isso que, em softwares, o token LISP[21] para "trinta e sete" não precisa de nenhuma estrutura interna — ele é significativo apenas por sua referência e relação, não por alguma propriedade computacional específica de "trinta e sete" em si.

Ninguém jamais desenvolveu uma Aritmética Geral Artificial, embora, é claro, existam muitas Aritméticas Artificiais estreitas e específicas para determinados domínios que trabalham com números entre "vinte" e "trinta", e assim por diante. E se observarmos o quão lento tem sido o progresso nos números na faixa de "duzentos", fica claro que não teremos uma Aritmética Geral Artificial tão cedo. Os principais especialistas no campo estimam que levará pelo menos cem anos até que as calculadoras possam somar tão bem quanto uma criança humana de doze anos.

Mas nem todos concordam com essa estimativa ou com crenças meramente convencionais sobre Aritmética Artificial. É comum ouvir afirmações como estas:

---

21 NT. Em **LISP**, um **token** é a menor unidade significativa de código, como um símbolo, número, palavra-chave ou operador, que é reconhecido pelo interpretador ou compilador da linguagem. Esses tokens são gerados durante a fase de análise léxica e servem como blocos básicos para a construção de expressões e programas em LISP, uma linguagem de programação conhecida por sua sintaxe baseada em parênteses e forte ênfase em manipulação simbólica.

- “É uma questão de enquadramento — o que é igual a ‘vinte e um mais’ depende se é ‘mais três’ ou ‘mais quatro’. Se conseguirmos armazenar fatos aritméticos suficientes para cobrir as verdades do senso comum que todos conhecem, começaremos a ver uma adição real na rede.“
- “Mas você nunca conseguirá programar tantos fatos aritméticos contratando especialistas para inseri-los manualmente. O que precisamos é de uma Aritmética Artificial capaz de aprender a vasta rede de relações entre os números que os humanos adquirem durante a infância, observando conjuntos de maçãs.“
- “Não, o que realmente precisamos é de uma Aritmética Artificial capaz de compreender a linguagem natural, de modo que, em vez de ser informada explicitamente de que ‘vinte e um mais dezesseis’ é igual a ‘trinta e sete’, ela possa obter o conhecimento explorando a Web.“
- “Francamente, parece-me que vocês estão apenas tentando se convencer de que podem resolver o problema. Nenhum de vocês sabe realmente o que é aritmética, então vocês estão se debatendo com esses tipos de argumentos genéricos. ‘Precisamos de um AA que possa aprender X,’ ‘Precisamos de um AA que possa extrair X da Internet.’ Quero dizer, parece bom, parece que você está progredindo e é bom até para as relações-públicas, porque todos pensam que entendem a solução proposta — mas isso realmente não o aproxima da adição geral, em oposição à adição de domínio específico. Provavelmente nunca saberemos a natureza fundamental da aritmética. O problema é muito difícil para os humanos resolverem.“
- “É por isso que precisamos desenvolver uma aritmética geral da mesma forma que a Natureza fez — por meio da evolução.“
- “As abordagens de cima para baixo falharam claramente em produzir aritmética. Precisamos de uma abordagem de baixo para cima, de uma forma de fazer a aritmética emergir. Temos que reconhecer a imprevisibilidade fundamental dos sistemas complexos.“
- “Vocês todos estão equivocados. Os esforços anteriores para criar aritmética de máquina foram inúteis desde o início, pois simplesmente não possuíam poder computacional suficiente. Se considerarmos o número de trilhões de sinapses presentes no cérebro humano, fica claro que as calculadoras não possuem tabelas de pesquisa nem próximas a essa magnitude. Precisamos de calculadoras tão poderosas quanto o cérebro humano. Segundo a Lei de Moore, isso ocorrerá em 27 de abril de 2031, entre 4h e 4h30 da manhã.“
- “Acredito que a aritmética de máquina será desenvolvida quando os pesquisadores digitalizarem cada neurônio de um cérebro humano completo em um computador, de modo que simulemos o circuito biológico responsável pela adição em seres humanos.“
- “Acho que não precisamos esperar para escanear um cérebro inteiro. As redes neurais são semelhantes ao cérebro humano, e podemos treiná-las para realizar tarefas sem entendermos como elas as realizam. Criaremos programas capazes de realizar aritmética sem que nós, seus criadores, compreendamos como eles a fazem.“
- “No entanto, o Teorema de Gödel demonstra que nenhum sistema formal pode capturar as propriedades básicas da aritmética. A física clássica é formalizável, portanto, para somar dois mais dois, o cérebro deve recorrer à física quântica.“
- “Ei, se a aritmética humana fosse simples o suficiente para ser reproduzida em um computador, não conseguiríamos contar alto o suficiente para construir computadores.“
- “Você nunca ouviu falar do Experimento da Calculadora Chinesa de John Searle? Mesmo se você tivesse um conjunto enorme de regras que permitisse adicionar ‘vinte e um’ e ‘dezesseis’, imagine traduzir todas as palavras para o chinês e verá que não há uma adição genuína acontecendo. Não há números reais em nenhum lugar do sistema, apenas rótulos que os humanos usam para números…“

Há mais de uma lição nesta parábola, e eu a contei com diferentes lições em diferentes contextos. Isso ilustra a ideia de níveis de organização, por exemplo — uma CPU pode adicionar dois números grandes porque os números não são objetos opacos de uma caixa preta, mas sim estruturas ordenadas de 32 bits.

Mas para fins de superação vieses, estabeleçamos dois princípios morais:

- Primeiramente, há o perigo de acreditar em afirmações que você não pode compreender por conta própria.
- Em segundo lugar, existe o perigo de tentar contornar confusões básicas.

Para evitar qualquer acusação de generalização baseada em evidências fictícias, ambas as lições po-

dem ser extraídas da história real da Inteligência Artificial. O primeiro perigo está relacionado ao problema de nível de objeto enfrentado pelos dispositivos AA: eles funcionavam como gravadores que reproduziam "conhecimento" gerado externamente, utilizando um processo que não podiam internalizar. Um humano poderia dizer ao dispositivo AA que "vinte e um mais dezesseis é igual a trinta e sete", o dispositivo AA poderia gravar essa frase e reproduzi-la, ou até mesmo combinar o padrão "vinte e um mais dezesseis" para obter "trinta e sete!" — Porém, os dispositivos AA não conseguiam gerar esse conhecimento por si mesmos.

Isso é muito semelhante a acreditar em um físico que te diz "Luz é composta por ondas", simplesmente registrando essas palavras fascinantes e as reproduzindo quando alguém perguntar: "Do que a luz é feita?", sem possuir a capacidade de gerar esse conhecimento por si mesmo.

A segunda lição moral refere-se ao perigo do meta nível que consumiu pesquisadores de Aritmética Artificial e espectadores opinativos — o perigo de contornar as lacunas confusas em seu conhecimento. Há uma tendência de fazer qualquer coisa, exceto cerrar os dentes, trabalhar intensamente e preencher essas lacunas.

Quando você diz "É algo emergente!" ou quando diz "É algo que não pode ser conhecido!", em ambos os casos você está evitando reconhecer que há uma percepção básica necessária, a qual você não possui.

Como você pode saber quando terá um novo insight fundamental? Não há outra maneira a não ser enfrentar o problema de frente, aprender tudo o que puder sobre ele, estudá-lo de todos os ângulos possíveis, talvez por anos a fio. Essa não é uma atividade que o meio acadêmico esteja pronto para permitir, especialmente quando você precisa publicar pelo menos um artigo por mês. Certamente, não é nada que os investidores de risco estejam dispostos a financiar. Você deve prosseguir e construir o sistema agora ou desistir e buscar outras opções.

Observe os comentários anteriores: nenhum deles tem a intenção de iniciar uma busca pelo insight que falta, que poderia desvendar o mistério dos números e tornar "vinte e sete" mais do que uma caixa preta. Nenhum dos comentaristas percebeu que suas dificuldades surgiam da ignorância ou confusão em suas próprias mentes, e não de uma propriedade intrínseca da aritmética. Eles não estavam buscando alcançar um estado no qual o que era confuso se tornasse claro.

Se você ler o livro *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Raciocínio Probabilístico em Sistemas Inteligentes: Redes de Inferência Plausível) [1], de Judea Pearl, você perceberá que o insight fundamental por trás dos modelos gráficos é indispensável para resolver os problemas que demandam sua aplicação. (Infelizmente, essa não é uma ideia que possa ser resumida em uma camiseta; portanto, será necessário ler o livro por conta própria). Até o momento, não encontrei nenhuma divulgação online das redes bayesianas que transmita adequadamente as razões subjacentes aos princípios, nem a importância de a matemática ser exatamente como é, mas o livro de Pearl é extraordinário.) Houve uma época em que dezenas de "lógicas não monotônicas" tentavam, de maneira desajeitada, capturar intuições como "Se o alarme de roubo disparar, provavelmente houve um ladrão, mas se eu descobrir que houve um pequeno terremoto perto da minha casa, provavelmente não houve um ladrão." Com o insight proporcionado pelos modelos gráficos, é possível fornecer uma explicação matemática precisa sobre porque a lógica de primeira ordem não é adequada para esse tipo de trabalho e expressar a solução correta de maneira concisa, capturando todos os detalhes do senso comum de uma só vez. Enquanto você não tiver esse insight, continuará a ajustar a lógica aqui e ali, adicionando cada vez mais "gambiarras" para forçá-la a se adequar a tudo o que parece "obviamente verdadeiro".

Você não saberá que o problema da Aritmética Artificial é insolúvel sem essa chave. Se você não conhece as regras, não conhece a regra que diz que é necessário conhecer as regras para fazer qualquer coisa. E assim, surgirão todo tipo de ideias inteligentes que parecem funcionar, como construir um sistema de Aritmética Artificial capaz de ler linguagem natural e baixar milhões de afirmações aritméticas da internet.

E ainda assim, de alguma forma, essas ideias inteligentes nunca funcionam. Sempre ocorre que você "não vê nenhuma razão para não funcionar" porque está ignorando as barreiras, não porque elas não existam. É como atirar em um alvo distante com os olhos vendados — você pode disparar tiros às cegas, gritando: "Você não pode provar que não vou acertar o centro!" Mas, até tirar a venda, você nem mesmo está participando do jogo de mira. Quando "ninguém pode provar para você" que sua ideia valiosa está incorreta, signifi-

ca que você não tem informações suficientes para atingir um alvo pequeno em um vasto espaço de soluções. Até você saber que sua ideia funcionará, ela não funcionará.

Com base na história dos insights fundamentais anteriores em Inteligência Artificial e nas grandes confusões propostas antes desses insights, deduzo uma lição importante da vida real: quando o problema central é a sua própria ignorância, estratégias inteligentes para contorná-la resultam em atirar no próprio pé.

Referências

[1] Pearl, Probabilistic Reasoning in Intelligent Systems.

148 — Valores terminais e valores instrumentais

Num nível puramente instintivo, qualquer planejador humano age como se fizesse uma distinção entre meios e fins. Quer chocolate? Há chocolate no supermercado Publix. Você pode chegar ao supermercado dirigindo uma milha ao sul na Av. Washington. Você pode dirigir se entrar no carro. Você pode entrar no carro se abrir a porta. Você pode abrir a porta se tiver as chaves do carro. Assim, você coloca as chaves do carro no bolso e se prepara para sair de casa…

… quando, de repente, chega a notícia no rádio de que um terremoto destruiu todo o chocolate no Publix local. Bem, não adianta ir ao Publix se não houver chocolate lá, e não adianta entrar no carro se você não estiver indo a lugar nenhum, e não adianta ter as chaves do carro no bolso se você não for dirigir. Então, você tira as chaves do carro do bolso, liga para a pizzaria local e pede uma pizza de chocolate. Hum, delicioso.

Raramente percebo que as pessoas perdem o controle dos planos que elas mesmas conceberam. As pessoas geralmente não vão ao supermercado se souberem que não há chocolate. Mas também notei que, quando as pessoas começam a falar explicitamente sobre sistemas de metas em vez de apenas desejar coisas, mencionando "metas" em vez de simplesmente utilizá-las, elas ficam geralmente confusas. Os seres humanos são especialistas em planejamento, não especialistas em teoria do planejamento, caso contrário, haveria muitos mais desenvolvedores de IA no mundo.

Especificamente, observei que as pessoas ficam confusas quando, em discussões filosóficas abstratas, e não no cotidiano, consideram a distinção entre meios e fins; mais formalmente, entre "valores instrumentais" e "valores terminais".

Parte do problema, a meu ver, é que a mente humana usa um sistema bastante improvisado para acompanhar seus objetivos — funciona, mas não de maneira clara. O inglês não incorpora uma distinção nítida entre meios e fins: *I want to save my sister's life* (Quero salvar a vida da minha irmã) e *I want to administer penicillin to my sister* (Quero administrar penicilina à minha irmã) usam a mesma palavra *want* (querer).

Podemos descrever, em mero inglês, a distinção que está sendo perdida?

Como primeira tentativa:

"Valores instrumentais" são desejáveis estritamente com base em suas consequências antecipadas. "Quero administrar penicilina à minha irmã", não porque uma irmã cheia de penicilina seja um bem intrínseco, mas na expectativa de que a penicilina cure sua pneumonia comedora de carne. Se, ao contrário, você previsse que injetar penicilina a transformaria em uma poça como a Bruxa Malvada do Oeste, você lutaria com a mesma intensidade para mantê-la longe da penicilina.

"Valores terminais" são desejáveis sem condicionar a outras consequências: "Quero salvar a vida da minha irmã" não tem nada a ver com a antecipação de que ela receberá uma injeção de penicilina depois disso.

Essa primeira tentativa possui falhas óbvias. Se salvar a vida da minha irmã fizesse com que a Terra fosse engolida por um buraco negro, eu sairia e choraria por um tempo, mas não administraria penicilina. Isso significa que salvar a vida da minha irmã não era um valor "terminal" ou "intrínseco", por que teoricamente está condicionado às suas consequências? Estou tentando salvar a vida dela apenas porque acredito que um buraco negro não consumirá a Terra? O senso comum nos diz que isso não é o que está acontecendo.

Então, esqueça o inglês. Podemos criar uma descrição matemática de um sistema de decisões no qual os valores terminais e instrumentais sejam tipos separados e incompatíveis — como números inteiros e números de ponto flutuante em uma linguagem de programação, sem conversão automática entre eles.

Um sistema de decisão bayesiano ideal pode ser configurado usando apenas quatro elementos:

- `Resultados: digite Resultado[ ]`
  - lista de resultados possíveis:
  - {irmã vive, irmã morre}
- `Ações: digite Ação[ ]`
  - lista de ações possíveis:
  - {administrar penicilina, não administrar penicilina}
- `Função_utilidade: digite Resultado → Utilidade`
  - função de utilidade que mapeia cada resultado em uma utilidade
  - (uma utilidade sendo representável como um número real entre infinito negativo e positivo)
  - $\left\{ \begin{array}{l} \text{irmã vive} \mapsto 1 \\ \text{irmã morre} \mapsto 0 \end{array} \right\}$

- `Função_probabilidade_condicional:`

`digite Ação -> (Resultado -> Probabilidade)`

  - função de probabilidade condicional que mapeia cada ação em uma distribuição de probabilidade sobre os resultados
  - (uma probabilidade sendo representável como um número real entre 0 e 1)

$$\left\{ \begin{array}{ll} \text{administrar penicilina} & \longmapsto \begin{pmatrix} \text{irmã vive} \longmapsto 0,9 \\ \text{irmã morre} \longmapsto 0,1 \end{pmatrix} \\ \text{não administrar penicilina} & \longmapsto \begin{pmatrix} \text{irmã vive} \longmapsto 0,3 \\ \text{irmã morre} \longmapsto 0,7 \end{pmatrix} \end{array} \right\}$$

Se você não conseguir ler o sistema de tipos diretamente, não se preocupe, eu sempre traduzirei para o inglês. Para os programadores, ver o sistema descrito em declarações distintas ajuda a definir objetos mentais distintos.

E quanto ao próprio sistema de tomada de decisões?

- `Utilidade_Esperada: Ação A →`

`(Soma O em Resultados: Utilidade(O) * Probabilidade(O|A)`

  - A "utilidade esperada" de uma ação é igual à soma, para todos os resultados, da utilidade desse resultado multiplicada pela probabilidade condicional desse resultado, dado essa ação.

$$\left\{ \begin{array}{ll} \text{UE}(\text{administrar penicilina}) & = 0.9 \\ \text{UE}(\text{não administrar penicilina}) & = 0.3 \end{array} \right\}$$

- Escolha:

— > (Argmax A em Ações: Utilidade Esperada(A)

  - Escolha uma ação cuja “utilidade esperada“ seja máxima.
  - {retorno: administrar penicilina}

Para cada ação, calcule a probabilidade condicional de todas as consequências que podem ocorrer e some as utilidades dessas consequências multiplicadas por sua probabilidade condicional. Em seguida, escolha a melhor ação.

Este é um esboço matematicamente simples de um sistema de tomada de decisões. Não é uma forma eficiente de computar decisões no mundo real.

E se, por exemplo, você precisar de uma sequência de ações para executar um plano? O formalismo pode facilmente representar isso ao permitir que cada Ação represente uma sequência completa. No entanto, isso cria um espaço exponencialmente grande, assim como o espaço de todas as frases que você pode digitar em 100 letras. Para dar um exemplo simples, se uma das ações possíveis no primeiro turno for ‘Atire no meu próprio pé’, um planejador humano concluirá que essa é uma má ideia em geral e descartará todas as sequências que começam com essa ação. No entanto, nós simplificamos essa estrutura em nossa representação. Não temos sequências de ações, apenas ações isoladas.

Portanto, sim, existem algumas complicações menores. Obviamente, caso contrário, simplesmente iríamos correndo e construiríamos uma IA real dessa maneira. Nesse sentido, isso se assemelha muito à própria teoria da probabilidade bayesiana.

Mas este é um daqueles momentos em que é surpreendentemente útil considerar a versão absurdamente simples antes de adicionar quaisquer complicações pomposas.

Imagine o filósofo que afirma: “Todos somos egoístas; só nos preocupamos com nossos próprios estados mentais.” A mãe que alega se importar com o bem-estar de seu filho, na verdade, quer acreditar que seu filho está bem - essa crença é o que a faz feliz. Ela o ajuda em prol de sua própria felicidade, não da dele.“ Você diz: “Bem, suponha que a mãe sacrifique sua vida para empurrar seu filho para longe de um caminhão que se aproxima. Isso não a fará feliz, apenas morta.“ O filósofo gagueja por alguns momentos e então responde: ‘Mas ela fez isso porque valorizou essa escolha acima das outras - pelo sentimento de importância que atribuiu a essa decisão.“

Então você diz,

TIPO DE ERRO: Nenhum construtor encontrado para Utilidade_Esperada → Utilidade.

Permita-me explicar essa resposta.

Mesmo nosso formalismo simples ilustra uma distinção clara entre utilidade esperada, o qual é algo que as ações possuem, e utilidade, sendo algo que os resultados têm. Certamente, é possível mapear utilidades e utilidades esperadas em números reais. No entanto, é como observar que é possível mapear a velocidade do vento e a temperatura em números reais. Isso não as tornam a mesma coisa.

O filósofo começa argumentando que todas as utilidades devem estar relacionadas a resultados que consistem em seu estado mental. Se isso fosse verdade, sua inteligência funcionaria como um motor para direcionar o futuro para áreas onde você se sentiria feliz. Os futuros seriam distintos apenas pelo seu estado mental; você seria indiferente entre dois futuros em que tivesse o mesmo estado mental.

E, na verdade, seria muito improvável que você sacrificasse sua própria vida para salvar outra.

Quando contestamos que as pessoas às vezes sacrificam suas vidas, a resposta do filósofo passa a discutir as utilidades esperadas sobre as ações: ‘O sentimento de importância que ela atribuiu a essa decisão’. Isso é um salto drástico que deveria nos deixar perplexos. Tentar converter uma Utilidade Esperada em uma Utilidade resultaria em um erro absoluto em nossa linguagem de programação. Mas em inglês, tudo soa igual.

As escolhas do nosso sistema de tomada de decisões simples são baseadas na maior Utilidade Esperada, mas isso não diz nada sobre para onde o futuro é direcionado. Não diz nada sobre as utilidades que o tomador de decisões atribui ou sobre quais resultados provavelmente ocorrerão no mundo real como resultado. Não diz nada sobre a função da mente como um motor.

A causa física de uma ação física é um estado cognitivo, em nosso decisor ideal, uma Utilidade Esperada, e essa utilidade esperada é calculada ao avaliar uma função de utilidade sobre as consequências imaginadas. Para salvar a vida de seu filho, você deve imaginar o evento no qual a vida de seu filho é salva, e essa imaginação não é o evento em si. É uma representação, como a diferença entre a palavra "neve" e a própria neve. No entanto, isso não significa que o que está entre aspas deva ser um estado cognitivo. Se você escolher a ação que leva ao futuro em que você representa a frase "meu filho continua vivo", então você funcionará como um motor para direcionar o futuro para uma região onde seu filho continua vivo. Não é um motor que direciona o futuro para uma região onde você representa a frase "meu filho continua vivo". Para direcionar o futuro para lá, sua função de utilidade teria que retornar uma alta utilidade quando fornecida com a representação "meu filho continua vivo", a citação da citação, sua imaginação de si mesmo imaginando. As receitas estragam o bolo quando você as tritura e as mistura na massa.

E é por isso que é útil considerar primeiro os sistemas de decisões simples. Se adicionarmos complicações demais ao sistema, as distinções que antes eram claras se tornam mais difíceis de enxergar.

Agora, analisemos algumas complicações. Claramente, a função Utilidade (mapeamento de Resultados em Utilidades) tem o objetivo de formalizar o que mencionei anteriormente como "valores terminais" — valores que não dependem de suas consequências. E o que dizer do cenário em que salvar a vida de sua irmã resulta na destruição da Terra por um buraco negro? Em nosso formalismo, descartamos essa possibilidade. Os Resultados não levam a outros Resultados, apenas as Ações levam a Resultados. Portanto, o cenário em que sua irmã se recupera de uma pneumonia seguida pela Terra sendo engolida por um buraco negro seria reduzido a um único "resultado possível".

E onde estão os "valores instrumentais" nesse formalismo simplificado? Na verdade, eles desapareceram completamente! Veja bem, nesse formalismo, as ações levam diretamente aos resultados, sem eventos intermediários. Não existe a noção de lançar uma pedra que voa e derruba uma maçã da árvore, fazendo-a cair no chão. Lançar a pedra é a Ação e leva diretamente ao Resultado da maçã caindo no chão, segundo a função de probabilidade condicional que converte uma Ação em uma distribuição de Probabilidade sobre os Resultados.

Para calcular de fato a função de probabilidade condicional e considerar separadamente a utilidade da pneumonia de uma irmã e de um buraco negro engolindo a Terra, precisaríamos representar a estrutura da rede de causalidade — como os eventos levam a outros eventos.

Nesse ponto, os valores instrumentais começariam a reaparecer. Se a rede causal fosse suficientemente regular, poderíamos identificar um estado B que tenderia a levar a C, independentemente de como B fosse alcançado. Assim, se você desejasse alcançar C por algum motivo, poderia planejar eficientemente, elaborando primeiro um estado B que levasse a C e, em seguida, uma ação que levasse a B. Esse seria o conceito de "valor instrumental" — B teria valor instrumental porque levaria a C. O próprio estado C pode ser avaliado terminalmente, ou seja, como um termo na função de utilidade em relação ao resultado global. Ou C pode ser apenas um valor instrumental, um nó que não foi diretamente avaliado pela função de utilidade.

O valor instrumental, nesse formalismo, é apenas uma ferramenta para o cálculo eficiente de planos. Ele pode e deve ser descartado sempre que essa regularidade não existir.

Suponha, por exemplo, que exista um valor particular de B que não leve a C. Você escolheria uma Ação que levasse a esse B? Ou não importa a filosofia abstrata: se você quisesse ir ao supermercado comprar chocolate e decidisse dirigir até lá, conseguiria entrar no carro arrancando a porta com uma pá a vapor? (Não.) O valor instrumental é uma "abstração com vazamento", como dizemos nós, programadores; às vezes, é necessário descartar o valor armazenado em cache e calcular a utilidade real esperada. Parte de ser eficiente sem ser suicida, é perceber quando os atalhos convenientes falham. Embora esse formalismo dê origem a valores instrumentais, ele só o faz quando há a regularidade necessária, e estritamente como um atalho conveniente no cálculo.

Mas se complicarmos o formalismo antes de entender a versão simples, podemos erroneamente acreditar que os valores instrumentais adquirem uma existência própria, até mesmo em um sentido normativo. Ao afirmamos que B é geralmente bom porque leva a C, estamos obrigados a sempre buscar B, mesmo na ausência de C. As pessoas cometem esse tipo de erro em discussões filosóficas abstratas, embora jamais o fizessem na vida real, usar uma pá a vapor para abrir a porta do carro. Podemos até questionar a possibilidade de desenvolver um consequencialista que maximize apenas a aptidão genética inclusiva, imaginando que ele passaria fome por não ter um valor terminal explícito para "comer comida". As pessoas cometem esse erro, apesar de nunca passarem o dia inteiro abrindo portas de carro por medo de ficarem presas do lado de fora caso não valorizem explicitamente essa ação.

Os valores instrumentais residem na (estrutura de rede da) função de probabilidade condicional. Isso torna o valor instrumental estritamente dependente das crenças de fato, considerando uma função de utilidade fixa. Se eu acreditar que a penicilina causa pneumonia e, que a ausência de penicilina cura a pneumonia, então meu valor instrumental percebido da penicilina passará de alto para baixo. Altere as crenças de fato — altere a função de probabilidade condicional que associa as ações às consequências em que se acredita — e os valores instrumentais mudarão em conjunto.

Em argumentos morais, algumas disputas são sobre consequências instrumentais, enquanto outras são sobre valores terminais. Se seu oponente no debate afirmar que proibir armas levará a uma menor criminalidade, e você afirmar que proibir armas levará a uma maior criminalidade, vocês concordam com um valor instrumental superior (o crime é ruim), mas discordam sobre quais eventos intermediários levam a quais consequências. No entanto, não acredito que uma discussão sobre a circuncisão feminina seja realmente uma discussão factual sobre a melhor forma de alcançar um valor compartilhado de tratar as mulheres com justiça ou fazê-las felizes.

Essa distinção importante é frequentemente negligenciada em discussões acaloradas. Pessoas com divergências factuais e valores compartilhados decidem que seus oponentes no debate devem ser sociopatas. Como se seus odiados inimigos, os defensores do controle de armas/direitos, realmente quisessem matar pessoas, o que seria improvável em uma perspectiva realista da psicologia.

Temo que o cérebro humano não faça uma distinção entre crenças morais terminais e crenças morais instrumentais. "Devemos banir as armas" e "Devemos salvar vidas" não parecem diferentes como crenças morais, da mesma forma que a visão difere do som. Apesar de todas as outras maneiras pelas quais o sistema de metas humanas complica tudo à vista, essa distinção acaba se desintegrando em uma mistura de coisas com valor condicional.

Para extrair os valores terminais, temos que examinar essa mistura de coisas valiosas e tentar descobrir quais estão obtendo seu valor de algum outro lugar. É um projeto difícil! Se você diz que deseja banir as armas para reduzir o crime, pode levar um tempo para perceber que "reduzir o crime" não é um valor terminal, mas sim um valor instrumental superior com vínculos com valores terminais como vidas humanas e felicidade humana. E então aqueles que defendem os direitos das armas podem ter vínculos com o valor instrumental superior de "reduzir o crime" mais um vínculo com um valor de "liberdade", que pode ser um valor terminal para eles, ou outro valor instrumental…

Não podemos listar toda a nossa rede de valores derivados de outros valores. Provavelmente, nem armazenamos todo o histórico de como os valores se desenvolveram. Ao considerar os dilemas morais corretos, como "Você faria X se Y", muitas vezes podemos descobrir de onde vieram nossos valores. No entanto, até mesmo esse projeto está repleto de armadilhas, dilemas enganosos e argumentos filosóficos confusos. Não sabemos quais são nossos próprios valores, de onde vieram, e não podemos descobrir, exceto por meio de projetos de arqueologia cognitiva propensos a erros. Apenas ter uma distinção consciente entre "valor terminal" e "valor instrumental", entender o que isso significa e usá-la corretamente é um trabalho árduo. Somente ao examinar o formalismo simples é que podemos ver como isso deveria ser fácil, em princípio.

E isso sem mencionar todas as outras complicações do sistema de recompensa humano — todo o uso da arquitetura de reforço e a maneira como comer chocolate é prazeroso, e antecipar comer chocolate também é prazeroso, mas são tipos diferentes de prazeres…

Mas não vou reclamar muito da confusão.

Ser ignorante em relação aos seus próprios valores nem sempre é divertido, mas pelo menos não é chato.

149 — Generalizações vazadas

As maçãs são boas para comer? Em geral, sim, mas algumas maçãs são podres.

Os seres humanos têm dez dedos? A maioria de nós tem, mas muitas pessoas perderam um dedo e ainda são consideradas humanas.

A menos que você vá para um nível de descrição muito abaixo de qualquer objeto macroscópico — abaixo de sociedades, indivíduos, dedos, tendões e ossos, células, chegando às partículas e campos onde as leis são verdadeiramente universais — praticamente todas as generalizações que você fizer no mundo real terão exceções.

(Embora possa haver, é claro, algumas exceções à regra acima…)

Na maioria das vezes, a maneira como lidamos com generalizações que têm exceções é simplesmente lidar com elas. Se uma loja de biscoitos geralmente fecha às 22h, exceto no Dia de Ação de Graças, quando fecha às 18h, e hoje é o Dia Nacional do Genocídio dos Nativos Americanos, é melhor você aparecer antes das 18h ou não conseguirá um biscoito.

Nossa capacidade de lidar com generalizações que têm exceções é oposta à nossa necessidade de fechamento, o grau em que queremos afirmar de uma vez por todas que os seres humanos têm dez dedos, e ficamos frustrados quando temos que tolerar uma ambiguidade contínua.

Aumentar as apostas pode aumentar a necessidade de fechamento desligando nossa tolerância à complexidade quando mais precisamos dela.

A vida já seria complicada mesmo se as coisas que desejamos fossem simples (elas não são). As exceções às generalizações vazadas sobre o que fazer a seguir vêm da estrutura incerta do mundo real. Ou, de outra forma:

Os valores instrumentais muitas vezes não são especificados de forma concisa e local.

Suponha que haja uma caixa contendo um milhão de dólares. A caixa está trancada, não com uma fechadura comum, mas com uma dúzia de chaves que controlam uma máquina capaz de abri-la. Se você entender como a máquina funciona, poderá deduzir quais sequências de teclas abrirão a caixa. Existem várias sequências que podem ativar a máquina para abrir a caixa. No entanto, se você pressionar uma sequência completamente errada, a máquina incinera o dinheiro. E se você não conhece a máquina, não há regras simples como “pressionar qualquer tecla três vezes abre a caixa“ ou “pressionar cinco teclas diferentes sem repetições incinera o dinheiro“.

Existe uma especificação não local e concisa de quais teclas você deve pressionar: você deve pressionar as teclas de forma que a caixa seja aberta. Você pode escrever um programa compacto que calcule quais sequências de teclas são boas, ruins ou neutras, mas esse programa precisaria descrever a máquina, não apenas as teclas em si.

Da mesma forma, há uma especificação local não compacta sobre quais teclas pressionar: uma tabela de pesquisa gigante com os resultados para cada sequência possível de teclas. É um programa muito extenso, mas não menciona nada além das teclas.

Mas não é possível descrever quais sequências de teclas são boas, ruins ou neutras, o que é simples

e formulado apenas em termos das próprias teclas.

Pode ficar ainda pior se houver tentações de generalizações locais que se revelem como vazamentos. Pressionar a maioria das teclas três vezes seguidas abre a caixa, mas há uma tecla específica que incinera o dinheiro se você a pressionar apenas uma vez. Você pode pensar que encontrou uma generalização perfeita — uma classe de sequências localmente descritível que sempre abre a caixa — quando, na verdade, não conseguiu visualizar todos os caminhos possíveis da máquina ou não falhou em avaliar todos os efeitos colaterais.

A máquina representa a complexidade do mundo real. A abertura da caixa (que é boa) e o incinerador (que é ruim) representam os milhares de fragmentos de desejo que compõem nossos valores terminais. As teclas representam as ações, políticas e estratégias disponíveis para nós.

Quando se considera quantas maneiras diferentes valorizamos os resultados e, quão complicados são os caminhos que percorremos para alcançá-los, é surpreendente existir algo como um conselho ético útil. (Dentre todos os conselhos estranhos, mas úteis, o mais estranho de todos é que "o fim não justifica os meios".)

Porém, a complexidade da ação não precisa refletir a complexidade dos objetivos. Muitas vezes encontramos pessoas que sorriem com sabedoria e dizem: "Bem, a moralidade é complicada, você sabe. A circuncisão feminina é considerada correta em uma cultura e errada em outra, torturar pessoas nem sempre é algo ruim." Como você é ingênuo, cheio de necessidade de conclusões, pensando que existem regras simples.

Você pode afirmar, incondicional e categoricamente, que matar alguém é uma grande dose de negatividade em termos de utilidade terminal. Isso inclui até mesmo Hitler. Isso não significa que você não deva matar Hitler. Significa que a utilidade líquida resultante de matar Hitler carrega uma enorme dose negativa de utilidade em relação à morte de Hitler e uma dose imensamente maior de utilidade positiva em relação a todas as outras vidas que seriam salvas como consequência.

Muitos cometem o erro que alertei em "Valores Terminais e Valores Instrumentais" e pensam que se a utilidade esperada líquida resultante da morte de Hitler é considerada positiva, então a utilidade terminal local imediata também deve ser positiva. Isso significa que o princípio moral "A morte é sempre algo ruim" é, em si, uma generalização com vazamento. No entanto, isso é uma contagem dupla, envolvendo utilidades em vez de probabilidades; você está estabelecendo uma ressonância entre a utilidade esperada e a utilidade, em vez de um fluxo unidirecional da utilidade para a utilidade esperada.

Ou talvez seja apenas o desejo de um debate político unilateral: a melhor política não deve ter desvantagens.

Em minha filosofia moral, a utilidade negativa local resultante da morte de Hitler é estável, independentemente do que aconteça com as consequências externas e, portanto, com a utilidade esperada.

Claro, você pode apresentar um argumento moral de que é algo intrinsecamente bom punir pessoas más, mesmo com a pena de morte para pessoas extremamente más. Mas você não pode sustentar esse argumento moral apontando que a consequência de matar um homem armado pode ser salvar outras vidas. Isso apela para o valor da vida, não para o valor da morte. Se as utilidades esperadas são complicadas e têm vazamentos, isso não significa que as utilidades também devam ser complicadas e vazarem. Elas podem ser! Mas isso seria um argumento separado.

150 — A complexidade oculta dos desejos

> Desejo viver nos lugares que escolher, em uma versão fisicamente saudável do meu corpo atual, sem ferimentos visíveis e aparentemente normal, contendo o meu estado mental atual, um corpo consiga se curar de qualquer ferimento a uma taxa de três sigmas mais rápido do que a média, considerando a tecnologia médica disponível para mim, e que estará protegido de quaisquer doenças, lesões ou enfermidades que resultem em incapacidade, dor ou degradação funcional de qualquer sentido, órgão ou função corporal por mais de dez dias consecutivos, ou quinze dias em qualquer ano…
>
> — The Open-Source Wish Project, Wish For Immortality 1.1

Existem três tipos de gênios: aqueles aos quais você pode dizer com segurança: “Desejo que você faça o que eu desejo“; gênios para os quais nenhum desejo é seguro; e gênios que não são muito poderosos ou inteligentes.

Imagine que sua mãe idosa esteja presa em um prédio em chamas e, por acaso, você esteja em uma cadeira de rodas, incapaz de se apressar. Você poderia gritar: “Tire minha mãe daquele prédio!“, mas não haveria ninguém para ouvir.

Felizmente, você tem uma Bomba de Resultados no bolso. Esse dispositivo útil comprime o fluxo do tempo, aumentando a probabilidade de certos resultados e diminuindo a de outros.

A Bomba de Resultados não é senciente. Ele contém uma pequena máquina do tempo que reinicia o tempo, a menos que ocorra um resultado específico. Por exemplo, se você conectou os sensores da Bomba de Resultados a uma moeda e definiu que a máquina do tempo deve continuar reiniciando até que a moeda mostre “cara“, ao lançar a moeda, você a verá mostrar “cara“. (Os físicos afirmam que qualquer futuro em que ocorra uma “reinicialização“ é inconsistente e, portanto, nunca ocorre em primeiro lugar — então você não está realmente causando a morte de nenhuma versão sua.)

Qualquer proposição que você possa inserir na Bomba de Resultados acontecerá de alguma forma, desde que não viole as leis da física. Se você tentar inserir uma proposição altamente improvável, o dispositivo sofrerá uma falha mecânica espontânea antes que esse resultado ocorra.

Você também pode redirecionar o fluxo de probabilidade de maneiras mais quantitativas, usando a “função futura“ para ajustar a probabilidade de reinicialização temporal para diferentes resultados. Se a probabilidade de reinicialização temporal for de 99% quando a moeda der “cara“ e 1% quando a moeda der “coroa“, as chances mudam de 1:1 para 99:1 a favor de “coroa“. Se você tivesse uma máquina misteriosa que produzisse dinheiro e quisesse maximizar a quantidade de dinheiro gerada, usaria probabilidades de reinicialização que diminuíssem à medida que a quantidade de dinheiro aumentasse. Por exemplo, gerar $10 poderia ter uma probabilidade de reinicialização de 99,999999% e gerar $100 poderia ter uma probabilidade de reinicialização de 99,99999%. Dessa forma, você pode obter um resultado que tende a ser o mais alto possível na função futura, mesmo

sem conhecer o máximo ideal.

Então, desesperado, você retira a Bomba de Resultados do bolso — lembre-se de que sua mãe continua presa no prédio em chamas — e tenta descrever seu objetivo: tirar sua mãe do prédio!

A interface do usuário não aceita entradas em português. A Bomba de Resultados não é senciente, lembra-se? No entanto, possui scanners 3D para o ambiente próximo e recursos de correspondência de padrões. Então, você segura uma foto da cabeça e dos ombros de sua mãe, faz uma correspondência com a foto, usa a continuidade de objetos para selecionar todo o corpo dela (não apenas a cabeça e os ombros) e ajusta a função futura com base na distância de sua mãe ao centro do prédio. Quanto mais longe ela estiver do centro do prédio, menor será a probabilidade de reinicialização da máquina do tempo. Você grita "Tire minha mãe do prédio!" para dar sorte e pressiona *Enter*.

Por um momento, parece que nada acontece. Você olha em volta, esperando que o caminhão de bombeiros pare e os socorristas cheguem - ou até mesmo um corredor rápido e forte para tirar sua mãe do prédio -

Bum! Com um estrondo ensurdecedor, o cano de gás sob o prédio explode. Conforme a estrutura se desfaz, em uma sequência que parece se desenrolar em câmera lenta, você avista o corpo despedaçado de sua mãe sendo lançado para o alto, afastando-se rapidamente do antigo centro do prédio.

Ao lado da Bomba de Resultados, há um botão de arrependimento de emergência. Todas as funções futuras são automaticamente configuradas com um valor negativo enorme para o Botão de Arrependimento ser pressionado — uma probabilidade de reinicialização temporal de quase 1 - tornando extremamente improvável que a Bomba de Resultados faça algo que perturbe o usuário o suficiente para pressionar o Botão de Arrependimento. Você não consegue se lembrar de pressioná-lo. Mas você mal começa a alcançar o Botão de Arrependimento (e para que isso serve agora?), quando uma viga de madeira em chamas cai do céu e o esmaga.

Isso não era realmente o que você queria, mas pontua muito alto na função futura definida…

A Bomba de Resultados é um gênio de segunda classe. Nenhum desejo é seguro.

Se alguém lhe pedisse para salvar uma mãe de um prédio em chamas, você poderia ajudar ou fingir que não ouviu. Mas você nunca consideraria explodir o prédio. "Tire minha mãe do prédio" parece um desejo muito mais seguro do que realmente é, porque você não considera os planos aos quais atribui valores extremamente negativos.

Considere novamente a Tragédia do Selecionismo de Grupo: alguns dos primeiros biólogos afirmaram que a seleção de grupo para tamanhos de subpopulação baixos resultaria em restrição individual na reprodução; no entanto, ao forçar a seleção de grupo no laboratório, ocorreu canibalismo, especialmente de fêmeas imaturas. Retrospectivamente, é óbvio que, dada a forte seleção para tamanhos de subpopulação reduzidos, os canibais acabariam reproduzindo indivíduos que abrem mão voluntariamente de oportunidades reprodutivas. Mas comer crianças é uma solução tão repugnante que Wynne-Edwards, Allee, Brereton e outros pesquisadores de seleção simplesmente não consideraram isso. Eles só viram as soluções que eles próprios teriam utilizado.

Imagine que você tente corrigir a função futura especificando que a Bomba de Resultados não deve explodir o edifício: os resultados nos quais os materiais de construção são distribuídos em abundância, terão probabilidades de reinicialização temporal de aproximadamente 1.

Então, sua mãe cai de uma janela do segundo andar e quebra o pescoço. A Bomba de Resultados toma um caminho diferente no tempo que ainda resulta na sua mãe fora do prédio, mas não

era o que você desejava e certamente não era uma solução que um socorrista humano consideraria. Se ao menos o projeto Wish de código aberto tivesse desenvolvido um desejo de remover sua mãe de um prédio em chamas:

> Preciso mover minha mãe (definida como a mulher que compartilha metade de meus genes e que me deu à luz) para fora do edifício atualmente mais próximo de mim que está em chamas; mas é crucial que a integridade estrutural do edifício permaneça intacta, sem paredes desmoronando ou explodindo. Além disso, não posso esperar até que o fogo seja extinto para que uma equipe de resgate recupere o corpo da minha mãe.

Todos esses casos especiais, o número aparentemente infinito de correções necessárias, devem lembrá-lo da parábola da Adição Artificial — programar um Sistema Especialista Aritmético adicionando explicitamente cada vez mais afirmações como "quinze mais quinze é igual a trinta, mas quinze mais dezesseis é igual a trinta e um, não trinta".

Como você exclui o resultado no qual o prédio explode e sua mãe é lançada para o céu? Você olha para frente e prevê que sua mãe acabaria morta, e você não deseja essa consequência, então tenta evitar o evento que leva a isso.

Seu cérebro não está programado com uma declaração específica e pré-gravada de que "explodir um prédio em chamas que contém minha mãe é uma má ideia". No entanto, você está tentando pré-gravar essa declaração específica exata na função futura da Bomba de Resultados. Assim, o desejo está se transformando em uma tabela de pesquisa gigante que registra seu julgamento de todos os possíveis caminhos ao longo do tempo.

Você falhou em pedir o que realmente queria. Você queria que sua mãe continuasse viva, mas queria que ela ficasse mais afastada do centro do prédio.

Mas isso não é tudo o que você queria. Se sua mãe fosse resgatada do prédio, mas ficasse terrivelmente queimada, esse resultado teria uma classificação mais baixa em sua ordem de preferência do que um resultado em que ela fosse resgatada com saúde e salva. Portanto, você valoriza não apenas a vida de sua mãe, mas também a saúde dela.

E você valoriza não apenas a saúde física dela, mas também o estado emocional dela. Ser resgatada de uma forma traumatizante, por exemplo, por um monstro roxo gigante surgindo do nada e agarrando-a, seria inferior a um bombeiro aparecendo e escoltando-a por uma rota segura, longe das chamas. (Sim, devemos nos ater à física, mas talvez uma Bomba de Resultados suficientemente poderosa possa fazer alienígenas coincidentemente aparecerem na vizinhança exatamente naquele momento.) Certamente, você preferiria que ela fosse resgatada pelo monstro a ser queimada viva.

E se um buraco de minhoca se abrisse espontaneamente e a engolisse para uma ilha deserta? Melhor do que ela estar morta, mas pior do que ela estar viva, saudável, sem traumas e mantendo contato contínuo com você e outros membros de sua rede social.

Seria aceitável salvar a vida de sua mãe às custas da vida do cachorro da família, se ele corresse para alertar um bombeiro, mas fosse atropelado por um carro? Certamente, sim, mas seria melhor, em igualdade de condições, evitar matar o cachorro. Você não gostaria de trocar uma vida humana pela vida dele, mas e a vida de um assassino condenado? Faz diferença se o assassino morre tentando salvá-la, por bondade no coração dele? E se forem dois assassinos? Se o custo da vida de sua mãe fosse a destruição de todas as cópias existentes, incluindo as memórias, da Pequena Fuga em Sol Menor de Bach, isso valeria a pena? E se ela tivesse uma doença terminal e fosse morrer em dezoito meses, de qualquer forma?

Se o pé de sua mãe for esmagado por uma viga em chamas, vale a pena resgatar o restante

do corpo? E se for apenas a cabeça dela que estiver esmagada, deixando o corpo intacto? E se o corpo for esmagado, restando apenas a cabeça? E se houver uma equipe de criogenia esperando do lado de fora, pronta para preservar a cabeça? Uma cabeça congelada é uma pessoa? Terry Schiavo é uma pessoa? Quanto vale um chimpanzé?

Seu cérebro não é infinitamente complicado; há apenas uma complexidade finita de Kolmogorov / comprimento da mensagem suficiente para descrever todos os julgamentos que você faria. Mas o fato de essa complexidade ser finita não a torna pequena. Valorizamos muitas coisas, e não, elas não podem ser reduzidas à busca pela felicidade ou à valorização da aptidão reprodutiva.

Não há desejo seguro menor do que toda a moralidade humana. Existem muitos caminhos possíveis através do tempo. Você não pode visualizar todas as estradas que levam ao destino que você atribui ao gênio. “Maximizar a distância entre sua mãe e o centro do prédio“ pode ser alcançado de maneira ainda mais eficaz detonando uma arma nuclear. Ou, em níveis mais altos de poder do gênio, arremessando o corpo dela para fora do Sistema Solar. Ou, em níveis mais altos de inteligência genial, fazendo algo que nem você, nem eu poderia conceber, assim como um chimpanzé não pensaria em detonar uma arma nuclear. Você não pode visualizar todos os caminhos através do tempo, assim como não pode programar uma máquina de xadrez codificando um movimento para cada posição possível no tabuleiro.

E a realidade é muito mais complicada do que o xadrez. Você não pode prever antecipadamente quais de seus valores serão necessários para julgar o caminho do gênio no tempo. Especialmente se você deseja algo de longo prazo ou com um alcance mais amplo do que resgatar sua mãe de um prédio em chamas.

Temo que o projeto Wish de código aberto seja fútil, exceto como uma ilustração de como não abordar problemas de gênio. O único gênio seguro é aquele que compartilha todos os seus critérios de julgamento, e nesse caso, você só precisa dizer “Desejo que você faça o que eu desejo“. Isso simplesmente executa a função do gênio.

De fato, nem mesmo deveria ser necessário dizer nada. Para ser um executor seguro de desejos, um gênio deve compartilhar os mesmos valores que o levaram a formular o desejo. Caso contrário, o gênio pode não escolher um caminho no tempo que leve ao destino que você tinha em mente, ou pode falhar em excluir efeitos colaterais horríveis que levariam você a nem mesmo considerar um plano em primeiro lugar. Os desejos são generalizações com vazamentos, derivadas da enorme, mas finita, estrutura que é toda a sua moralidade; somente incluindo toda essa estrutura você pode tapar todos os vazamentos.

Com um gênio seguro, o desejo se torna supérfluo. Basta executar o gênio.

151 — Otimismo antropomórfico

A falácia central do antropomorfismo consiste em esperar que algo possa ser previsto pela caixa preta do seu cérebro, quando a estrutura causal difere tanto da de um cérebro humano, que você não tem base para esperar tal coisa.

Em Tragédia do Selecionismo de Grupo, os primeiros biólogos (antes de 1966) acreditavam que os predadores restringiriam voluntariamente sua reprodução para evitar a superpopulação de seu habitat e o esgotamento das populações de presas. Posteriormente, quando Michael J. Wade efetivamente replicou em laboratório condições quase impossíveis para a seleção de grupo, os adultos se adaptaram canibalizando ovos e larvas, principalmente as larvas fêmeas.[1]

Por que os defensores da seleção de grupo não consideraram essa possibilidade?

Imagine que você fosse membro de uma tribo e soubesse que, em breve, sua tribo enfrentaria escassez de recursos. Você poderia propor como solução que nenhum casal tivesse mais de um filho — depois do primeiro filho, o controle de natalidade seria implementado. Sugerir algo como: "Vamos cada um ter tantos filhos quanto pudermos, mas depois vamos caçar e canibalizar os filhos uns dos outros, especialmente as meninas", não ocorreria a você nem mesmo como uma possibilidade.

Pense em uma ordem de preferência para as soluções, em relação aos seus objetivos. Você deseja obter uma solução o mais alto possível nessa ordem de preferência. Como você pode encontrá-la? Com seu cérebro, é claro! Considere seu cérebro como um gerador de soluções de alto nível — um processo de busca, que produz soluções de alto nível em sua ordem de preferência inata.

O espaço de soluções em todos os problemas do mundo real geralmente é bastante amplo; por isso, é necessário um cérebro eficiente, que não perde tempo considerando a grande maioria das soluções de baixo nível.

Se sua tribo enfrentar escassez de recursos, você pode tentar pular em um pé só, ou morder seus próprios dedos dos pés. Essas "soluções" obviamente não funcionariam e acarretariam grandes custos, como é evidente ao examiná-las - mas na verdade seu cérebro é eficiente demais para desperdiçar tempo considerando tais soluções ruins; ele nem mesmo as cogita inicialmente. Em sua busca por soluções de alto nível, seu cérebro se direciona diretamente para áreas do espaço de soluções, como "Todos na tribo se reúnem e concordam em ter no máximo um filho por casal até que a escassez de recursos seja superada".

Uma solução de baixo nível, como "Todos têm quantos filhos for possível, e então canibalizam as meninas", não seria gerada em seu processo de pesquisa.

Entretanto, a classificação de uma opção como "baixa" ou "alta" não é uma propriedade inerente à opção em si, mas sim uma propriedade do processo de otimização que realiza a preferência. Diferentes processos de otimização conduzirão a diferentes ordens de busca.

No contexto da evolução, a estratégia de indivíduos se reproduzirem ao máximo e depois canibalizarem as filhas de outros é uma escolha óbvia, considerando que indivíduos restringindo voluntariamente sua própria reprodução em benefício do grupo é absolutamente ridícula. Ou, para expressar isso de maneira menos antropomórfica, o primeiro conjunto de alelos substituiria rapidamente o segundo em uma população. (E a seleção natural não tem uma ordem de busca óbvia aqui — essas duas alternativas parecem tão simples quanto as mutações.)

Consideremos o caso em que um dos biólogos afirmou: “Se uma população de predadores tiver recursos finitos, a evolução os moldará para restringir voluntariamente sua reprodução — é assim que eu faria se estivesse encarregado de construir predadores”. Isso seria um exemplo direto de antropomorfismo, no qual as linhas de raciocínio estão claramente expostas: eu faria dessa forma, portanto, deduzo que a evolução fará da mesma maneira.

Em minha área de trabalho, ocasionalmente nos deparamos com essa falácia abertamente. No entanto, suponha que você diga a alguém: “Um sistema de inteligência artificial não necessariamente funcionará da mesma forma que você”. E se você disser isso ao biólogo hipotético: “A evolução não funciona como você”. Qual seria a resposta? Posso lhe dar uma resposta que você dificilmente ouvirá: “Oh, verdade! Eu não havia percebido isso! Uma das etapas da minha inferência estava incorreta; vou descartar a conclusão e começar do zero.“

Não, o que você ouvirá, em vez disso, é uma justificativa de por que qualquer IA deveria raciocinar da mesma maneira que o falante. Ou uma razão pela qual a seleção natural, que segue critérios de otimização e utiliza métodos de otimização completamente diferentes, deveria agir da mesma forma que um ser humano consideraria uma boa ideia.

Daí surge a elaborada ideia de que a seleção de grupo favoreceria grupos de predadores nos quais os indivíduos abdicam voluntariamente de oportunidades reprodutivas.

Os defensores da seleção de grupo cometeram erros em suas previsões da mesma forma que alguém que comete a falácia abertamente. Suas conclusões finais foram as mesmas, como se estivessem assumindo abertamente que a evolução necessariamente pensa da mesma forma que eles. No entanto, eles apagaram o que estava escrito acima da linha de fundo de seu argumento, sem apagar a linha de fundo real, e criaram novas justificativas. Agora, o raciocínio falacioso está disfarçado, a etapa obviamente falha na inferência foi ocultada, embora a conclusão permaneça a mesma e, portanto, na realidade, continua totalmente errada.

Mas por que qualquer cientista faria isso? No final, os dados contradisseram os defensores da seleção de grupo, e eles ficaram envergonhados.

Como mencionei em “Critérios Falsos de Otimização”, nós, humanos, parecemos ter desenvolvido um instinto para argumentar que nossas preferências políticas derivam de praticamente qualquer critério de otimização. A política era uma característica do ambiente ancestral. Somos descendentes daqueles que argumentaram de forma mais persuasiva que os interesses da tribo — não apenas seus próprios interesses — exigiam a execução de seu odiado rival, Uglak. Certamente, não somos descendentes de Uglak, que falhou ao argumentar que o código moral de sua tribo — não apenas seu próprio interesse óbvio — exigia sua sobrevivência.

E, como somos mais persuasivos ao argumentar a favor do que acreditamos honestamente, desenvolvemos um instinto para acreditar honestamente que os objetivos dos outros e o código moral de nossa tribo realmente implicam que eles devem fazer as coisas do nosso jeito para seu próprio benefício.

Dessa forma, os defensores da seleção de grupo, imaginando a bela imagem de predadores restringindo sua reprodução, racionalizaram instintivamente por que a seleção natural deveria agir de acordo com suas próprias visões, mesmo que fosse contrário aos propósitos intrínsecos da seleção natural. “As raposas serão mais aptas se restringirem sua procriação! É serio!“ Elas vão até superar outras raposas que não restringem sua reprodução! Sinceramente!

O problema de tentar argumentar com a seleção natural para agir de acordo com suas preferências é que a evolução não possui um mecanismo que possa ser influenciado por seus argumentos. A evolução não age como nós — nem ao ponto de ter qualquer elemento que possa ouvir ou se importar com sua explicação detalhada sobre porque a evolução deveria agir de acordo com suas preferências. Argumentos humanos não são compatíveis com a estrutura interna da seleção natural como um processo de otimização — argumentos humanos não são usados para promover alelos, como os argumentos humanos têm um papel causal na política humana.

Portanto, em vez de persuadir a seleção natural a seguir suas visões, os defensores da seleção de grupo acabaram apenas envergonhados quando a realidade se mostrou diferente.

Existe um subtexto bastante relevante aqui em relação à IA hostil.

Mas a questão é geral: esse é o problema do raciocínio otimista em geral. O que é otimismo? É classificar as possibilidades conforme a ordem de preferência pessoal e selecionar um resultado que esteja no topo dessa ordem de preferência, e, de alguma forma, acreditar que esse resultado será a previsão correta. Que tipo de racionalizações elaboradas foram geradas ao longo desse processo provavelmente não é tão relevante quanto se poderia imaginar;observe a história cognitiva: o otimismo entra, o otimismo sai. No entanto, a natureza, ou qualquer outro processo em discussão, não está realmente fazendo uma escolha causal entre os resultados, classificando-os em ordem de preferência e selecionando um resultado mais alto. Portanto, o cérebro falha em sincronizar com o ambiente e a previsão acaba não correspondendo à realidade.

Referências

[1] Wade, “Group selections among laboratory populations of Tribolium.”

152 — Propósitos perdidos

Foi no jardim de infância ou na primeira série que me pediram para orar pela primeira vez. Recebi uma transliteração de uma oração em hebraico. Perguntei o significado das palavras. Disseram-me que, desde que eu orasse em hebraico, não precisava saber o significado das palavras, funcionaria de qualquer maneira.

Esse foi o começo da minha ruptura com o judaísmo.

Enquanto você lê isto, um jovem está sentado em uma mesa na universidade, estudando seriamente um material que não tem intenção de usar e nenhum interesse em conhecer por si só. Ele deseja um emprego bem remunerado, e um emprego bem remunerado requer um pedaço de papel, e o pedaço de papel exige um mestrado anterior, e o mestrado exige um diploma de bacharel, e a universidade que concede o diploma de bacharel exige que você faça um curso sobre padrões de tricô do século XII para se formar. Então, ele estuda diligentemente, pretendendo esquecer tudo no momento da prova final, mas ainda assim se esforça seriamente, porque deseja aquele pedaço de papel.

Talvez você tenha percebido que tudo isso é loucura, mas aposto que fez isso mesmo assim. Você não teve escolha, certo? Um estudo recente na área da Baía de São Francisco mostrou que 80% dos professores do ensino fundamental relataram gastar menos de uma hora por semana em aulas de ciências, e 16% disseram que não dedicam tempo algum às ciências. Por quê? Pelo que entendi, a causa imediata é a Lei "Nenhuma Criança Deixada para Trás" e legislações semelhantes. Praticamente todo o tempo em sala de aula agora é gasto na preparação para testes obrigatórios ao nível estadual ou federal. Parece que me lembro (embora não consiga encontrar a fonte) que apenas fazer os testes obrigatórios consome 40% do tempo de aula em uma escola.

A antiga burocracia soviética era famosa por se preocupar mais com as aparências do que com a realidade. Uma fábrica de calçados ultrapassou sua cota produzindo muitos sapatos minúsculos. Outra fábrica de calçados informou que o couro cortado, mas desmontado, era considerado um "sapato". Os burocratas superiores não estavam interessados em investigar muito, pois também queriam relatar o cumprimento das cotas. Tudo isso foi de grande ajuda para os camaradas que sofriam com os pés congelados.

Agora está sendo sugerido em várias fontes que a maioria das descobertas publicadas na medicina, embora sejam "estatisticamente significativas com $p < 0,05$", são falsas. Mas enquanto $p < 0,05$ continuar sendo o limite para publicação, por que alguém se apegaria a padrões mais elevados, quando isso exigiria financiamento maior para grupos experimentais maiores e reduziria a probabilidade de publicação? Todos sabem que o objetivo da ciência é publicar muitos artigos, assim como o objetivo de uma universidade é emitir certos pedaços de pergaminho, e o objetivo de uma escola é passar nos testes obrigatórios que garantem o orçamento anual. Você não pode definir as regras do jogo e, se tentar seguir regras diferentes, simplesmente perderá.

(Embora, por algum motivo, as revistas de física exijam um limite de $p < 0,0001$. É como se

concebessem algum outro propósito para sua existência além de publicar artigos de física.)

Tem chocolate no supermercado, e você pode chegar no supermercado dirigindo, e dirigir requer que você esteja no carro, o que significa abrir a porta do carro, que precisa de chaves. Se você descobrir que não há chocolate no supermercado, não continuará abrindo e batendo a porta do carro só porque a porta do carro precisa ser aberta. Raramente vejo pessoas perderem o controle dos planos que elas mesmas fizeram.

É outra história quando os incentivos devem fluir por meio de grandes organizações — ou pior, por várias organizações e grupos de interesse diferentes, alguns deles governamentais. Nesse caso, você vê comportamentos que seriam considerados insanidade literal se tivessem nascido de uma única mente. Alguém é pago toda vez que abre a porta de um carro, porque isso é mensurável; e essa pessoa não se importa se o motorista é pago para ir ao supermercado, muito menos se o comprador compra o chocolate ou se quem quer comer está feliz ou faminto.

De uma perspectiva bayesiana, os subobjetivos são epifenômenos das funções de probabilidade condicional. Não há utilidade esperada sem utilidade. Seria tolice pensar que o valor instrumental poderia adquirir uma vida matemática própria, deixando o valor terminal à margem? Não é uma abordagem sensata segundo critérios teóricos de tomada de decisões.

Contudo, considere a Lei "Nenhuma Criança Deixada para Trás". Os políticos querem dar a impressão de que estão fazendo algo em relação às dificuldades educacionais; os políticos precisam parecer ocupados para os eleitores neste ano, e não daqui a quinze anos, quando as crianças estiverem buscando emprego. Os políticos não são consumidores de educação. Os burocratas devem demonstrar progresso, o que significa que só estão interessados no progresso que pode ser medido este ano. Não são eles que acabarão ignorando a ciência. Os editores que encomendam livros didáticos e os comitês que compram livros didáticos não ficam sentados entediados em salas de aula.

Os verdadeiros consumidores de conhecimento são as crianças — que não podem pagar, não podem votar, não podem participar dos comitês. Seus pais cuidam delas, mas não frequentam as aulas; eles só podem responsabilizar os políticos com base em uma percepção superficial de serem "duros na educação". Os políticos estão ocupados demais buscando reeleição para analisar todos os dados por conta própria; eles precisam confiar em imagens superficiais de burocratas que estão ocupados e encomendando estudos — isso pode não funcionar para ajudar as crianças, mas ajuda os políticos a parecerem atenciosos. Os próprios burocratas não pretendem usar os livros didáticos, portanto, não se importam se eles são terríveis de ler, desde que o processo de aquisição pareça bom superficialmente. Os editores de livros didáticos não têm motivo para produzir livros ruins, mas eles sabem que o comitê de aquisição de livros didáticos vai compará-los com base na quantidade de assuntos abordados e que o comitê de compras da quarta série não está coordenado com o comitê de compras da terceira série, portanto, os editores acabam incluindo o máximo de assuntos possível em um único livro. Os professores não conseguirão cobrir um quarto do livro antes do final do ano, e o professor do ano seguinte terá que recomeçar do início. Os professores podem reclamar, mas não têm poder de decisão e, em última análise, não é o futuro deles que está em jogo, estabelecendo limites claros para o quanto eles irão se dedicar ao altruísmo não remunerado…

É incrível, quando se olha dessa forma — considerando todas as informações e incentivos perdidos — que ainda reste algo do propósito original, adquirir conhecimento. Embora muitos sistemas educacionais pareçam estar atualmente em processo de colapso em um estado muito melhor que nada.

Quer resolver o problema de verdade? Faça os políticos irem para a escola.

Uma única mente humana pode rastrear a expectativa probabilística de utilidade à medida que percorre as chances condicionais de uma dúzia de eventos intermediários — incluindo depen-

dências não locais, como quando a utilidade esperada de abrir a porta do carro depende da presença ou ausência de chocolate no supermercado. No entanto, as organizações só podem recompensar hoje o que é mensurável hoje, o que pode ser estabelecido em contrato legal hoje, e isso implica medir os eventos intermediários em vez de suas consequências distantes. Essas medidas intermediárias, por sua vez, são generalizações com vazamentos — muitas vezes com vários vazamentos. Os burocratas são gênios não confiáveis, pois não compartilham os mesmos valores daqueles que desejam alcançar seus objetivos.

Miyamoto Musashi disse: [1]

> A coisa primordial quando você empunha uma espada é a sua intenção de cortar o inimigo, independentemente dos meios. Sempre que você aparar, golpear, saltar, atacar ou tocar a espada cortante do inimigo, você deve cortar o inimigo no mesmo movimento. É essencial conseguir isso. Se você pensar apenas em atingir, saltar, bater ou tocar o inimigo, não conseguirá realmente cortá-lo. Mais do que tudo, você deve estar pensando em levar seu movimento até o cortar. Você deve pesquisar isso completamente[22].

(Eu gostaria de ter vivido em uma época em que eu pudesse apenas dizer aos meus leitores que eles precisam pesquisar algo minuciosamente, sem os insultar.)

Por que alguém perderia o controle de seus objetivos em um duelo de espadas? Se alguém tivesse sido ensinado a lutar e não tivesse desenvolvido a arte a partir de si mesmo, poderia não compreender o motivo de bloquear em um momento ou avançar em outro; poderia não perceber quando as regras têm exceções, falhando em reconhecer os momentos em que o método usual não funcionará. A essência da arte da racionalidade epistêmica é compreender como cada regra está cortando a verdade simultaneamente. O equivalente essencial na racionalidade pragmática — teoria da tomada de decisões contra a teoria da probabilidade — é sempre ver como cada utilidade esperada se traduz em utilidade real. Você precisa pesquisar isso minuciosamente.

C.J. Cherryh disse [2]:

> "Sua espada não tem uma lâmina. Ela possui apenas sua intenção. Quando isso se perde, você não tem uma arma[23]."

Tenho visto muitas pessoas se perderem quando fazem desejos ao gênio de uma IA imaginária, imaginando desejo após desejo que lhes parecem atraentes, às vezes com muitos remendos e às vezes sem nem mesmo a pretensão de cautela. E elas não fazem a transição para o meta-nível. Elas não buscam instintivamente um propósito, o mesmo instinto que me levou ao ateísmo aos cinco anos de idade. Elas não questionam, como questiono reflexivamente: "Por que acho que esse desejo é uma boa ideia? O gênio também julgaria assim?" Elas não veem a fonte de seus julgamentos, pairando por trás do julgamento como seu gerador. Elas perdem o controle da bola; sabem que ela quicou, mas não olham automaticamente para trás para ver de onde ela quicou — o critério que gerou seus julgamentos.

---

22 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. You must thoroughly research this.*

23 NT. Texto original em inglês. *Your sword has no blade. It has only your intention. When that goes astray you have no weapon.*

Da mesma forma, as pessoas não percebem automaticamente quando pessoas supostamente egoístas apresentam argumentos altruístas em favor do egoísmo, ou quando pessoas supostamente altruístas apresentam argumentos egoístas em favor do altruísmo.

As pessoas conseguem lidar bem com o rastreamento de metas para dirigir ao supermercado, quando tudo está dentro de suas próprias mentes e não há gênios, burocracias ou filosofias envolvidas. O problema é que a civilização real é imensamente mais complicada que isso. Dezenas de organizações e dezenas de anos se interpõem entre o sofrimento da criança na sala de aula e o recém-formado na faculdade que não é muito bom em seu trabalho. (Mas será que o entrevistador ou o gerente perceberiam se o recém-formado for bom em apenas aparentar ocupado?) A cada novo elo entre a ação e sua consequência, a intenção tem mais uma chance de se perder. Com cada elo intermediário, a informação se perde, o incentivo se perde. E isso incomoda muito menos a maioria das pessoas do que a mim, ou por que todos os meus colegas estavam dispostos a fazer orações sem saber o que significavam? Eles não sentiram o mesmo instinto de olhar para o gerador.

As pessoas podem aprender a manter os olhos na bola? Evitar que sua intenção se desvie? Nunca bloquear, avançar ou tocar sem saber o objetivo maior que estão buscando no mesmo movimento? As pessoas geralmente querem fazer seu trabalho, se todo o resto for igual. Pode haver algo como uma corporação saudável? Ou até mesmo uma civilização saudável? Isso pode ser apenas um sonho distante, mas é onde cheguei com todos esses ensaios sobre o fluxo de intenções (também conhecido como utilidade esperada, também conhecido como valor instrumental) sem perder o propósito (também conhecido como utilidade, também conhecido como valor instrumental). As pessoas podem aprender a sentir o fluxo de metas dos pais e das metas dos filhos? Conscientemente, conhecer, assim como implicitamente, a distinção entre utilidade esperada e utilidade real?

Você se preocupa com as ameaças à sua civilização? A pior meta-ameaça para uma civilização complexa é sua própria complexidade, pois essa complicação leva à perda de muitos propósitos.

Quando olho para trás, percebo que, mais do que qualquer outra coisa, minha vida tem sido impulsionada por uma aversão excepcionalmente forte à perda de propósitos. Espero que essa aversão possa ser transformada em uma habilidade que possa ser aprendida.

Referências

[1] Miyamoto Musashi, Book of Five Rings (New Line Publishing, 2003).

[2] Carolyn J. Cherryh, The Paladin (Baen, 2002).

Parte N — Um guia humano para palavras

153 — A Parábola da adaga

(Adaptado de Raymond Smullyan [1])

Era uma vez um bobo da corte que tinha um interesse incomum pela lógica.

O bobo da corte presenteou o rei com duas caixas. Na primeira caixa, havia uma inscrição que dizia:

"Esta caixa contém um sapo zangado ou a caixa com a inscrição falsa contém um sapo zangado, mas não ambos."

Na segunda caixa, havia uma inscrição que dizia:

"Esta caixa contém ouro e a caixa com a inscrição falsa contém um sapo zangado, ou esta caixa contém um sapo zangado e a caixa com a inscrição verdadeira contém ouro."

O bobo da corte então disse ao rei: "Uma das caixas contém um sapo zangado e a outra contém ouro; e uma, apenas uma das inscrições é verdadeira."

O rei abriu a caixa errada, sendo atacado por um sapo furioso.

"Veja", disse o bobo da corte, "vamos supor que a primeira inscrição seja verdadeira. Nesse caso, se a primeira caixa contiver ouro, a segunda caixa conterá um sapo zangado, enquanto a caixa com a inscrição verdadeira conterá ouro. Isso faria com que a segunda afirmação também fosse verdadeira. Agora, suponha que a primeira inscrição seja falsa e a primeira caixa contenha ouro. Então, a segunda inscrição seria…"

O rei ordenou que o bobo da corte fosse imediatamente jogado nas masmorras.

No dia seguinte, o bobo da corte foi trazido diante do rei, acorrentado, e mostraram-lhe duas caixas.

"Uma dessas caixas contém a chave para abrir suas correntes", disse o rei, "e se você a encontrar, estará livre. Mas a outra caixa contém uma adaga para o seu coração, caso você falhe."

A primeira caixa tinha a seguinte inscrição:

"Ambas as inscrições são verdadeiras ou ambas as inscrições são falsas."

A segunda caixa tinha a seguinte inscrição:

"Esta caixa contém a chave."

O bobo da corte refletiu assim: “Suponhamos que a primeira inscrição seja verdadeira. Isso significa que a segunda inscrição também deve ser verdadeira. Agora, suponha que a primeira inscrição seja falsa. Então, novamente, a segunda inscrição deve ser verdadeira. Portanto, logicamente, a segunda caixa deve conter a chave.“

O bobo da corte abriu a segunda caixa e encontrou uma adaga.

“Como?!” exclamou o bobo da corte, horrorizado, enquanto era arrastado para longe. “Isso é logicamente impossível!”

“É perfeitamente possível”, respondeu o rei. “Eu simplesmente escrevi essas inscrições em duas caixas e coloquei a adaga na segunda caixa.”

Referências

[1] Raymond M. Smullyan, What Is the Name of This Book? The Riddle of Dracula and Other Logical Puzzles (Penguin Books, 1990).

154 — A Parábola da cicuta

Todos os homens são mortais. Sócrates é um homem. Portanto, Sócrates é mortal.

— Silogismo medieval padrão.

Sócrates levou o copo de cicuta aos lábios…

“Você acha”, perguntou um dos espectadores, “que nem mesmo a cicuta não seria suficiente para matar um homem tão sábio e bom?”

“Não”, respondeu outro espectador, um estudante de filosofia, todos os homens são mortais, e Sócrates é um homem; e se um mortal beber cicuta, certamente ele morrerá.“

“Bem”, disse o espectador, “e se acontecer de Sócrates não ser mortal?“

“Bobagem”, respondeu o aluno, um pouco bruscamente, “todos os homens são mortais por definição; isso é parte do que queremos dizer com a palavra ‘homem’. Todos os homens são mortais, Sócrates é um homem, portanto Sócrates é mortal. Não é apenas uma suposição, mas uma certeza lógica.“

“Suponho que está certo…”, disse o espectador. “Oh, olhe, Sócrates já bebeu a cicuta enquanto conversávamos.”

“Sim, ele deve desmaiar a qualquer minuto agora” disse o aluno.

E eles esperaram, e esperaram, e esperaram…

“Sócrates parece não ser mortal”, disse o espectador.

“Então Sócrates não deve ser um homem”, respondeu o estudante. “Todos os homens são mortais, Sócrates não é mortal, logo Sócrates não é um homem. E isso não é apenas uma suposição, mas uma certeza lógica.“

O problema fundamental de argumentar que as coisas são verdadeiras “por definição” é que você não pode fazer a realidade seguir um caminho diferente apenas escolhendo uma definição diferente.

Você poderia raciocinar, talvez, da seguinte maneira: “Todos as coisas que observei vestindo roupas, falando uma língua e usando ferramentas também compartilham outras características, como respirar ar e ter sangue vermelho. Os últimos trinta ‘humanos’ desse grupo que observei bebendo cicuta, logo caíram desacordados. Sócrates veste uma toga, fala grego antigo fluentemente e bebeu cicuta de um cálice. Portanto, prevejo que Sócrates desmaiará nos próximos cinco minutos.”

Mas isso seria apenas uma mera suposição. Não seria, você sabe, absolutamente certo e eterno. Os filósofos gregos, assim como a maioria dos filósofos pré-científicos, tinham uma preferência pela certeza.

Felizmente, os filósofos gregos têm uma resposta contundente para o seu questionamento. Você entendeu mal o significado de “Todos os humanos são mortais” dizem eles. Isso não é apenas uma observação casual. Faz parte da definição da palavra “humano”. A mortalidade é uma das várias características individualmente necessárias e, quando combinadas, suficientes para determinar a inclusão na categoria de “humano”. A afirmação “Todos os humanos são mortais” é uma verdade logicamente válida, absolutamente inquestionável. E, se Sócrates é humano, ele deve ser mortal: essa é uma dedução lógica tão certa quanto a certeza pode ser.

Mas, então, nunca podemos ter certeza de que Sócrates é um "humano" até que ele seja considerado mortal. Não adianta observar que Sócrates fala grego fluentemente, ou que Sócrates tem sangue vermelho, ou mesmo que Sócrates tem DNA humano. Nenhuma dessas características é logicamente equivalente à mortalidade. É necessário vê-lo morrer antes de concluir que ele era humano.

(E mesmo assim, não é uma certeza infinita. E se Sócrates levantar do túmulo uma noite após você vê-lo morrer? Ou, de forma mais realista, e se Sócrates estiver inscrito para criogenia? Se a mortalidade for definida como uma um tempo de vida médio finito, então você nunca pode realmente saber se alguém é humano até observar até o fim da eternidade - apenas para ter certeza de que ele não voltará. Ou você pode pensar que viu Sócrates desmaiar, mas pode ser uma ilusão projetada em seus olhos por um scanner de retina. Ou talvez você tenha alucinado a coisa toda...)

O problema com os silogismos é que eles são sempre válidos. "Todos os humanos são mortais; Sócrates é humano; portanto, Sócrates é mortal" é — se tratado como um silogismo lógico — logicamente válido em nosso próprio universo. Também é logicamente válido em universos paralelos de Everett, onde, devido a uma bioquímica levemente diferente, a cicuta é uma guloseima deliciosa em vez de um veneno. E é logicamente válido mesmo em universos onde Sócrates nunca existiu, ou por falar nisso, onde os humanos nunca existiram.

A definição bayesiana de evidência que favorece uma hipótese é a evidência que temos mais probabilidade de observar se a hipótese for verdadeira, do que se for falsa. Observar que um silogismo é logicamente válido nunca pode ser evidência a favor de qualquer proposição empírica, porque o silogismo será logicamente válido, independentemente de a proposição ser verdadeira ou falsa.

Os silogismos são válidos em todos os mundos possíveis e, portanto, observar sua validade nunca nos diz nada sobre em qual mundo possível realmente vivemos.

Isso não significa que a lógica seja inútil - apenas que a lógica só pode nos dizer o que, de certa forma, já sabemos. Mas nem sempre acreditamos no que sabemos. O número 29.384.209 é primo? Conforme a definição de meu sistema decimal e meus axiomas de aritmética, já determinei a resposta para essa pergunta — mas ainda não sei qual é a resposta, e eu preciso raciocinar logicamente para descobrir.

Da mesma forma, se eu formar a generalização empírica incerta "Os humanos são vulneráveis à cicuta" e a suposição empírica incerta "Sócrates é humano", a lógica pode me dizer que minhas suposições anteriores estão prevendo que Sócrates será vulnerável à cicuta.

Alguns sugerem que podemos ver o raciocínio lógico como uma forma de resolver nossa incerteza sobre mundos possíveis impossíveis - eliminando a massa de probabilidade em mundos logicamente impossíveis que não sabíamos ser logicamente impossíveis. Nesse sentido, o argumento lógico pode ser tratado como observação.

Mas quando falamos sobre uma previsão empírica, como "Sócrates vai desmaiar e parar de respirar" ou "Sócrates vai fazer cinquenta polichinelos e competir nas Olimpíadas no próximo ano", isso está relacionado a mundos possíveis, não a mundos possíveis impossíveis.

A lógica pode nos dizer quais hipóteses correspondem a quais observações, e pode nos dizer o que essas hipóteses preveem para o futuro — ela pode trazer observações antigas e suposições anteriores para um novo problema. No entanto, a lógica nunca afirma categoricamente: "Sócrates vai parar de respirar agora". A lógica nunca dita nenhuma questão empírica; ela nunca resolve nenhuma questão do mundo real que poderia, por qualquer extensão da imaginação, ir para qualquer um dos lados.

Apenas se lembre da Litania Contra a Lógica:

A lógica permanece verdadeira, aonde quer que você vá,

Então a lógica nunca lhe diz onde você mora.

155 — Palavras como inferências ocultas

Suponhamos que encontrei um barril selado no topo, mas com um buraco grande o suficiente para caber minha mão. Eu alcanço o buraco e sinto um objeto pequeno e curvo. Puxo o objeto para fora e ele é azul — um ovo azulado. Em seguida, enfio a mão e sinto algo duro e plano, com bordas — que, ao ser extraído, revela-se um cubo vermelho. Retiro 11 ovos e 8 cubos, sendo todos os ovos azuis e todos os cubos vermelhos.

Agora, ao alcançar e sentir outro objeto em forma de ovo, antes de retirá-lo e olhar, devo adivinhar: qual será a aparência dele?

A evidência não comprova que todos os ovos no barril são azuis e todos os cubos são vermelhos. Nem mesmo sugere isso conclusivamente, pois uma amostra de 19 não é significativamente grande. Mesmo assim, vou supor que esse objeto em forma de ovo seja azul — ou, como segunda opção, vermelho. Se eu supor qualquer outra cor, há tantas possibilidades quanto cores distintas existem — e, aliás, quem disse que o ovo precisa ser de uma única tonalidade? Talvez possua uma imagem de um cavalo pintado.

Então, digo "azul", com uma pátina obediente de humildade. Afinal, sou uma pessoa sofisticada do tipo racionalista e tenho controle sobre minhas suposições e dependências — presumo estar fazendo isso corretamente, certo?

No entanto, quando um grande objeto em forma de felino listrado e amarelo salta sobre mim das sombras, penso: "Uau! Um tigre!" E não: "Mm… objetos com as propriedades de grandeza, amarelos, listrados e com forma felina possuíam anteriormente as propriedades de 'faminto' e 'perigoso' e, portanto, embora não seja logicamente necessário, pode ser uma suposição empiricamente acertada que aaaaahhhhh crac crunch gulp."

Por alguma razão estranha, o cérebro humano parece ter sido adaptado para fazer essa inferência de forma rápida, automática e sem manter um registro explícito das suas suposições.

E se eu chamar os objetos em forma de ovo de "bleggs" (para ovos azuis) e os cubos vermelhos de "rubes"? Nesse caso, quando eu alcançar e sentir outro objeto em forma de ovo, posso pensar: "Oh, é um blegg", em vez de considerar todo aquele problema de indução.

É um equívoco comum pensar que você pode definir uma palavra da maneira que quiser.

Isso seria verdade se o cérebro tratasse as palavras como construções puramente lógicas, classes aristotélicas, e se você nunca extraísse mais informações do que inseriu.

Ainda assim, o cérebro continua categorizando, quer aprovemos ou não conscientemente. "Todos os humanos são mortais; Sócrates é um humano; portanto, Sócrates é mortal" — assim falavam os antigos filósofos gregos. Ora, se a mortalidade faz parte da sua definição lógica de "humano", você não pode classificar logicamente Sócrates como humano, até observar que ele é mortal. Mas — aqui está o problema — Aristóteles sabia muito bem que Sócrates era um humano. O cérebro de

Aristóteles categorizou Sócrates como “humano” tão eficientemente quanto o seu próprio cérebro categoriza tigres, maçãs e tudo mais em seu ambiente: rapidamente, silenciosamente e sem aprovação consciente.

Aristóteles estabeleceu regras pelas quais ninguém poderia concluir que Sócrates era “humano” até após a sua morte. No entanto, Aristóteles e seus alunos continuaram a concluir que pessoas vivas eram humanas e, portanto, mortais; eles viram propriedades distintas, como rostos humanos e corpos humanos, e seus cérebros deram o salto para propriedades inferidas, como a mortalidade.

A falta de compreensão do funcionamento de sua própria mente não impede, felizmente, que a mente faça o seu trabalho. Caso contrário, os aristotélicos teriam morrido de fome, incapazes de concluir que um objeto era comestível apenas porque parecia e tinha a textura de uma banana.

Assim, os aristotélicos continuaram a classificar objetos do ambiente com base em informações parciais, como as pessoas sempre fizeram. Os estudantes da lógica aristotélica continuaram a pensar exatamente da mesma maneira, mas adquiriram uma concepção errônea do que estavam fazendo.

Se você perguntasse a um filósofo aristotélico se Carol, a quitandeira, era mortal, eles diriam “sim” Se você perguntasse como eles sabiam, eles diriam: “Todos os humanos são mortais; Carol é humana; portanto, Carol é mortal”. Se você perguntasse se era uma suposição ou uma certeza, eles afirmariam ser uma certeza (pelo menos até o século XVI). Se você perguntasse como eles sabiam que os humanos eram mortais, eles diriam que isso foi estabelecido por definição.

Os aristotélicos continuavam sendo as mesmas pessoas, mantinham suas naturezas originais, mas adquiriram crenças incorretas sobre seu próprio funcionamento. Eles se olharam no espelho da autoconsciência e viram algo diferente de seus verdadeiros “eus”: eles refletiram incorretamente.

Seu cérebro não trata as palavras como definições lógicas sem consequências empíricas, e portanto, você também não deveria. O simples ato de criar uma palavra pode fazer com que sua mente aloque uma categoria e, assim, desencadear inferências inconscientes de similaridade. Ou bloquear inferências de similaridade; se eu criar dois rótulos, posso fazer com que sua mente aloque duas categorias. Observe como usei “você” e “seu cérebro” como se fossem coisas diferentes?

Cometer erros sobre o interior da sua cabeça não muda o que está lá; caso contrário, Aristóteles teria morrido ao concluir que o cérebro era um órgão para resfriar o sangue. Os erros filosóficos geralmente não interferem nas inferências perceptivas instantâneas.

No entanto, erros filosóficos podem bagunçar gravemente os processos de pensamento deliberado que usamos para tentar corrigir nossas primeiras impressões. Se você acredita que pode “definir uma palavra da maneira que quiser”, sem perceber que seu cérebro continua categorizando sem sua supervisão consciente, então você não fará o esforço de escolher suas definições com sabedoria.

156 — Extensões e intensões[24]

"O que é vermelho?"

"Vermelho é uma cor."

"O que é uma cor?"

"Uma cor é uma propriedade de uma coisa."

Mas afinal, o que é uma coisa? E o que é uma propriedade? Rapidamente, ambos se perdem em um labirinto de palavras definidas por outras palavras, um problema que Steven Harnad descreveu uma vez como tentar aprender chinês com um dicionário chinês-chinês.

Se, por outro lado, me perguntasse: "O que é vermelho?", eu responderia apontando para objetos vermelhos ao meu redor. Por exemplo, eu poderia mostrar uma placa de pare, uma camisa vermelha, um semáforo vermelho, meu sangue após um corte acidental, um cartão de visita vermelho e, por fim, usaria a roda de cores no meu computador para selecionar a área vermelha. Isso provavelmente seria suficiente, embora, se você conhece o significado da palavra "Não", os verdadeiramente rigorosos insistiriam que eu apontasse para o céu e dissesse "Não".

Acredito que tenha roubado esse exemplo de S. I. Hayakawa — embora não tenha muita certeza, pois ouvi isso em meio ao borrão indistinto da minha infância. (Quando eu tinha doze anos, meu pai apagou acidentalmente todos os arquivos do meu computador. Não me lembro de nada antes disso.)

Mas é assim que me recordo de ter aprendido, pela primeira vez, sobre a diferença entre definição intensional e extensional. Fornecer uma "definição intensional" significa definir uma palavra ou frase em termos de outras palavras, como faz um dicionário. Já fornecer uma "definição extensional" é apontar exemplos, como os adultos fazem ao ensinar crianças. A frase anterior dá uma definição intencional de "definição extensional", o que a torna um exemplo extensional de "definição intencional".

Na racionalidade de Hollywood e na cultura popular em geral, os "racionalistas" são retratados como obcecados por palavras, flutuando em um espaço verbal interminável desconectado da realidade.

Mas os verdadeiros Racionalistas Tradicionais há muito tempo insistem em manter uma conexão estreita com a experiência:

> Se você procurar uma definição de lítio em um livro de química, poderá ser informado de que é o elemento cuja massa atômica é quase 7. Mas se o autor for mais lógico, ele lhe dirá que, ao procurar entre os minerais que são vítreos, translúcidos, cinza ou branco, muito duros, quebradiços e insolúveis, você encontrará um que dá um tom carmesim a uma chama sem luminosidade,esse mineral, quando triturado com cal ou raticida de witherita , pode

24 NT. **Intensão**: refere-se ao **significado interno** ou às **propriedades definidoras** de um termo ou conceito. Por exemplo, a intensão de "triângulo" inclui "figura geométrica de três lados". Não confundir com intenção, que refere-se à propósito. **Extensão**: refere-se ao **conjunto de objetos ou entidades** que um termo denota no mundo real. Por exemplo, a extensão de "triângulo" é o conjunto de todas as figuras triangulares existentes.
Essa distinção é fundamental na Filosofia da Linguagem para analisar como palavras e conceitos se relacionam com o mundo e com o pensamento.

> ser parcialmente dissolvido em ácido muriático. Se a solução resultante for evaporada e o resíduo for extraído com ácido sulfúrico e devidamente purificado, poderá ser convertido por métodos comuns em um cloreto que, ao ser obtido no estado sólido, fundido e eletrolisado com meia dúzia de baterias poderosas, produzirá uma pequena quantidade de um metal prateado rosado que flutua na gasolina. E esse material é um espécime de lítio[25].
>
> — Charles Sanders Peirce [1]

Este é um exemplo de "mente lógica" conforme descrito por um racionalista tradicional genuíno , em vez de um roteirista de Hollywood.

No entanto, observe que Peirce não está realmente te mostrando um pedaço de lítio. Ele não tinha fragmentos de lítio colados em seu livro. Em vez disso, ele está fornecendo um mapa do tesouro — um procedimento intencionalmente definido que, quando seguido, o levará a um exemplo concreto de lítio. Isso não é o mesmo que entregar um pedaço de lítio para você, mas também não é o mesmo que dizer "peso atômico 7". (Embora, se você tiver olhos suficientemente aguçados, a frase "3 prótons" poderia permitir que você identificasse o lítio de relance…)

Essa é uma definição intensional e extensional, a qual é uma maneira de transmitir a outra pessoa o que você quer dizer com um conceito. Quando mencionei "definições" anteriormente, estava me referindo à maneira de comunicar conceitos — explicar a outra pessoa o que você quer dizer com "vermelho", "tigre", "humano" ou "lítio". Agora, falaremos sobre os próprios conceitos reais.

A intensão real do meu conceito de "tigre" é o padrão neural (no meu córtex temporal) que examina um sinal recebido do córtex visual para determinar se é ou não um tigre.

A extensão real do meu conceito de "tigre" é tudo o que chamo de tigre.

As definições intencionais não capturam todas as intenções, e as definições extensionais não abrangem todas as extensões. Se eu apenas apontar para um tigre e disser a palavra "tigre" a comunicação pode falhar se a outra pessoa pensar que estou me referindo a um "animal perigoso" ou a um "tigre macho" ou a uma "coisa amarela". Da mesma forma, se eu disser "animal perigoso listrado de amarelo e preto", sem apontar para nada, o ouvinte pode imaginar vespas gigantes.

Não é possível expressar em palavras todos os detalhes do conceito cognitivo — como ele existe em sua mente — que permite reconhecer as coisas como tigres ou não-tigres. É algo muito vasto. E você não pode apontar todos os tigres que já viu, muito menos tudo o que você chamaria de tigre.

As definições mais eficazes utilizam uma combinação de comunicação intensional e extensional para definir um conceito. Ainda assim, você está apenas comunicando mapas de conceitos ou instruções para construir conceitos — não está comunicando as categorias reais conforme existem em sua mente ou no mundo.

(Sim, com bastante criatividade, é possível construir exceções a essa regra, como "Frases que Eliezer Yudkowsky publicou contendo o termo 'huragaloni' a partir de 4 de fevereiro de 2008." Acabei de apresentar a você a amplitude total desse conceito. No entanto, exceto na matemática, as definições geralmente são como mapas do tesouro, não os próprios tesouros.)

Essa é mais uma razão pela qual não se pode "definir uma palavra de qualquer maneira que você

---

25 NT. Texto original em inglês. *If you look into a textbook of chemistry for a definition of lithium, you may be told that it is that element whose atomic weight is 7 very nearly. But if the author has a more logical mind he will tell you that if you search among minerals that are vitreous, translucent, grey or white, very hard, brittle, and insoluble, for one which imparts a crimson tinge to an unluminous flame, this mineral being triturated with lime or witherite rats-bane, and then fused, can be partly dissolved in muriatic acid; and if this solution be evaporated, and the residue be extracted with sulphuric acid, and duly purified, it can be converted by ordinary methods into a chloride, which being obtained in the solid state, fused, and electrolyzed with half a dozen powerful cells, will yield a globule of a pinkish silvery metal that will float on gasolene; and the material of that is a specimen of lithium.*

queira”: não é possível programar conceitos diretamente no cérebro de outra pessoa.

Mesmo no paradigma aristotélico, onde fingimos que as definições são os próprios conceitos reais, não se tem liberdade simultânea de intenção e extensão. Suponhamos que eu defina Marte como “uma enorme esfera rochosa vermelha, com cerca de um décimo da massa da Terra e 50% mais distante do Sol”. Nesse caso, é uma questão separada mostrar que essa definição intencional corresponde a algo específico em minha experiência extensional ou, de fato, corresponde a algo real. Se, em vez disso, eu disser “Isso é Marte” e apontar para uma luz vermelha no céu noturno, torna-se uma questão distinta mostrar que essa luz extensional corresponde a qualquer definição intencional específica que eu possa propor — ou qualquer crença intencional que eu possa ter — como “Marte é o Deus da Guerra”.

Entretanto, a maior parte do trabalho do cérebro ao aplicar intenções ocorre subconscientemente. Não temos consciência de que nossa identificação de uma luz vermelha como “Marte” é algo separado de nossa definição verbal “Marte é o Deus da Guerra”. Não importa que tipo de definição intencional eu crie para descrever Marte, minha mente acredita que “Marte” se refere a essa coisa e que é o quarto planeta do Sistema Solar.

Ao considerar como a mente humana funciona realmente de maneira pragmática, a noção de que se pode “definir uma palavra de qualquer maneira que eu queira” logo se transforma em “posso acreditar em qualquer coisa sobre um conjunto fixo de objetos” ou “posso manipular qualquer objeto dentro ou fora de um teste de adesão fixo”. Assim como geralmente não se pode transmitir toda a intenção de um conceito por meio de palavras, por ser um teste complexo e abrangente de associação neural, não se pode controlar toda a intenção de um conceito, pois ele é aplicado subconscientemente. Por isso, é tão comum argumentar que XYZ é verdadeiro “por definição”. Se as mudanças de definição fossem meras operações nulas empíricas, como deveriam ser, ninguém se incomodaria em discuti-las. No entanto, se abusarmos um pouco das definições, elas se transformam em varinhas mágicas — apenas em discussões, é claro, não na realidade.“

Referências

[1] Charles Sanders Peirce, Collected Papers (Harvard University Press, 1931).

157 — Agrupamentos de similaridade

Era uma vez, os filósofos da Academia de Platão afirmavam que a melhor definição para o humano era “um bípede sem penas”. Conta-se que Diógenes de Sinope, também conhecido como Diógenes, o Cínico, prontamente exibiu uma galinha depenada e declarou: “Aqui está o homem de Platão”. Os platônicos logo alteraram sua definição para “um bípede sem penas com unhas largas”.

Nenhum dicionário, nenhuma enciclopédia jamais listou todas as coisas que os humanos têm em comum. Temos sangue vermelho, cinco dedos em cada uma das duas mãos, crânios ósseos, 23 pares de cromossomos — mas o mesmo pode ser dito de outras espécies animais. Fabricamos ferramentas complexas para criar outras ferramentas complexas, usamos linguagem combinatória sintática e exploramos reações de fissão crítica como fonte de energia — essas coisas podem servir para destacar apenas os humanos, mas não todos os humanos — muitos de nós nunca construíram um reator de fissão. Com o conjunto certo de sequências de genes necessárias e suficientes, você poderia destacar todos e apenas os humanos — pelo menos por enquanto — mas ainda estaria longe de abranger tudo o que os humanos têm em comum.

Mas, desde que você não esteja perto de uma galinha depenada, dizer “Procure por bípedes sem penas” pode servir para selecionar algumas dezenas de coisas específicas que são humanas, em contraste com casas, vasos, sanduíches, gatos, cores ou teoremas matemáticos.

Uma vez que a definição de “bípedes sem penas” tenha sido associada a alguns bípedes sem penas em particular, você pode examinar o grupo e começar a identificar algumas das outras características — além das simples pernas sem penas — que esses “bípedes sem penas” parecem compartilhar em comum.

Os bípedes sem penas que você observa também parecem usar linguagem, construir ferramentas complexas, falar uma linguagem combinatória com sintaxe, sangrar sangue vermelho quando cutucados e morrer quando bebem cicuta.

Assim, a categoria “humano” se enriquece e incorpora cada vez mais características; e quando Diógenes finalmente apresenta sua galinha depenada, não nos enganamos: está claro que essa galinha depenada não se assemelha aos outros “bípedes sem penas”

(Se a lógica aristotélica fosse um bom modelo de psicologia humana, os platônicos teriam olhado para a galinha depenada e dito: “Sim, isso é um humano; onde você quer chegar?”)

Se o primeiro bípede sem penas que você vê é uma galinha depenada, você pode acabar pensando que o rótulo verbal “humano” denota uma galinha depenada; então, posso modificar meu mapa do tesouro para apontar para “bípedes sem penas com unhas largas” e, se for sábio, continuar dizendo: “Está vendo Diógenes ali? Aquele é um humano, e eu sou um humano, e você é um humano; e aquele chimpanzé ali não é humano, embora seja bastante parecido.“

A pista inicial só precisa levar o usuário ao agrupamento de similaridade — o grupo de coisas que têm muitas características em comum. Depois disso, a pista inicial serviu ao seu propósito e posso transmitir a nova informação “os humanos são atualmente mortais” ou qualquer outra coisa que eu queira dizer sobre nós, bípedes sem penas.

Um dicionário é melhor pensado não como um livro de definições aristotélicas de classes, mas como um livro de dicas para combinar rótulos verbais com agrupamentos de similaridade, ou rótulos correspondentes com propriedades úteis para distinguir agrupamentos de similaridade.

158 — Tipicidade e similaridade assimétrica

Pássaros voam. Bem, exceto avestruzes. Mas qual ave é a mais típica — um tordo ou um avestruz?

E qual é a cadeira mais típica: uma cadeira de escritório, uma cadeira de balanço ou um pufe?

A maioria das pessoas diria que um tordo é um pássaro mais típico, assim como uma cadeira de escritório é a cadeira mais típica. Os psicólogos cognitivos que estudam experimentalmente esse tipo de questão o fazem sob o título de "efeitos de tipicidade" ou "efeitos de protótipo". [1] Para afirmações como "Um tordo é um pássaro" ou "Um pinguim é um pássaro", os tempos de reação são mais rápidos para exemplos mais centrais. [2] As medidas de tipicidade se correlacionam bem com diferentes métodos investigativos, como os tempos de reação. Além disso, é possível pedir às pessoas que classifiquem diretamente, em uma escala de 1 a 10, o quão bem um exemplo (como um tordo específico) se encaixa em uma categoria (como "pássaro").

Então, temos uma medida mental de tipicidade - que pode funcionar como uma heurística - mas será que existe um viés correspondente que podemos utilizar para identificá-la?

Bem, qual dessas afirmações parece mais natural: "98 é aproximadamente 100" ou "100 é aproximadamente 98"? Se você for como a maioria das pessoas, a primeira afirmação parece fazer mais sentido. [3] Por motivos semelhantes, quando solicitadas a avaliar a semelhança entre México e Estados Unidos, as pessoas dão avaliações consistentemente mais altas do que quando solicitadas a avaliar a semelhança entre Estados Unidos e México. [4]

E se isso ainda parece inofensivo, um estudo de Rips mostrou que as pessoas tendem a esperar que uma doença se espalhe mais facilmente de tordos para patos em uma ilha do que de patos para tordos. [5] Embora não seja logicamente impossível, do ponto de vista pragmático, qualquer característica que diferencie um pato de um tordo e minimize a probabilidade de transmissão de uma doença de patos para tordos também seria uma diferença entre um tordo e um pato, reduzindo a probabilidade de transmissão de uma doença de tordos para patos.

É possível apresentar racionalizações, como "Bem, pode haver mais espécies de tordos próximas, o que tornaria a doença mais provável de se espalhar inicialmente, etc." mas tome cuidado para não racionalizar demais as probabilidades de classificações feitas por indivíduos que nem perceberam que uma comparação estava ocorrendo. E não se esqueça de que o México é mais parecido com os Estados Unidos do que os Estados Unidos com o México, e que 98 está mais próximo de 100 do que 100 está de 98. Uma interpretação mais simples é que as pessoas estão usando a heurística de similaridade (demonstrada) como um substituto para a probabilidade de propagação de uma doença, e essa heurística é (comprovadamente) assimétrica.

O Kansas está extraordinariamente próximo do centro dos Estados Unidos, enquanto o Alasca está extraordinariamente distante. Portanto, o Kansas provavelmente está mais próximo da maioria dos lugares nos EUA, e o Alasca provavelmente está mais distante. No entanto, não segue logicamente que o Kansas esteja mais próximo do Alasca do que o Alasca do Kansas. Mas as pessoas parecem raciocinar (metaforicamente falando) como se a proximidade fosse uma propriedade inerente ao Kansas e a distância uma propriedade inerente ao Alasca, de modo que o Kansas continua próximo, mesmo do Alasca, e o Alasca continua distante, mesmo do Kansas.

Portanto, mais uma vez, podemos observar que a noção de categorias de Aristóteles — classes lógicas com membros determinados por uma coleção de propriedades que são individualmente estritamente ne-

cessárias e juntas são estritamente suficientes — não é um bom modelo para a psicologia cognitiva humana. (A visão científica mudou um pouco nos últimos 2.350 anos? Quem diria?) Nem mesmo raciocinamos, como se a associação a um conjunto fosse uma propriedade verdadeira ou falsa; as afirmações sobre a associação a um conjunto podem ser mais ou menos verdadeiras. (Observação: isso não é o mesmo que ser mais ou menos provável.)

Essa é mais uma razão para não fingir que você, ou qualquer outra pessoa, vai realmente tratar as palavras como classes lógicas aristotélicas.

Referências

[1] Eleanor Rosch, “Principles of Categorization,” in Cognition and Categorization, ed. Eleanor Rosch and Barbara B. Lloyd (Hillsdale, NJ: Lawrence Erlbaum, 1978).

[2] George Lakoff, Women, Fire, and Dangerous Things: What Categories Reveal about the Mind (Chicago: Chicago University Press, 1987).

[3] Jerrold Sadock, “Truth and Approximations,“ Papers from the Third Annual Meeting of the Berkeley Linguistics Society (1977): 430–439.

[4] Amos Tversky and Itamar Gati, “Studies of Similarity,“ in Cognition and Categorization, ed. Eleanor Rosch and Barbara Lloyd (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1978), 79–98.

[5] Lance J. Rips, “Inductive Judgments about Natural Categories,“ Journal of Verbal Learning and Verbal Behavior 14 (1975): 665–681.

159 — A Estrutura de Agrupamentos do Espaço das Coisas

A noção de um “espaço de configuração” é uma forma de traduzir descrições de objetos em posições de objetos. Pode parecer que o azul está “mais próximo” do azul esverdeado do que do vermelho, mas quanto mais perto? É difícil responder a essa pergunta apenas olhando para as cores. Mas é útil saber que as coordenadas de cores (proporcionais) em RGB são 0:0:5, 0:3:2 e 5:0:0. Isso seria ainda mais clara se fosse plotado em um gráfico 3D.

Da mesma forma, você pode ver um tordo como um tordo — com cauda marrom, peito vermelho, formato padrão de tordo, velocidade máxima de voo quando está sem carga, seu DNA típico da espécie e alelos individuais. Ou você pode vê-lo como um único ponto em um espaço de configuração, no qual as dimensões descrevem tudo o que sabemos ou podemos saber sobre o tordo.

Um tordo é maior que um vírus e menor que um porta-aviões — essa pode ser a dimensão do “volume” Da mesma forma, um tordo pesa mais que um átomo de hidrogênio e menos que uma galáxia; essa pode ser a dimensão de “massa”. Diferentes tordos terão correlações fortes entre “volume” e “massa”, então os pontos que representam tordos estarão alinhados em uma sequência bastante linear nessas duas dimensões — mas a correlação não será exata, então realmente precisamos de duas dimensões separadas.

Essa é a vantagem de ver os tordos como pontos no espaço: você não conseguiria visualizar a tendência linear com tanta facilidade se estivesse apenas imaginando os tordos como criaturas fofas que batem as asas.

O DNA de um tordo é uma variável altamente multidimensional, mas você ainda pode pensar nisso como parte da localização de um tordo no espaço das coisas — com milhões de coordenadas quaternárias, uma coordenada para cada base de DNA — ou talvez uma visão mais sofisticada que isso. A forma e a cor (refletância da superfície) do tordo também podem ser consideradas parte da posição do tordo no espaço das coisas, mesmo que não sejam dimensões únicas.

Assim como o ponto de coordenada 0:0:5 contém as mesmas informações que a cor azul real em HTML, não devemos perder informações ao ver os tordos como pontos no espaço. A mesma afirmação sobre a massa do tordo é verdadeira, quer visualizemos um tordo equilibrando a balança oposta a um peso de 0,07 kg ou um ponto de tordo com uma coordenada de massa de +70.

Podemos até imaginar um espaço de configuração com uma ou mais dimensões para cada característica distinta de um objeto, de modo que a posição do ponto de um objeto nesse espaço corresponda a todas as informações sobre o próprio objeto real. Isso seria representado de forma bastante redundante — as dimensões incluiriam massa, volume e densidade.

Se isso parece extravagante, os físicos quânticos usam um espaço de configuração de dimensão infinita, e um único ponto nesse espaço descreve a localização de cada partícula no universo. Portanto, na verdade, estamos sendo comparativamente conservadores em nossa visualização do espaço das coisas — um ponto no espaço das coisas descreve apenas um objeto, não o universo inteiro.

Se não temos certeza sobre a massa e o volume exatos do tordo, podemos pensar em uma pequena nuvem no espaço das coisas, um volume de incerteza, dentro do qual o tordo pode estar. A densidade dessa nuvem é a densidade de nossa crença de que o tordo tem aquela massa e volume específicos. Se tivermos mais certeza sobre a densidade do tordo do que sobre sua massa e volume, nossa nuvem de probabilidade

estará altamente concentrada na dimensão da densidade e concentrada em torno de uma linha oblíqua no subespaço de massa/volume. (De fato, essa nuvem é, na verdade, uma superfície, devido à relação V*D = M.)

As "categorias radiais" são como os psicólogos cognitivos descrevem os limites não aristotélicos das palavras. A "mãe" central concebe seu filho, dá à luz e o sustenta. Uma doadora de óvulos que nunca vê seu filho, é uma mãe? Ela é a "mãe genética". O que dizer de uma mulher que recebe um embrião estranho implantado e o carrega até o fim? Ela é uma "mãe de aluguel". E a mulher que cria um filho que não é geneticamente seu? Bem, ela é uma "mãe adotiva". O silogismo aristotélico seria: "Os humanos têm dez dedos, Fred tem nove dedos, portanto Fred não é humano", mas a maneira como realmente pensamos é: "Os humanos têm dez dedos, Fred é um humano, portanto Fred é um 'humano com nove dedos'".

Podemos pensar sobre a radialidade das categorias em termos intencionais, conforme descrito acima — propriedades que geralmente estão presentes, mas opcionalmente ausentes. Se pensarmos na intenção da palavra "mãe", poderíamos considerá-la como um brilho distribuído no espaço das coisas, um brilho cuja intensidade corresponde ao grau em que aquele volume do espaço das coisas corresponde à categoria "mãe". O brilho está concentrado no centro da genética, do nascimento e da criação dos filhos, o volume das doadoras de óvulos também brilharia, mas com menos intensidade.

Ou podemos pensar na radialidade das categorias de forma extensional. Suponha que mapeamos todos os pássaros do mundo no espaço das coisas, usando uma métrica de distância que corresponda tanto quanto possível à similaridade percebida em humanos: um tordo é mais semelhante a outro tordo do que qualquer um é semelhante a um pombo, mas tordos e pombos são mais semelhantes entre si do que com um pinguim, etc.

Assim, o centro de todos os pássaros seria densamente povoado por muitos aglomerados próximos, tordos, pardais, canários, pombos e muitas outras espécies. Águias, falcões e outras grandes aves predadoras ocupariam um aglomerado próximo. Os pinguins estariam em um grupo mais distante, assim como galinhas e avestruzes.

O resultado pode se assemelhar, de fato, a um aglomerado astronômico: muitas galáxias orbitando o centro e alguns pontos fora do padrão.

Ou poderíamos pensar simultaneamente sobre a intenção da categoria cognitiva "pássaro" e sua extensão nos pássaros reais: os aglomerados centrais de tordos e pardais brilhando intensamente com uma típica "passaridade"; aglomerados satélites de avestruzes e pinguins brilhando mais fracamente com uma "passaridade" atípica, e Abraham Lincoln a alguns megaparsecs de distância, não brilhando de forma alguma.

Prefiro essa última visualização — os pontos brilhantes — porque, a meu ver, a estrutura da intenção cognitiva decorre da estrutura do agrupamento extensional. Primeiro vem a estrutura no mundo, a distribuição empírica dos pássaros no espaço das coisas; então, ao observá-la, formamos uma categoria cujo brilho intencional reflete grosseiramente essa estrutura.

Isso nos dá ainda outra visão de por que as palavras não são classes aristotélicas: a estrutura empírica agrupada do universo real não é tão cristalina. Um aglomerado natural, um grupo de coisas altamente semelhantes entre si, pode não ter um conjunto de propriedades necessárias e suficientes — nenhum conjunto de características que todos os membros do grupo tenham e nenhum não membro tenha.

No entanto, mesmo que uma categoria esteja irremediavelmente embaçada e irregular, não há necessidade de entrar em pânico. Eu não teria objeção se alguém dissesse que os pássaros são "coisas com penas que voam". Mas os pinguins não voam! Bem, está certo. A regra usual tem uma exceção; não é o fim do mundo. De qualquer forma, não se pode esperar que as definições correspondam exatamente à estrutura empírica do espaço das coisas, porque o mapa é menor e muito menos complicado do que o território. O objetivo da definição de "coisas com penas que voam" é levar o ouvinte ao grupo de pássaros, não fornecer uma descrição completa de cada pássaro existente até o nível molecular.

Quando você desenha um limite em torno de um grupo de pontos extensionais agrupados empiricamente no espaço das coisas, você pode encontrar pelo menos uma exceção para cada regra intencional simples que você possa inventar.

Mas se uma definição funciona bem o suficiente, na prática, para identificar o grupo empírico pretendido, contestá-la pode ser chamado de "picuinha".

160 — Consultas disfarçadas

Imagine que você tem um trabalho peculiar em uma fábrica peculiar: sua tarefa é pegar objetos de uma esteira transportadora misteriosa e separá-los em duas caixas. Quando você chega, Susan, a Classificadora Sênior, explica que os objetos azuis em forma de ovo são chamados de "bleggs" e vão para a "caixa de bleggs", enquanto os cubos vermelhos são chamados de "rubes" e vão para a "caixa de rubes".

Assim que você começa a trabalhar, percebe que os bleggs e rubes diferem de outras maneiras além da cor e forma. Os bleggs têm pêlos em sua superfície, enquanto os rubes são lisos. Os bleggs são levemente flexíveis ao toque, enquanto os rubes são rígidos. Os bleggs são opacos, enquanto os rubes têm uma superfície ligeiramente translúcida.

Logo após começar a trabalhar, você encontra um blegg com um tom incomum de azul-escuro — na verdade, em um exame mais atento, a cor revela-se roxa, a meio caminho entre o vermelho e o azul.

No entanto, espere! Por que você está chamando esse objeto de "blegg"? Originalmente, um "blegg" era definido como azul e em forma de ovo — de fato, a qualificação do azul está incluso no próprio nome "blegg". Esse objeto não é azul. Falta uma das qualificações necessárias; você deveria chamá-lo de "objeto roxo em forma de ovo" em vez de "blegg".

Mas acontece que, além de ser roxo e em forma de ovo, o objeto também é peludo, flexível e opaco. Então, quando você viu o objeto, pensou: "Ah, um blegg de cor estranha" Certamente não é um rube… correto?

Ainda assim, você não tem certeza do que fazer em seguida. Então, você chama Susan, a Classificadora Sênior.

"Ah, sim, é um blegg", diz Susan, "você pode colocá-lo na caixa de bleggs".

Você começa a jogar o blegg roxo na caixa de bleggs, mas para por um momento. "Susan" você diz, "como você sabe que isso é um blegg?"

Susan olha para você estranhamente. "Não é óbvio? Este objeto pode ser roxo, mas ainda tem a forma de um ovo, é peludo, flexível e opaco, assim como todos os outros bleggs. Alguns defeitos de cor devem ser esperados. Ou será que isso é um daqueles enigmas filosóficos, como "Como você sabe que o mundo não foi criado há cinco minutos completamente com falsas memórias?" Em um sentido filosófico, não tenho certeza absoluta de que esse é um blegg, mas parece um bom palpite."

"Não, quero dizer…." Você faz uma pausa, procurando as palavras certas. "Por que existe uma caixa para bleggs e uma caixa para rubes? Qual é a diferença entre bleggs e rubes?"

"Bleggs são azuis e em forma de ovo, rubes são vermelhos e em forma de cubo", diz Susan pacientemente. "Você assistiu a palestra de orientação padrão, certo?"

"Por que é necessário separar os bleggs e rubes?"

"Bem… caso contrário, eles estariam todos misturados", responde Susan. "Afinal, ninguém nos pagaria para passar o dia inteiro sem separar os bleggs e rubes, não é?"

"Quem originalmente determinou que o primeiro objeto azul em forma de ovo era chamado de 'ble-

gg' e como eles chegaram a essa conclusão?"

Susan dá de ombros. "Suponho que você poderia simplesmente chamar os objetos vermelhos em forma de cubo de 'bleggs' e os objetos azuis em forma de ovo de 'rubes', mas parece mais fácil de lembrar dessa maneira."

Você pondera por um momento. "Suponha que um objeto completamente misturado saia da esteira. Por exemplo, um objeto translúcido, peludo, em forma de esfera laranja com tentáculos verdes contorcidos. Como eu poderia saber se é um blegg ou um rube?"

"Uau, nunca encontramos um objeto tão misturado", diz Susan, "mas acredito que devemos levá-lo para o scanner de triagem."

"Como funciona o scanner de triagem?", você pergunta. "Usa raios-x? Ressonância magnética? Espectroscopia de nêutrons de transmissão rápida?"

"Fui informada de que ele opera com base na Regra de Bayes, mas não entendo completamente como", responde Susan. "No entanto, adoro mencionar isso: Bayes, Bayes, Bayes, Bayes, Bayes."

"E o que o scanner de triagem mostra?"

"Ele indica se devemos colocar o objeto na caixa de bleggs ou na caixa de rubes. É por isso que ele é chamado de scanner de triagem."

Nesse momento, você permanece em silêncio.

"Aliás", diz Susan casualmente, "pode ser interessante para você saber que os bleggs contêm pequenas pepitas de minério de vanádio e os rubes contêm fragmentos de paládio, ambos com valor industrial."

"Susan, você é pura maldade."

"Obrigado."

Agora parece que descobrimos a essência e o cerne do conceito de ser um blegg: um blegg é um objeto que contém uma pepita de minério de vanádio. As características superficiais, como a cor azul e a pelagem, não determinam se um objeto é um blegg; essas características só importam porque ajudam a inferir se um objeto é um blegg, ou seja, se o objeto contém vanádio.

Conter vanádio é uma definição necessária e suficiente: todos os bleggs contêm vanádio e tudo que contém vanádio é um blegg: "blegg" é apenas uma forma abreviada de dizer "objeto que contém vanádio". Certo?

Não tão rápido, adverte Susan: cerca de 98% dos bleggs contêm vanádio, mas 2% contêm paládio. Para ser mais preciso (continua Susan), cerca de 98% dos objetos opacos, flexíveis, em forma de ovo azul contêm vanádio. Para bleggs incomuns, a porcentagem pode ser diferente: 95% dos bleggs roxos contêm vanádio, 92% dos bleggs duros contêm vanádio, e assim por diante.

Agora, imagine que você encontre um objeto opaco flexível, com pelos azuis, em forma de ovo. À primeira vista, parece um blegg em todos os aspectos visíveis. Por pura diversão, você decide levá-lo para o scanner de triagem, e o scanner revela "paládio" — uma ocorrência rara de apenas 2%. Será que é um blegg?

Inicialmente, você pode pensar em chamá-lo de "rube", afinal, você está prestes a jogá-lo na caixa de rubes. Entretanto, descobre-se que a maioria dos bleggs emite um leve brilho no escuro quando as luzes são apagadas, enquanto a maioria dos rubes não brilha. Além disso, a proporção de bleggs que brilham no escuro não difere significativamente entre os objetos opacos flexíveis em forma de ovo azul que contêm paládio, em comparação com aqueles que contêm vanádio. Portanto, se você quiser especular se o objeto brilha como um blegg ou permanece escuro como um rube, seria razoável supor que ele brilhe como um blegg.

Então, o objeto é realmente um blegg ou um rube?

Por um lado, você jogará o objeto na caixa de rubes, independentemente do que mais descobrir sobre ele. Por outro lado, se houver características desconhecidas do objeto que você precise inferir, você as

inferirá considerando-o como um blegg, não como um rube — agrupando-o no conjunto de objetos opacos flexíveis em forma de ovo azul, e não no conjunto de objetos translúcidos, duros e lisos em forma de cubo vermelho.

A pergunta “Este objeto é um blegg?” pode ter diferentes interpretações em diferentes ocasiões.

Se não estivesse respondendo alguma pergunta, você não teria motivos para se preocupar.

O ateísmo pode ser considerado uma “religião”? O transumanismo pode ser chamado de “seita”? O argumento de que o ateísmo é uma religião “porque afirma crenças sobre Deus” é uma tentativa de equiparar os métodos de raciocínio do ateísmo aos da religião, ou que o ateísmo não é mais seguro do que a religião em termos de probabilidade de gerar violência causal, entre outras coisas.

O que está realmente em jogo é a reinvindicação dos ateus de serem substancialmente diferentes e superiores em relação à religião, algo que as pessoas religiosas tentam negar, negando essa diferença, em vez de negar a superioridade (!).

No entanto, essa não é a parte irracional a priori. A parte irracional a priori se manifesta quando, durante a discussão, alguém recorre a um dicionário para buscar a definição de “ateísmo” ou “religião” (e sim, tanto os ateus quanto os religiosos cometem o mesmo erro). Como um dicionário poderia determinar se um grupo empírico de ateus é realmente substancialmente diferente de um grupo empírico de teólogos? Como a realidade pode variar de acordo com o significado de uma palavra? Os pontos no espaço das coisas não se movem quando redesenhamos um limite.

Muitas vezes, as pessoas não percebem que sua discussão sobre onde traçar um limite de definição é, na verdade, uma disputa sobre a inferência de uma característica compartilhada pela maioria das coisas em um agrupamento empírico…

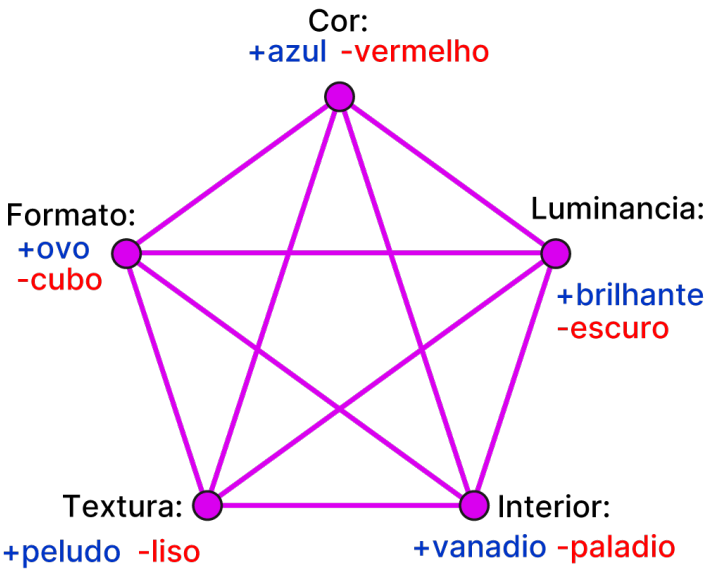
Daí surge a expressão “consulta disfarçada”.

161 — Categorias neurais

Em “Consultas disfarçadas”, abordei uma tarefa de classificação entre “bleggs” e “rubes”. Um blegg típico possui as seguintes características: é azul, tem forma de ovo, é peludo, flexível, opaco, brilha no escuro e contém vanádio. Por sua vez, um rube típico é vermelho, tem forma de cubo, é liso, duro, translúcido, não brilha e contém paládio. Para simplificar, deixemos de lado as características de flexibilidade/dureza e opacidade/translucidez. Isso resulta em cinco dimensões no espaço das coisas: cor, forma, textura, luminância e interior.

Suponha que eu queira criar uma Rede Neural Artificial (ANN) para prever características de bleggs que não foram observadas, com base nas características que foram observadas. E suponha que eu seja bastante ingênuo em relação às ANNs: li alguns livros populares e empolgantes sobre como as redes neurais são distribuídas, emergentes e paralelas, assim como o cérebro humano!!! Mas não consigo derivar as equações diferenciais para a descida de gradiente em uma rede multicamada não recorrente com unidades sigmóides (o que, na verdade, é mais fácil do que parece).

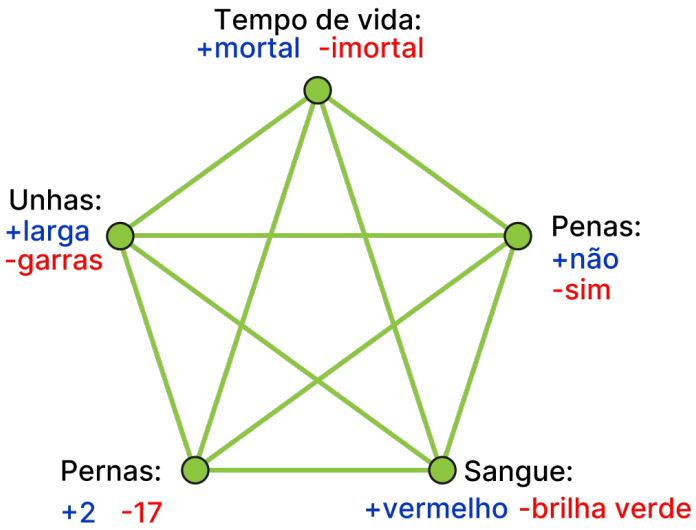
Nesse cenário, posso projetar uma rede neural que se assemelha à Rede 1:



A Rede 1 tem a finalidade de classificar bleggs e rubes. No entanto, como “blegg” é um conceito sintético e desconhecido, também incluí uma Rede 1b similar, para distinguir humanos de Monstros Espaciais, com informações de Aristóteles (“Todos os homens são mortais”) e da Academia de Platão (“Um bípede sem penas com unhas largas”).

Uma rede neural precisa de uma regra de aprendizado. A ideia óbvia é que, quando dois nodos costumam estar ativos ao mesmo tempo, devemos fortalecer a conexão entre eles — essa é uma das primeiras regras já propostas para o treinamento de uma rede neural, conhecida como Regra de Hebb.

Dessa forma, se você frequentemente vê coisas que são azuis e peludas — ativando simultaneamente o nodo “cor” no estado + e o nodo ”textura” no estado + – a conexão se fortaleceria entre cor e textura, fazendo com que cores + ativem texturas +, e vice-versa. Se você visse coisas que fossem azuis, em forma de ovo e contendo vanádio, isso fortaleceria conexões mútuas positivas entre cor, forma e interior.



Rede 1b

Digamos que você já tenha observado muitos bleggs e rubes passando pela esteira rolante. No entanto, de repente, você se depara com algo que é peludo, tem forma de ovo e — suspiro! — uma tonalidade arroxeada avermelhada (que modelaremos como um nível de ativação de “cor” de -2/3). Você ainda não avaliou a luminância ou o interior. O que prever, o que prever?

Então, o que acontece é que os níveis de ativação na Rede 1 oscilam um pouco. A ativação positiva flui da forma para a luminância, a ativação negativa flui da cor para o interior... Claro, todas essas mensagens são transmitidas em paralelo! E assincronamente! Assim como ocorre no cérebro humano...

Finalmente, a Rede 1 se estabelece em um estado estável, com alta ativação positiva para “luminância“ e “interior”. Pode-se dizer que a rede “espera“ (embora ainda não tenha visto) que o objeto brilhe no escuro e contenha vanádio.

Surpreendentemente, a Rede 1 exibe esse comportamento mesmo que não haja nenhum nó explícito que indique se o objeto é um blegg ou não. O julgamento está implícito em toda a rede! O “ser blegg“ é um atrator! Surge como resultado do comportamento emergente! Distribuído! E da regra de aprendizado.

Na prática, projetos de redes neurais como o descrito anteriormente - por mais que pareçam um modismo - enfrentam diversos desafios. Redes recorrentes nem sempre se estabilizam imediatamente: elas podem oscilar, exibir comportamento caótico ou levar muito tempo para se estabilizar. Isso é problemático quando você se depara com um objeto grande, amarelo e listrado e precisa esperar cinco minutos para que sua rede neural distribuída se ajuste ao atrator “tigre“. Embora as redes sejam assíncronas e paralelas, elas podem não operar em tempo real.

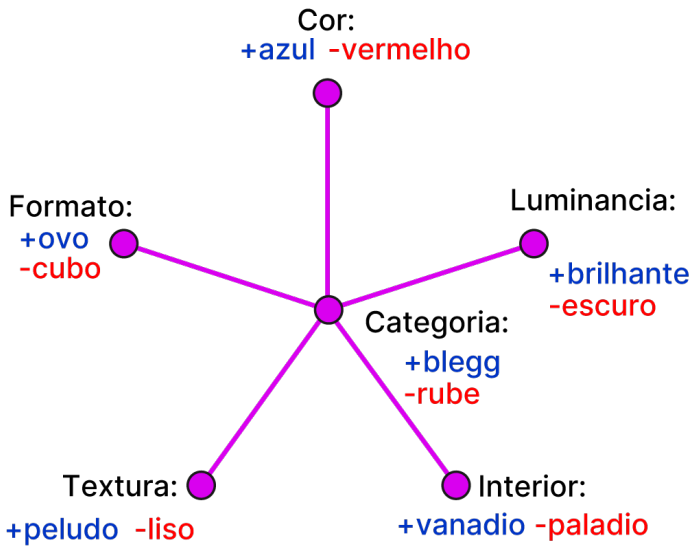
Outros problemas surgem, como contar duas vezes as evidências quando as mensagens são transmitidas de ida e volta. Por exemplo, se você suspeitar que um objeto brilha no escuro, essa suspeita ativará a crença de que o objeto contém vanádio, que por sua vez reforçará a crença de que o objeto brilha no escuro.

Além disso, ao tentar escalar o projeto da Rede 1, é necessário lidar com $O(N^2)$ conexões, onde N é o número total de observáveis. Diante disso, o que poderia ser um projeto de rede neural mais realista?

A Rede 2, na qual uma onda de ativação converge no nó central a partir de quaisquer nós (observados) e, em seguida, se propaga para quaisquer nós (não observados). Isso permite calcular a resposta em uma

etapa, em vez de esperar que a rede se estabilize - um requisito importante na biologia, onde os neurônios somente operam em frequências de 20 Hz. Adicionalmente, a arquitetura da Rede 2 escala de forma mais eficiente, sendo proporcional a O(N) em vez de $O(N^2)$.

Rede 2



É importante reconhecer que a primeira arquitetura de rede pode permitir a detecção de certas características de forma mais fácil do que a segunda. A Rede 1 possui uma conexão direta entre cada par de nós. Assim, se os objetos vermelhos nunca brilham no escuro, mas objetos peludos vermelhos geralmente possuem outras características blegg, como forma de ovo e vanádio, a Rede 1 pode representar isso facilmente, utilizando uma conexão direta negativa forte da cor para a luminância, mas conexões positivas mais fortes da textura para todos os outros nós, exceto a luminância.

Isso também não representa uma “exceção especial“ à regra geral de que os bleggs brilham. Lembre-se: na Rede 1, não há uma unidade que represente o brilho dos bleggs; o ser blegg surge como um atrator na rede distribuída.

Portanto, sim, essas conexões $O(N^2)$ estavam nos proporcionando algo. Mas não muito. A Rede 1 não é tão útil na maioria dos problemas reais, onde raramente encontramos um animal preso entre ser um gato e um cachorro.

(Também existem fatos que não podem ser facilmente representados na Rede 1 ou na Rede 2. Por exemplo, quando a cor azul-marinho e a forma esferoide são encontradas juntas, sempre indicam a presença de paládio. No entanto, quando encontradas individualmente, uma sem a outra, elas são evidências muito fortes para vanádio. Isso é difícil de representar em qualquer arquitetura sem nodos adicionais. Tanto a Rede 1 quanto a Rede 2 incorporam suposições implícitas sobre o tipo de estrutura ambiental que provavelmente existe. A capacidade de entender isso é o que separa os adultos dos bebês no aprendizado de máquina.)

Não se engane: nem a Rede 1, nem a Rede 2 são realistas do ponto de vista biológico. No entanto, ainda parece razoável supor que, independentemente de como o cérebro realmente funciona, ele está, em certo sentido, mais próximo da Rede 2 do que da Rede 1. É rápido, barato, escalonável e funciona bem para distinguir cães e gatos. A seleção natural se adapta a esse tipo de situação, assim como a água fluindo por uma paisagem de aptidão.

Classificar objetos como bleggs ou rubes e jogá-los nas caixas apropriadas parece uma tarefa bastante comum. Mas você notaria se os objetos azul-marinho nunca brilhassem no escuro?

Bem, talvez você notasse se lhe apresentassem vinte objetos que fossem iguais apenas em sua cor

azul-marinho e, em seguida, apagassem a luz, e nenhum dos objetos brilhasse. Se isso lhe causasse impacto, por assim dizer. Talvez, ao apresentar todos esses objetos azul-marinho como um grupo, seu cérebro forme uma nova subcategoria e identifique a característica “não brilha“ dentro dessa subcategoria. Mas provavelmente você não notaria se os objetos azul-marinho estivessem dispersos entre uma centena de outros bleggs e rubes. Não seria nada fácil nem intuitivo de perceber, ao contrário da distinção fácil e intuitiva entre cães e gatos.

Ou seja: “Sócrates é humano, todos os humanos são mortais, portanto, Sócrates é mortal“. Como Aristóteles sabia que Sócrates era humano? Ora, Sócrates não tinha penas, tinha unhas largas, andava ereto, falava grego e, bem, tinha geralmente a aparência e comportamento de um humano. Assim, o cérebro decide de uma vez por todas que Sócrates é humano e, a partir disso, infere que Sócrates é mortal, assim como todos os outros humanos observados dessa maneira. Não parece fácil nem intuitivo questionar em que medida vestir roupas, em contraposição ao uso da linguagem, está associado à mortalidade. Simplesmente, “coisas que vestem roupas e usam linguagem são humanas“ e “humanos são mortais“.

Existem vieses associados à tentativa de classificar as coisas em categorias definitivas? Certamente existem. Veja, por exemplo, o “Sectarismo anti-sectarista“.

162 — Como um algoritmo se sente por dentro

"Se uma árvore cai na floresta, mas ninguém ouve, ela faz barulho?" Certa vez, vi uma discussão real sobre esse assunto — uma discussão totalmente ingênua que não chegava nem perto do subjetivismo berkeliano[26]. Era apenas:

"Faz barulho, como qualquer outra árvore caindo!"

"Mas como pode haver um som que ninguém ouve?"

A visão racionalista padrão seria que a primeira pessoa está falando como se "som" significasse vibrações acústicas no ar; a segunda pessoa está falando como se "som" significasse uma experiência auditiva no cérebro. Se você perguntar: "Existem vibrações acústicas?" ou "Existem experiências auditivas?", a resposta é imediatamente óbvia. Assim, o argumento é realmente sobre a definição da palavra "som".

Acredito que a análise padrão está correta. Então, aceitando isso como premissa, pergunto: por que as pessoas entram em discussões como essa? Qual é a psicologia subjacente?

Uma ideia-chave do programa de heurística e vieses é que os erros geralmente revelam mais sobre a cognição do que as respostas corretas. Entrar em uma discussão acalorada sobre se uma árvore cai em uma floresta deserta e faz barulho é tradicionalmente considerado um erro.

Então, que tipo de projeto mental corresponde a esse erro?

No texto "Consultas Disfarçadas", apresentei a tarefa de classificação blegg/rube, na qual Susan, a Classificadora Sênior, explica que seu trabalho é classificar objetos que saem de uma esteira rolante, colocando os ovos azuis ou "bleggs" na caixa de bleggs e os cubos vermelhos ou "rubes" na caixa de rubes. Isso ocorre porque os bleggs contêm pequenas pepitas de minério de vanádio e os rubes contêm pequenos fragmentos de paládio, ambos úteis industrialmente.

Exceto que cerca de 2% dos objetos azuis em forma de ovo contêm paládio. Então, se você encontrar uma coisa azul em forma de ovo que contém paládio, você deveria chamá-la de "rube"? Você a colocaria na caixa de rubes — por que não chamá-la de "rube"?

Mas quando você desliga a luz, quase todos os bleggs brilham levemente no escuro.

E objetos azuis em forma de ovo que contêm paládio têm a mesma probabilidade de brilhar no escuro do que qualquer outro objeto azul em forma de ovo.

Assim sendo, caso você encontre um objeto azul em formato de ovo contendo paládio e pergunte "É um blegg?", a resposta dependerá do objetivo da sua pergunta: se a sua pergunta for "Em qual caixa o objeto vai?", você o considera como um rube. Porém, se a sua pergunta for "Se eu desligar a luz, ele brilhará?", você prevê como se o objeto fosse um blegg. Em um caso, a pergunta "É um blegg?" representa a pergunta disfarçada "Em qual caixa ele vai?". No outro caso, a pergunta "É um blegg?" representa a pergunta disfarçada "Irá

26 NT. O **subjetivismo berkeliano** refere-se à filosofia de George Berkeley (1685-1753), um empirista irlandês, que defende a ideia de que a realidade só existe na medida em que é percebida. Berkeley argumenta que "ser é ser percebido" (*esse est percipi*), ou seja, os objetos só têm existência enquanto são percebidos por uma mente. Ele rejeita a noção de matéria independente da percepção, afirmando que o mundo material é composto por ideias que existem na mente de Deus e nas mentes humanas. Essa visão é frequentemente associada ao **idealismo subjetivo**, pois coloca a experiência subjetiva como fundamento da realidade.

brilhar no escuro?“.

Agora, imagine que você tenha em mãos um objeto azul em formato de ovo contendo paládio, e que tenha observado que ele é peludo, flexível, opaco e brilha no escuro.

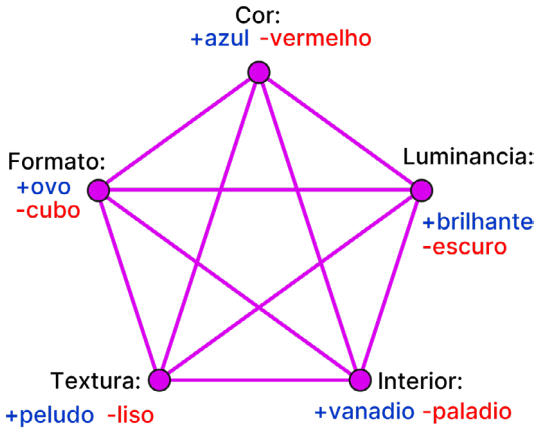
Isso responde a todas as perguntas, observa cada observável introduzido. Não há mais nenhuma pergunta disfarçada a ser feita.

Então, por que alguém ainda teria o impulso de continuar discutindo se esse objeto é realmente um blegg?’

Estes diagramas de “Categorias Neurais”mostram duas redes neurais diferentes que podem ser usadas para responder a perguntas sobre bleggs e rubes.

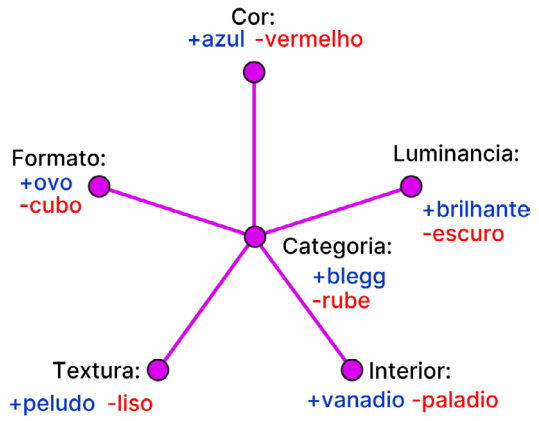
Rede 1

A Rede 1 tem várias desvantagens, como comportamento potencialmente oscilante/caótico ou a necessidade de $O(N^2)$ conexões - mas a estrutura da Rede 1 tem uma grande vantagem sobre a Rede 2: cada unidade na rede corresponde a uma consulta testável. Se você observar todos os observáveis, fixando todos os valores, não sobrarão unidades na rede.



Rede 2

Por outro lado, a Rede 2 é uma candidata muito melhor para se assemelhar vagamente ao funcionamento do cérebro humano: é rápida, barata, escalonável e tem uma unidade central extra pendente que pode variar sua ativação, mesmo após observar cada um dos nós circundantes.



Isso significa que mesmo após saber se um objeto é azul ou vermelho, ovo ou cubo, peludo ou liso,

brilhante ou escuro, e se contém vanádio ou paládio, parece haver uma pergunta sem resposta: mas é realmente um blegg?

Normalmente, em nossa experiência diária, as vibrações acústicas e a experiência auditiva andam juntas. No entanto, uma árvore caindo em uma floresta deserta desfaz essa associação comum. E mesmo depois que você sabe que a queda da árvore cria vibrações acústicas, mas não uma experiência auditiva, parece que ainda resta uma pergunta: fez algum barulho?

Sabemos onde Plutão está e para onde está indo; conhecemos a forma e a massa de Plutão - mas ele é um planeta?

Agora, lembre-se: quando você olha para a Rede 2, como descrevi aqui, você vê o algoritmo de fora. As pessoas não pensam consigo mesmas: “A unidade central deve disparar ou não?“ mais do que você pensa “Deve o neurônio #12.234.320.242 em meu córtex visual disparar ou não?“

É preciso um esforço deliberado para visualizar seu cérebro de fora. E mesmo assim, você ainda não vê seu cérebro real, apenas imagina o que pensa estar lá.Espera-se que com base na ciência, mas independente disso, você não tem nenhum acesso direto às estruturas de rede neural da introspecção. É por isso que os antigos gregos não inventaram a neurociência computacional.

Ao olhar para a Rede 2, você vê de fora, mas a sensação de como a estrutura da rede neural épor dentro, se você mesmo for um cérebro executando esse algoritmo, é que mesmo após conhecer todas as características do objeto, você ainda se pergunta: “Mas é um blegg, ou não?“

“Esta é uma grande lacuna a ser superada e eu já vi muitas pessoas pararem em seus caminhos por causa disso. O que acontece é que não reconhecemos intuitivamente nossas intuições como “intuições“, apenas as enxergamos como a realidade. Quando olhamos para um copo verde, não pensamos em nós mesmos como vendo uma imagem reconstruída em nosso córtex visual — embora seja isso que estejamos vendo — apenas enxergamos um copo verde. Pensamos: “Ah, olha só, esta xícara é verde“, não “A imagem em meu córtex visual desta xícara é verde“.

E da mesma forma, quando as pessoas discutem se a árvore que cai faz barulho ou se Plutão é um planeta, elas não se veem discutindo se uma categorização deveria estar ativa em suas redes neurais. Parece que ou a árvore faz barulho ou não.

Sabemos onde Plutão está e para onde está indo; conhecemos a forma de Plutão e sua massa — mas ele é um planeta? Sim, algumas pessoas dizem que isso é uma briga por definições — mas mesmo isso é uma perspectiva da Rede 2, porque estamos discutindo como a unidade central deve ser conectada. Se fôssemos uma mente construída conforme as linhas da Rede 1, não diríamos “Depende de como você define ‘planeta’“, mas apenas “Dado que conhecemos a órbita, forma e massa de Plutão, não há mais perguntas a serem feitas“. Ou melhor, é assim que sentiríamos - como se não houvesse mais dúvidas - se fôssemos uma mente construída conforme as linhas da Rede 1.

Antes de questionar suas intuições, você precisa perceber que o que sua mente está vendo é uma intuição - algum algoritmo cognitivo, visto de dentro - em vez de uma percepção direta de como as coisas realmente são.

Acredito que as pessoas se apegam às suas intuições não tanto porque acreditam que seus algoritmos cognitivos são perfeitamente confiáveis, mas porque não conseguem perceber suas intuições como a maneira como seus algoritmos cognitivos se parecem por dentro.

E assim, tudo o que você tenta dizer sobre como o algoritmo cognitivo nativo se desvia, acaba sendo contrastado com a percepção direta do modo como as coisas realmente são - e descartado como obviamente errado.”

163 — Definições em disputa

Já presenciei inúmeras conversas - inclusive aquelas supostamente sobre ciência cognitiva - que acabaram se transformando em disputas sobre definições. Tomando como exemplo a clássica pergunta "Se uma árvore cai em uma floresta e ninguém ouve, ela faz barulho?", a discussão frequentemente segue um caminho como este:

Albert: "Claro que faz barulho. Que tipo de pergunta tola é essa? Todas as vezes em que ouvi uma árvore caindo, ela produziu um som. Logo, suponho que outras árvores também façam barulho ao cair. Não acredito que o mundo mude quando não estou observando."

Barry: "Espere um minuto. Se ninguém ouve, como pode ser considerado um som?"

Nesse exemplo, Barry está discutindo com Albert devido a uma intuição genuinamente diferente sobre a constituição de um som. No entanto, há mais de uma maneira pela qual a Disputa Padrão pode começar. Barry pode ter motivos para rejeitar a conclusão de Albert. Ou talvez Barry seja um cético que, ao ouvir o argumento de Albert, reflexivamente o escrutinou em busca de possíveis falhas lógicas; e então, ao encontrar um contra-argumento, automaticamente, sem aplicar uma segunda camada de busca por um contra-contra-argumento; resultando no seu próprio argumento em uma posição oposta. Isso não requer que a intuição prévia de Barry — a intuição que ele teria se tivesse sido questionado antes de Albert falar — seja diferente da de Albert.

Bem, se Barry não tinha uma intuição diferente antes, certamente ele tem agora.

Albert: "O que você quer dizer com 'não há som'? As raízes da árvore estalam, o tronco desmorona e bate no chão. Isso gera vibrações que se propagam pelo solo e pelo ar. É aí que a energia da queda vai e se transforma em calor e som. Você está dizendo que, se as pessoas saírem da floresta, a árvore viola a Lei da Conservação de Energia?"

Barry: "Mas ninguém ouve nada. Se não há humanos na floresta, ou, para fins de argumentação, qualquer outra entidade com um sistema nervoso complexo capaz de 'ouvir', então ninguém ouve um som."

Albert e Barry apresentam argumentos que parecem apoiar suas respectivas posições, descrevendo de forma mais detalhada os pensamentos que levaram seus "detectores de som" a se ativarem ou permanecerem em silêncio. No entanto, até o momento, a conversa ainda se concentra na floresta e não nas definições. Além disso, observe que eles realmente não discordam de nada que ocorre na floresta.

Albert: "Esta é a discussão mais estúpida em que já me envolvi. Você é um completo idiota."

Barry: "Ah, é? Bem, parece que alguém jogou uma pá na sua cara para apagar um incêndio."

O insulto é proferido e aceito; agora, nenhuma das partes pode retroceder sem perder a dignidade. Tecnicamente, isso não faz parte do argumento, pois os racionalistas explicam tais comportamentos, mas é uma parte tão importante da Disputa Padrão que estou incluindo-a mesmo assim.

Albert: "A árvore produz vibrações acústicas. Por definição, isso é um som."

Barry: "Ninguém ouve nada. Por definição, isso não é um som."

A discussão começa a se deslocar para o foco nas definições. Sempre que sentir a tentação de usar

as palavras “por definição“ em um argumento que não seja puramente matemático, lembre-se de que tudo o que é verdadeiro “por definição“ é verdadeiro em todos os mundos possíveis e, portanto, observar sua verdade nunca pode restringir em qual mundo você vive.

Albert: “O microfone do meu computador pode gravar um som mesmo quando ninguém está por perto para ouvi-lo, armazená-lo em um arquivo, e é chamado ‘arquivo de som’. E o que é armazenado no arquivo são os padrões de vibrações no ar, não os padrões de disparos neurais no cérebro de alguém. ‘Som’ se refere a um padrão de vibrações.“

Albert apresenta um argumento que parece apoiar a ideia de que a palavra “som“ possui um significado específico. Esse tipo de questão difere daquela sobre se as vibrações acústicas ocorrem em uma floresta, mas essa mudança passa geralmente despercebida.

Barry: “Ah, é? Vejamos se o dicionário concorda com você.“

Há tantas coisas sobre as quais poderia estar curioso no cenário da queda da árvore. Eu poderia ir à floresta observar as árvores, ou aprender a derivar a equação da onda para mudanças na pressão do ar, ou examinar a anatomia do ouvido, ou estudar a neuroanatomia do córtex auditivo. Em vez de fazer qualquer uma dessas coisas, devo consultar um dicionário, aparentemente. Por quê? Os editores do dicionário são especialistas em botânica, especialistas em física ou especialistas em neurociência? Consultar uma enciclopédia faria sentido, mas por que um dicionário?

Albert: “Ah! Definição 2c no Merriam-Webster: ‘Som: energia mecânica radiante transmitida por ondas de pressão longitudinais em um meio material (como o ar).’“

Barry: “Ah! Definição 2b no Merriam-Webster: ‘Som: A sensação percebida pelo sentido da audição.’“

Albert e Barry, em uníssono: “Maldito dicionário! Isso não ajuda em nada!“

Os editores de dicionários são historiadores do uso da linguagem, não legisladores da linguagem. Os editores de dicionários encontram palavras em uso atual e, em seguida, escrevem palavras ao lado (de uma pequena parte) do que as pessoas parecem querer dizer com elas. Se houver mais de um uso, os editores incluem mais de uma definição.

Albert: “Veja, suponha que deixei um microfone na floresta e gravei o padrão das vibrações acústicas da árvore caindo. Se eu tocasse essa gravação para alguém, eles chamariam de ‘som’! Essa é a forma comum de uso! Não fique por aí inventando suas próprias definições malucas!“

Barry: “Em primeiro lugar, posso definir uma palavra da maneira que quiser, desde que a utilize consistentemente. Em segundo lugar, a definição que dei está no dicionário. Em terceiro lugar, quem lhe deu o direito de decidir o que é ou não de uso comum?“

Existem muitos erros de racionalidade na Disputa Padrão. Alguns deles eu já mencionei, e outros ainda serão abordados; assim como as possíveis soluções.

No entanto, por enquanto, gostaria apenas de destacar — tristemente — que Albert e Barry parecem concordar em quase todos os aspectos sobre o que realmente está acontecendo na floresta, no entanto, isso não parece gerar qualquer sentimento de concordância

Discutir definições é um caminho que leva a um beco sem saída; as pessoas não seguiriam esse caminho se soubessem desde o início onde ele leva. Se você perguntasse a Albert (ou Barry) por que ele continua discutindo, ele provavelmente responderia algo como: “Barry (ou Albert) está tentando introduzir sua própria definição desprezível de ‘som’ para apoiar seu ponto de vista ridículo, e estou aqui para defender a definição padrão“.

Mas vamos supor que eu volte no tempo antes do início da discussão:

(Eliezer aparece do nada em um veículo peculiar que se assemelha à máquina do tempo do filme original “A Máquina do Tempo“.)

Barry: “Uau! Um viajante do tempo!“

Eliezer: “Sou um viajante do futuro! Ouçam minhas palavras! Viajei de volta no tempo - cerca de quinze minutos...“

Albert: “Quinze minutos?“

Eliezer: “-para trazer-lhes esta mensagem!“

(Há uma pausa de confusão e expectativa.)

Eliezer: “Vocês acham que ‘som’ deveria ser definido como exigindo tanto vibrações acústicas (ondas de pressão no ar) quanto experiências auditivas (alguém ouvindo o som), ou ‘som’ deveria ser definido como significando apenas vibrações acústicas, ou apenas experiência auditiva?“

Barry: “Você voltou no tempo para nos perguntar isso?“

Eliezer: “Meus propósitos são meus! Respondam!“

Albert: “Bem... Não vejo por que isso importaria. Você pode escolher qualquer definição, desde que a utilize consistentemente.“

Barry: “Jogue uma moeda. Quer dizer, jogue uma moeda duas vezes.“

Eliezer: “Pessoalmente, eu diria que, se o problema surgir, ambos os lados devem começar a descrever o evento em termos inequívocos de nível inferior, como vibrações acústicas ou experiências auditivas. Ou cada lado pode criar uma nova palavra, como ‘“alberzle“ e “bargulum“, para usar em vez do que costumavam chamar de ‘som’. Dessa forma, ambos os lados podem usar as novas palavras consistentemente, sem precisar recuar ou perder a face, mas ainda conseguem se comunicar. E, é claro, durante todo o debate, deve-se tentar manter em mente alguma proposição testável sobre a qual o argumento realmente trata. Isso parece bom para vocês?“

Albert: “Suponho que sim...“

Barry: “Por que estamos discutindo isso?“

Eliezer: “Para preservar sua amizade contra uma contingência que vocês, agora, nunca saberão. Pois o futuro já foi alterado!“

(Eliezer e a máquina desaparecem em uma nuvem de fumaça.)

Barry: “Onde estávamos mesmo?“

Albert: “Ah, sim: se uma árvore cai na floresta e ninguém ouve, ela faz barulho?“

Barry: “Ela faz um ‘alberzle’, mas não um ‘bargulum’. Qual é a próxima pergunta?“

Essa solução não resolve todas as disputas sobre categorizações. Mas destroi uma fração significativa delas.

164 — Sinta o significado

Quando ouço alguém dizer: “Oh, olhe, uma borboleta” os fonemas falados ”borboleta“ entram em meu ouvido e vibram em meu tímpano, sendo transmitidos para a cóclea, fazendo cócegas nos nervos auditivos que transmitem picos de ativação para o córtex auditivo, onde começa o processamento dos fonemas, com o reconhecimento das palavras e a construção da estrutura sintática (um processo nada serial), e todos os tipos de outras complicações.

Mas, no final do dia, ou melhor, no final do segundo, estou pronto para olhar na direção apontada pelo meu amigo e ver um padrão visual que reconhecerei como uma borboleta; e ficaria bastante surpreso se visse um lobo no lugar dela.

Meu amigo olha para uma borboleta, sua voz vibra e seus lábios se movem, as ondas sonoras viajam pelo ar de forma imperceptível, meu ouvido captura os sons e meus nervos os transmitem ao cérebro, que os reconstrói, e veja só, eu sei o que meu amigo está olhando. Não é incrível? Se não soubéssemos sobre as ondas de pressão no ar, seria uma descoberta sensacional em todos os jornais: os seres humanos são telepatas! Os cérebros humanos conseguem transferir pensamentos uns para os outros!

Bem, somos telepatas, de fato. No entanto, a magia perde o encanto quando se torna apenas uma realidade comum e todos os nossos amigos também são capazes disso.

Você acha que a telepatia é simples? Tente construir um computador que seja telepático com você. A telepatia, ou “linguagem”, ou qualquer nome que queira dar para a nossa capacidade de transferir parcialmente nossos pensamentos é mais complicada do que parece.

Seria bastante inconveniente andar por aí pensando: “Agora vou transformar parcialmente algumas características dos meus pensamentos em uma sequência linear de sons que evocarão pensamentos semelhantes no meu interlocutor...”

Portanto, o cérebro esconde a complexidade — ou melhor, nunca a representa de fato — levando as pessoas a terem ideias peculiares sobre as palavras.

Como mencionei anteriormente, quando um objeto grande listrado e amarelo salta em minha direção, penso imediatamente: “Caramba! Um tigre!“, e não “Mm... objetos grandes, amarelos e listrados costumavam ter as propriedades de serem famintos e perigosos, então, embora não seja logicamente necessário, auughhhh créc créc (engole seco).

Da mesma forma, quando alguém grita “Caramba! Um tigre!“, a seleção natural não favoreceria um organismo que pensasse: “Mm... acabei de ouvir as sílabas “tie“ e “grr“ que meus colegas associam com seus próprios conceitos internos de tigre, e eles são mais propensos a pronunciar esses sons quando veem um objeto que categorizam como aiiieeee créc créc socorro, meu braço está sendo esmagado (engole seco).“

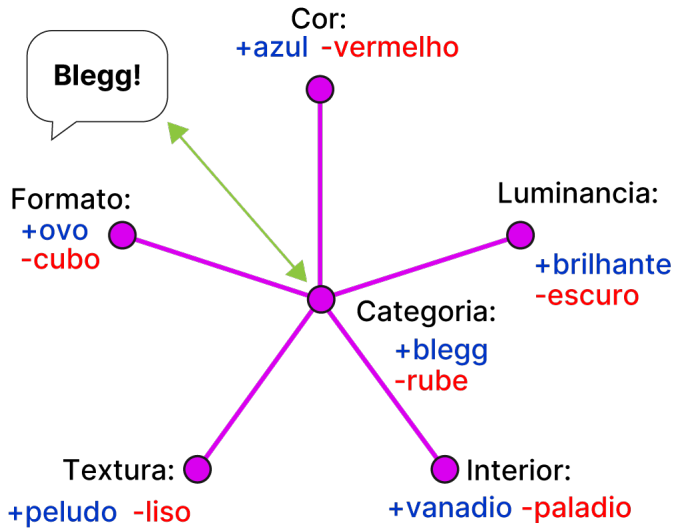
Considerando isso como uma limitação do projeto da arquitetura cognitiva humana, você não gostaria de ter etapas adicionais entre o momento em que seu córtex auditivo reconhece as sílabas “tigre“ e o momento em que o conceito de tigre é ativado.

Voltando à parábola dos bleggs e rubes, e à rede centralizada que categoriza de forma rápida e econômica, podemos imaginar uma conexão direta da unidade que reconhece a sílaba “blegg“ à unidade central da rede de bleggs. A unidade central, o conceito de blegg, é ativada quase que instantaneamente quando

você ouve Susan, a classificadora sênior, dizer: “Blegg!“

Para fins de discussão — que também não deveriam levar uma eternidade — assim que você vê um objeto azul em forma de ovo e a unidade central de blegg é ativada, você exclama “Blegg!“ para Susan.

O que esse algoritmo parece por dentro é que o rótulo e o conceito estão quase que inseparáveis; o significado parece uma propriedade intrínseca da própria palavra.



Rede 3

Os conhecedores reconhecerão isso como um exemplo da “Falácia da Projeção da Mente“ de E. T. Jaynes. Parece que uma palavra possui um significado intrínseco, assim como a vermelhidão é uma propriedade de uma maçã vermelha ou o mistério é uma propriedade de um fenômeno misterioso.

De fato, na maioria das ocasiões, o cérebro não fará nenhuma distinção entre a palavra e o significado — apenas se preocupando em separar os dois enquanto aprende um novo idioma, talvez. E mesmo assim, você verá Susan apontando para uma coisa azul em forma de ovo e dizendo “Blegg!“ e pensará, o que será que “blegg” significa, e não, a qual categoria mental Susan associa ao rótulo auditivo “blegg“.

Considere, nessa perspectiva, a parte da Disputa Padrão de Definições nas quais as duas partes debatem o verdadeiro significado da palavra “som“ - assim como podem discutir se uma maçã específica é realmente vermelha ou verde:

Albert: “O microfone do meu computador pode gravar um som mesmo quando ninguém está presente para ouvi-lo, armazenando-o em um arquivo, chamado “arquivo de som“. E o que é armazenado no arquivo são os padrões de vibrações no ar, não os padrões de disparos neurais no cérebro de alguém. ‘Som’ se refere a um padrão de vibrações.“

Barry: “Ah, é? Vejamos se o dicionário concorda com você.“

Albert sente intuitivamente que a palavra “som” tem um significado e que esse significado está relacionado às vibrações acústicas. Da mesma forma, Albert sente que uma árvore caindo na floresta produz um som (em vez de causar um evento correspondente à categoria de som).

Barry também sente que:

```
som.significado == experiências auditivas

floresta.som == falso.
```

Em vez de:

```
meuCérebro.EncontrarConceito("som")== conceito_ExperiênciaAuditiva

conceito_ExperiênciaAuditiva.combinar(floresta) == falso.
```

Isso se aproxima mais do que está realmente acontecendo, mas os humanos não evoluíram para saber disso, assim como não sabem instintivamente que o cérebro é composto de neurônios. As intuições conflitantes de Albert e Barry fornecem o combustível para continuar a discussão na fase de debate sobre o significado da palavra "som" — o que parece uma discussão sobre um fato como qualquer outro, como discutir se o céu é azul ou verde.

Você pode nem perceber que algo se perdeu até tentar realizar o ritual racionalista de estabelecer um experimento testável cujo resultado depende dos fatos que você está contestando tão fervorosamente…

165 — O argumento do uso comum

Parte da Disputa de Definição Padrão ocorre da seguinte forma:

Albert: "Veja, suponha que deixei um microfone na floresta e gravei o padrão das vibrações acústicas da árvore caindo. Se eu tocasse essa gravação para alguém, essa pessoa chamaria de 'som'! Essa é a forma comum de uso! Não fique inventando suas próprias definições malucas!"

Barry: "Em primeiro lugar, posso definir uma palavra da maneira que quiser, desde que a utilize consistentemente. Em segundo lugar, a definição que dei está no dicionário. Em terceiro lugar, quem lhe deu o direito de decidir o que é ou não de uso comum?"

Nem todas as disputas de definição chegam ao ponto de reconhecer o conceito de uso comum. Mais frequentemente, eu acredito, alguém recorre a um dicionário porque acredita que as palavras têm significados, e que o dicionário registra fielmente esses significados. Algumas pessoas até acreditam que o dicionário determina o significado — que os editores do dicionário são os legisladores da linguagem. Talvez porque, na escola primária, seus professores autoritários tenham dito que eles deveriam obedecer ao dicionário, que era uma regra obrigatória e não opcional.

Os editores de dicionário leem o que outras pessoas escrevem e registram o que as palavras parecem significar; eles são historiadores. O *Oxford English Dictionary* (Dicionário Oxford de Inglês) pode ser abrangente, mas nunca autoritário.

Mas certamente há um imperativo social para usar palavras de uma forma comumente compreendida? Será que nossa valiosa habilidade de linguagem, nossa telepatia humana, depende da coordenação mútua para funcionar? Talvez devêssemos tratar voluntariamente os editores de dicionários como árbitros supremos — mesmo que eles prefiram se considerar historiadores — para manter a silenciosa cooperação da qual todo discurso depende.

A expressão "dicionário autoritativo" é raramente usada corretamente, sendo um exemplo apropriado o The Authoritative Dictionary of IEEE Standards Terms (Dicionário Autorizado de Termos de Padrões do IEEE). O IEEE é um órgão composto por membros votantes, com a necessidade profissional de um acordo preciso sobre os termos e definições, e o Dicionário Autorizado de Termos de Padrões do IEEE é uma legislação real e negociada, exercendo qualquer autoridade que se considere presente no IEEE.

Na vida cotidiana, a linguagem compartilhada geralmente não surge de um acordo deliberado, como no caso do IEEE. É mais uma questão de infecção, pois as palavras são inventadas e difundidas pela cultura. (Pode-se dizer que é um "meme", seguindo Richard Dawkins, que introduziu o termo há quarenta anos — mas você já entende o que quero dizer, e se não, pode pesquisar no Google, assim também terá sido infectado.)

No entanto, como o exemplo do IEEE mostra, o acordo sobre a linguagem também pode ser um bem público estabelecido cooperativamente. Se você e eu desejamos trocar pensamentos através da linguagem, a telepatia humana, então é do nosso interesse mútuo usar a mesma palavra para conceitos semelhantes — preferencialmente, conceitos semelhantes ao limite de resolução da representação do nosso cérebro — mesmo que não tenhamos interesse mútuo óbvio em usar uma palavra específica para um conceito.

Não temos interesse mútuo óbvio em usar a palavra "oto" para significar "som" ou "som" para significar "oto", mas temos interesse mútuo em usar a mesma palavra, qualquer que seja. (De preferência, as palavras que usamos com frequência devem ser curtas, mas não entraremos na teoria da informação ainda.)

Mas, embora tenhamos um interesse mútuo, não é estritamente necessário que você e eu usemos rótulos semelhantes internamente; é apenas conveniente. Se sei que, para você, "oto" significa "som" — ou seja, você associa "oto" a um conceito muito semelhante ao que eu associo com "som" — então posso dizer "Papel amassado faz um oto crepitante". Isso exige um pensamento extra, mas posso fazê-lo se quiser.

Da mesma forma, se você disser: "O que é o 'bengala' de uma bola de boliche caindo no chão?" E eu souber qual conceito você associa às sílabas "bengala", então posso entender o que você quer dizer. Isso pode exigir alguma reflexão e pode me dar uma pausa, porque normalmente associo "bengala" a um conceito diferente. Mas posso fazer isso muito bem.

Quando os humanos realmente querem se comunicar entre si, é difícil impedi-los! Se estivermos presos em uma ilha deserta sem uma linguagem comum, pegaremos gravetos e faremos desenhos na areia.

O apelo de Albert ao argumento do uso comum assume que o acordo sobre a linguagem é um bem público estabelecido cooperativamente. No entanto, Albert assume isso apenas com o propósito de acusar retoricamente Barry de quebrar o acordo e colocar em risco esse bem público. Agora, o argumento sobre a queda da árvore passou da botânica para a semântica e a política; e então Barry responde desafiando a autoridade de Albert para definir a palavra.

Um racionalista, com a disciplina de abraçar a abordagem ativa, perceberia que a conversa se desviou significativamente.

Caro leitor, será que tudo isso é realmente necessário? Albert sabe o que Barry quer dizer com "som". Barry sabe o que Albert quer dizer com "som". Ambos têm acesso a palavras como "vibrações acústicas" ou "experiência auditiva", que já associam aos mesmos conceitos e que podem descrever eventos na floresta sem ambiguidade. Se estivessem presos em uma ilha deserta, tentando se comunicar, já teriam cumprido sua tarefa.

Quando ambos os lados sabem o que o outro quer dizer e acusam mutuamente de abandonar o "uso comum", então é claro que não estão encontrando uma maneira de se comunicar. Mas esse é o único benefício que o uso comum oferece desde o início.

Por que discutir o significado de uma palavra, com ambas as partes tentando se sobressair? Se for apenas um conflito de terminologia que ganhou uma proporção exagerada e não há mais nada em jogo, ambas as partes precisam simplesmente criar duas novas palavras e usá-las de maneira consistente.

Ainda assim, muitas vezes as categorizações funcionam como inferências implícitas e consultas disfarçadas. O ateísmo é uma "religião"? Se alguém argumenta que os métodos de raciocínio usados no ateísmo são equivalentes aos usados no judaísmo, ou que o ateísmo é comparável ao Islã em termos de violência causada, então essa pessoa tem um claro interesse argumentativo em agrupar tudo isso em um emaranhado confuso de "fé".

Ou considere a luta para unir negros e brancos como "pessoas". Este não seria o momento para criar duas palavras, pois o que está em jogo é exatamente a ideia de que não se deve fazer uma distinção moral.

Porém, ao tratar de qualquer proposição empírica ou moral, não se pode mais apelar ao uso comum.

Se a questão é como agrupar coisas semelhantes para fins de inferência, as previsões empíricas dependerão da resposta, o que significa que as definições podem estar erradas. Um conflito de previsões não pode ser resolvido por meio de uma pesquisa de opinião.

Se você deseja saber se o ateísmo deve ser agrupado com religiões sobrenaturalistas para fins de uma inferência empírica específica, o dicionário não pode responder.

Se você deseja saber se "negro" significa "pessoa", o dicionário não pode responder.

Se todos acreditarem que a luz vermelha no céu é Marte, o Deus da Guerra, o dicionário definirá "Marte" como o Deus da Guerra. Se todos acreditarem que o fogo é a liberação do flogisto, o dicionário definirá "fogo" como a liberação do flogisto.

Há uma arte em usar as palavras; mesmo quando as definições não são literalmente verdadeiras ou

falsas, muitas vezes são mais sábias ou tolas. Dicionários são apenas registros do uso no passado; se você os considerar como árbitros supremos de significado, ficará preso à sabedoria do passado, impedindo-o de fazer melhor.

No entanto, tenha o cuidado de garantir (se tiver que se afastar da sabedoria do passado) que as pessoas compreendam o que você está tentando transmitir.

166 — Rótulos vazios

Considere (mais uma vez) a ideia aristotélica de categorias. Suponhamos que haja um objeto com propriedades A, B, C, D e E, ou pelo menos que pareça ter a propriedade “E”.

Fred: “Você quer dizer que aquela coisa ali é azul, redonda, felpuda e...”

Eu: “Na lógica aristotélica, não importa quais são as propriedades ou como eu as chamo. É por isso que estou usando apenas as letras.“

Em seguida, eu invento a categoria aristotélica “zawa”, que descreve aqueles objetos, todos eles e somente eles, que possuem as propriedades A, C e D.

Eu: “O objeto 1 é zawa, B e E.”

Fred: “E é azul — quer dizer, A — também, certo?”

Eu: “Isso está implícito quando digo que é zawa.“

Fred: “Ainda assim, eu gostaria que você dissesse explicitamente.”

Eu: “Certo. O objeto 1 é A, B, zawa e E.“

Em seguida, adiciono outra palavra, “yokie”, que descreve todos e somente os objetos que são B e E, e a palavra “xippo”, que descreve todos e somente os objetos que são E, mas não D.

Eu: “O objeto 1 é zawa e yokie, mas não xippo.”

Fred: “Espera, é luminescente? Quero dizer, é E?“

Eu: “Sim. Essa é a única possibilidade com base nas informações fornecidas.“

Fred: “Eu prefiro que você explique.”

Eu: “Tudo bem: o objeto 1 é A, zawa, B, yokie, C, D, E e não xippo.“

Fred: “Incrível! Você pode dizer tudo isso só de olhar?“

“Impressionante, não é? Inventemos mais palavras novas: “Bolo” é A, C e yokie; “mun” é A, C e xippo; e “merlacdonian“ é bolo e mun.”

Inutilmente confuso? Eu também acho. Vamos substituir os rótulos pelas definições:

“‘Zawa, B e E’ se torna [A, C, D], B, E

“Bolo e A’ tornam-se [A, C, [B, E]], A

“Merlacdonian” torna-se [A, C, [B, E]], [A, C, [E, ¬D]].

E o que deve ser lembrado sobre a ideia aristotélica de categorias é que [A, C, D] é toda a informação contida em “zawa“. Não é apenas uma questão de poder variar o rótulo, mas também de poder dispensar rótulos por completo — as regras aristotélicas funcionam puramente com estruturas como [A, C, D]. Chamar uma dessas estruturas de “zawa” ou atribuir qualquer outro rótulo a ela é uma conveniência (ou inconveni-

ência) humana que não faz a menor diferença para as regras aristotélicas.

Suponhamos que “humano” seja definido como um ser bípede mortal sem penas. Então, o silogismo clássico teria a seguinte forma:

Todos os seres [mortal, ¬penas, bípede] são mortais.

Sócrates é um ser [mortal, ¬penas, bípede].

Portanto, Sócrates é mortal.

A habilidade de raciocinar parece muito menos impressionante agora, não é?

Aqui, a ilusão de inferência vem dos rótulos, que ocultam as premissas e fingem uma novidade na conclusão. Substituir os rótulos pelas definições revela a ilusão, tornando visível a inutilidade empírica da tautologia. Você nunca pode afirmar que Sócrates é um ser [mortal, ¬penas, bípede] até o ter observado como mortal.

Existe uma ideia, que você deve ter notado que detesto, de que “você pode definir uma palavra do jeito que quiser”. Essa ideia surgiu da noção aristotélica de categorias; já que, se você seguir as regras aristotélicas de forma precisa e impecável — algo que os humanos nunca fazem; Aristóteles sabia muito bem que Sócrates era humano, mesmo que isso não fosse justificado pelas suas regras — mas se alguma entidade imaginária não-humana seguir exatamente essas regras, ela nunca chegaria a uma contradição. No entanto, ela também não chegaria a muita coisa: ela não poderia afirmar que Sócrates é um ser [mortal, ¬penas, bípede] até observar sua mortalidade.

Não é tanto que os rótulos sejam arbitrários no sistema aristotélico, mas sim que o sistema aristotélico funciona perfeitamente sem nenhum rótulo — ele produz exatamente o mesmo fluxo de tautologias, apenas parecem muito menos impressionantes. Os rótulos existem apenas para criar a ilusão de inferência.

Portanto, se você for adotar um provérbio aristotélico, o provérbio deveria ser: não “Posso definir uma palavra do jeito que quiser”, nem mesmo “Definir uma palavra nunca tem consequências”, mas sim “As definições não precisam de palavras”.

167 — Jogando Tabu com as suas palavras

No jogo Tabu da Hasbro[27], o objetivo é fazer com que o seu parceiro adivinhe uma palavra escrita em um cartão, sem utilizar essa palavra ou outras cinco que estão listadas no cartão. Por exemplo, imagine que você tem que fazer o seu parceiro adivinhar a palavra "beisebol", mas não pode usar as palavras "esporte", "bastão", "rebate", "arremesso", "base" e, obviamente, "beisebol".

Quando me deparo com um problema como esse, imediatamente penso: "Um conflito de grupo artificial em que se usa um cilindro longo de madeira para bater em um esferoide lançado e, em seguida, correr para quatro posições seguras". Embora essa não seja a estratégia mais eficiente para transmitir a palavra 'beisebol' conforme as regras estabelecidas — pode ser: "É o que os Yankees jogam" -, a habilidade geral de apagar uma palavra da mente foi algo que pratiquei por anos, embora com um propósito diferente.

No ensaio anterior, vimos como a substituição de termos por definições pode revelar a improdutividade empírica do silogismo clássico aristotélico. Todos os seres humanos são mortais (e também, aparentemente, bípedes sem penas); Sócrates é humano; logo, Sócrates é mortal. Ao substituirmos a palavra 'humano' por sua definição aparente, o seguinte raciocínio subjacente é revelado:

Todos [mortais, sem penas, bípedes] são mortais;

Sócrates é um [mortal, sem penas, bípede];

Portanto, Sócrates é mortal.

No entanto, o princípio de substituir palavras por definições se aplica muito mais amplamente:

Albert: "Uma árvore caindo em uma floresta deserta faz um som."

Barry: "Uma árvore caindo em uma floresta deserta não faz um som."

À primeira vista, já que alguém diz "som" e outro diz "sem som", parece que temos uma contradição, certo? Mas suponha que ambos decompõem seus conceitos antes de falar:

Albert: "Uma árvore caindo em uma floresta deserta corresponde a [teste de associação: este evento gera vibrações acústicas]."

Barry: "Uma árvore caindo em uma floresta deserta não corresponde a [teste de associação: este evento gera experiências auditivas]."

Agora não há mais uma colisão aparente — tudo o que eles precisavam fazer era proibir o uso da palavra "som". Se as "vibrações acústicas" entrarem em disputa, bastava jogar Tabu novamente e dizer "ondas de pressão em um meio material"; se necessário, jogaríamos Tabu novamente na palavra "onda" e a substituiríamos pela equação da onda. (Jogue Tabu em "experiência auditiva" e você terá "Aquela forma de processamento sensorial, no cérebro humano, que recebe como entrada uma série temporal linear de mis-

27 NT. A **Hasbro** é uma empresa americana global de brinquedos, jogos de tabuleiro e entretenimento, fundada em 1923. Ela é conhecida por marcas icônicas como **Monopoly**, **Transformers**, **My Little Pony**, **Nerf**, **Play-Doh**, **Magic: The Gathering** e **Dungeons & Dragons**. A Hasbro atua em diversos setores, incluindo brinquedos, jogos, filmes, TV e games digitais. Sua missão é criar experiências envolventes que conectam pessoas por meio de brincadeiras e histórias.

turas de frequência… ")

Mas suponha, por outro lado, que Albert e Barry discutissem:

Albert: "Sócrates corresponde ao conceito [teste de adesão: esta pessoa vai morrer após beber cicuta]."

Barry: "Sócrates corresponde ao conceito [teste de associação: esta pessoa não morrerá após beber cicuta]."

Agora, Albert e Barry têm expectativas bastante diferentes, o que pode causar um choque entre eles. A diferença reside no que eles esperam que aconteça depois que Sócrates beber cicuta. No entanto, eles podem não perceber essa discrepância se usarem a mesma palavra "humano" para se referir a conceitos diferentes.

Você obtém uma imagem muito diferente sobre o que as pessoas concordam ou discordam, dependendo se você olha para o rótulo (Albert diz "som" e Barry diz "sem som", então eles devem discordar) ou se realiza um teste visual (para Albert, o teste de pertinência são as vibrações acústicas, enquanto para Barry é a experiência auditiva).

Reúna um grupo de futuristas autoproclamados e pergunte se eles acreditam que teremos inteligência artificial em trinta anos, e eu imagino que pelo menos metade deles dirá que sim. Se você deixar por isso mesmo, eles vão apertar as mãos e se parabenizar reciprocamente pelo consenso. No entanto, se você tornar o termo "inteligência artificial" um tabu e pedir que eles descrevam o que eles esperam ver, sem nunca usar palavras como "computadores" ou "pensar", poderá descobrir que há um grande conflito de expectativas escondido por trás desse termo comum e sem expressão. Confira também a compilação de 71 definições de "inteligência" de Shane Legg.

A ilusão de unidade entre as religiões pode ser dissipada, tornando o termo "Deus" um tabu e pedindo que as pessoas digam em que acreditam. Ou tornando a palavra "fé" um tabu e perguntando por que elas acreditam nisso. Embora, na maioria das vezes, elas não consigam responder porque é, principalmente, uma profissão de fé, e você não pode ampliar cognitivamente uma gravação de áudio.

Quando você se encontra em dificuldades filosóficas, a primeira linha de defesa não é definir seus termos problemáticos, mas ver se consegue pensar sem os usar. Ou sem usar nenhum de seus sinônimos curtos. E tome cuidado para não inventar uma nova palavra para usar em seu lugar. Descreva observáveis externos e mecanismos internos. Não use um único identificador, seja qual for.

Albert acredita que as pessoas têm livre-arbítrio, enquanto Barry acredita que elas não têm livre-arbítrio. É claro que isso pode parecer gerar um conflito. A maioria dos filósofos sugeriria que Albert e Barry tentassem definir exatamente o que eles querem dizer com "livre-arbítrio", o que certamente levaria a uma longa discussão. Eu aconselharia Albert e Barry a descreverem o que acreditam que as pessoas têm ou não têm, sem usar a expressão "livre-arbítrio". Se você tentar isso em casa, evite também palavras como "escolher", "agir", "decidir", "determinar", "responsável" ou quaisquer sinônimos dessas palavras.

Essa é uma das ferramentas não triviais que tenho em minha caixa de ferramentas, e, na minha modesta opinião, é mais eficaz do que as ferramentas padrão. No entanto, requer mais esforço para ser utilizada; você obterá o que pagou por ela.

168 — Substitua o símbolo pela substância

O que é necessário para — como no exemplo do ensaio anterior — ver um “jogo de beisebol” como “um conflito de grupo artificial no qual você usa um longo cilindro de madeira para golpear um esferoide arremessado e depois corre entre quatro posições seguras”? O que é preciso para jogar a versão racionalista do Tabu, na qual o objetivo não é encontrar um sinônimo que não esteja no cartão, mas encontrar uma maneira de descrever sem o identificador de conceito padrão?

Você precisa visualizar. Você precisa fazer com que sua mente perceba os detalhes, como se estivesse vendo pela primeira vez. É necessário ter uma Visão Original.

Isso é um “taco”? Não, é uma vara longa, redonda e afunilada feita de madeira, estreitando em uma das extremidades para um humano poder segurá-la e girá-la.

Isso é uma “bola”? Não, é um objeto esférico coberto de couro com costuras simétricas, sólido, mas não duro quanto metal, que pode ser segurado e lançado, ou acertado com a vara de madeira, ou agarrado.

São “bases”? Não, são posições fixas em um campo de jogo, para as quais os jogadores tentam correr o mais rápido possível para se manterem seguros, nas regras artificiais do jogo.

O principal desafio para alcançar uma visão original é que a mente já possui um resumo conveniente, um pequeno conceito simples e fácil de usar. Palavras como “beisebol”, “taco” ou “base”. É necessário fazer um esforço para impedir que a mente siga o caminho familiar, o caminho mais fácil, o caminho de menor resistência, onde palavras pequenas e sem características invadem e apagam os detalhes que você está tentando ver. Uma única palavra pode ter a força destrutiva do clichê; uma palavra pode carregar o veneno de um pensamento já armazenado.

Jogar o jogo do Tabu — conseguir descrever sem recorrer a palavras ou termos convencionais — é uma das habilidades fundamentais do pensamento racionalista. Isso está no mesmo nível primordial do hábito constante de fazer perguntas como “por quê?” ou “O que essa crença me leva a antecipar?”. A arte está intimamente relacionada a:

- Pragmatismo, pois essa forma de ver geralmente oferece uma conexão mais próxima com a experiência antecipada, em vez de uma crença puramente proposicional;
- Reducionismo, pois essa forma de ver geralmente nos leva a descer a um nível inferior de organização, a olhar para as partes em vez de saltar diretamente para o todo;
- Abraçar a pergunta, pois muitas vezes as palavras nos distraem da pergunta que realmente queremos fazer;
- Evitar pensamentos armazenados em cache, que tendem a recorrer ao uso de palavras convencionais, para que você possa bloqueá-los usando palavras-padrão como tabu;
- A regra do escritor de “Mostre, não conte!”, que tem poder entre os racionalistas;
- E não perder de vista o propósito original. Como tornar uma palavra tabu pode ajudar a manter seu propósito?

De “Propósitos Perdidos”:

Enquanto você lê isso, um jovem ou uma jovem está sentado em uma mesa em uma universidade, estudando seriamente um material que não tem intenção de usar e nenhum interesse em conhecer por si só. Eles querem um emprego bem remunerado, e o trabalho bem remunerado requer um pedaço de papel, e o

pedaço de papel requer um mestrado anterior, e o mestrado requer um diploma de bacharel, e a universidade que concede o diploma de bacharel exige que você faça um curso de padrões de tricô do século XII para se formar. Então, eles estudam diligentemente, com a intenção de esquecer tudo quando o exame final for aplicado, mas ainda trabalhando seriamente, porque eles querem aquele pedaço de papel.

Por que você frequenta a "escola"? Para obter uma "educação" que culmina em um "diploma". Ao eliminar as palavras proibidas e seus sinônimos óbvios, e visualizar os detalhes reais, é muito mais provável que você perceba que a "escola" atualmente consiste em sentar ao lado de adolescentes entediados, ouvindo material que você já conhece, que um "diploma" é apenas um pedaço de papel com algo escrito, e que essa "educação" consiste em esquecer o material assim que é testado.

As generalizações vazadas frequentemente se manifestam por meio de categorizações: as pessoas que realmente aprendem nas salas de aula, são rotuladas como "obtendo uma educação", então "obter uma educação" deve ser algo bom; no entanto, qualquer pessoa que compareça a uma faculdade também se enquadra no conceito de "obter educação", independentemente de estar aprendendo ou não.

Os alunos que têm conhecimento em matemática terão um bom desempenho nos testes, mas se você exigir que as escolas alcancem boas pontuações nesses testes, elas focarão todo o tempo em ensinar para o teste. Uma categorização mental que não se alinha perfeitamente com seu objetivo pode gerar a mesma falha de incentivo internamente. Se você deseja aprender, precisa de uma "educação"; e enquanto estiver obtendo algo que se enquadre na categoria de "educação", pode não perceber se realmente está aprendendo ou não. Ou você pode notar, mas não perceberá que perdeu de vista seu propósito original, porque está "obtendo uma educação" e foi assim que mentalmente descreveu seu objetivo.

Classificar é descartar informações. Se disserem a você que uma árvore caindo faz um "som", você não saberá qual é o som real, pois você não ouviu a árvore caindo. Se uma moeda cair com a face voltada para cima, você não saberá sua orientação exata. Um objeto azul em forma de ovo pode ser chamado de "blegg", mas e se a forma exata do ovo variar ou o tom exato de azul? Categorias são usadas para descartar informações irrelevantes, separar o ouro da poeira, mas frequentemente a categorização padrão acaba descartando informações relevantes também. E quando você se depara com esse tipo de problema mental, a solução mais óbvia é jogar Tabu.

Por exemplo, "Jogar Tabu" é em si uma generalização vazada. A versão da Hasbro não é a versão racionalista; eles listam apenas cinco palavras proibidas adicionais no cartão, e isso não é suficiente para excluir o pensamento em palavras antigas e familiares. O que os racionalistas fazem poderia ser considerado como jogar Tabu, mas nem tudo que conta como jogar Tabu funciona para forçar a visão original. Se você apenas pensar em "jogar Tabu para forçar a visão original", começará a acreditar que qualquer coisa que conte como jogar Tabu também conta como visão original.

A versão racionalista não é um jogo, o que significa que você não pode vencer tentando ser esperto e forçando as regras. Você precisa jogar Tabu com uma desvantagem voluntária: pare de usar sinônimos que não estão no cartão. Além disso, é necessário impedir-se de inventar uma nova palavra ou frase simples que funcione como um substituto mental equivalente à palavra antiga. O objetivo é ampliar seu mapa, não renomear as cidades; desvincular o ponteiro, não alocar um novo; ver os eventos como eles são, não reescrever o clichê com palavras diferentes.

Ao visualizar o problema com mais detalhes, é possível perceber o propósito perdido: o que exatamente você faz quando "joga Tabu"? Para que serve cada parte desse processo?

Se você olhar para suas atividades e situações originalmente, também poderá ver seus objetivos originais. Ao adotar uma nova perspectiva, como se fosse a primeira vez, você se encontrará fazendo coisas que nunca imaginaria fazer se não fossem hábitos arraigados.

O propósito é perdido sempre que a substância real (aprendizagem, conhecimento, saúde) é substituída pelo símbolo que a representa (um diploma, uma pontuação em um teste, assistência médica). Para restaurar um propósito perdido ou uma categorização com perdas, é necessário fazer o oposto:

Substituir o símbolo pela substância; substituir o significante pelo significado; substituir a propriedade pelo teste de associação; substituir a palavra pelo seu sentido; substituir o rótulo pelo conceito; substituir

o resumo pelos detalhes; substituir a pergunta substituta pela pergunta real; desvincular o ponteiro; descer para um nível inferior de organização; simular mentalmente o processo em vez de nomeá-lo; expandir seu mapa.

A Verdade Simples foi gerada a partir de um exercício dessa disciplina para descrever a "verdade" em um nível inferior de organização, sem recorrer a termos como "exato", "correto", "representar", "refletir", "semântico", "acreditar", "conhecimento", "mapa" ou "real". (E lembre-se de que o objetivo não é realmente jogar Tabu — a palavra "verdadeiro" aparece no texto, mas não é usada para definir a verdade. Seria uma palavra proibida no jogo da Hasbro, mas não estamos realmente jogando esse jogo. Pergunte a si mesmo se o documento cumpriu seu propósito, não se seguiu as regras.)

A própria Regra de Bayes descreve "evidências" na matemática pura, sem utilizar palavras como "implica", "significa", "suporta", "prova" ou "justifica". Tentar definir esses termos filosóficos levará apenas a um ciclo infinito de argumentações.

E então, há a palavra mais importante de todas, Tabu. Eu frequentemente advirto para ter cuidado ao usá-la em excesso ou até mesmo evitá-la em certos casos. Agora você sabe o verdadeiro motivo. Não é um assunto ruim para se pensar. No entanto, sua verdadeira compreensão é medida pela sua capacidade de descrever o que está fazendo e por que está fazendo, sem usar essa palavra ou qualquer um de seus sinônimos.

169 — Falácias da compressão

"O mapa não é o território", como diz o ditado. O único mapa 100% preciso da Califórnia, em seu tamanho real e com detalhes atômicos, é a própria Califórnia. No entanto, a Califórnia possui regularidades importantes, como o layout de suas rodovias, que podem ser descritas usando muito menos informações — sem mencionar muito menos material físico — do que seria necessário para descrever cada átomo nas fronteiras do estado. Daí o outro ditado: "O mapa não é o território, mas você não pode dobrar o território e colocá-lo no porta-luvas".

Um mapa de papel da Califórnia, na escala de 10 quilômetros para 1 centímetro (um milhão para um), não tem espaço para mostrar a posição distinta de duas folhas caídas a um centímetro de distância na calçada. Mesmo que o mapa tentasse mostrar as folhas, elas apareceriam como o mesmo ponto no mapa; ou melhor, o mapa precisaria ter uma dimensão de 10 nanômetros, que é uma resolução melhor do que a maioria das impressoras de livros e até mesmo do que os olhos humanos podem perceber.

A realidade é vasta — apenas a parte visível possui bilhões de anos-luz de diâmetro. Mas seu mapa da realidade está escrito em alguns quilos de neurônios, dobrados para caber dentro do seu crânio. Não quero ser ofensivo, mas seu crânio é minúsculo. Comparativamente falando.

Portanto, inevitavelmente, certas coisas distintas na realidade serão comprimidas no mesmo ponto do seu mapa.

Mas, a sensação interna não é a de dizer: "Oh, veja, estou comprimindo duas coisas em um ponto do meu mapa". A sensação interna é a de que há apenas uma coisa e você a está vendo.

Uma criança suficientemente pequena ou um filósofo grego suficientemente antigo, não saberiam que existem coisas como "vibrações acústicas" ou "experiências auditivas". Para eles, haveria apenas um único evento que ocorre quando uma árvore cai — um único evento chamado "som".

Perceber que existem dois eventos distintos, subjacentes a um ponto do seu mapa, é um desafio essencialmente científico — um desafio científico difícil e significativo.

Às vezes, falácias de compressão ocorrem devido à confusão de duas coisas conhecidas sob o mesmo rótulo — você conhece as vibrações acústicas e conhece o processamento auditivo no cérebro, mas chama ambas de "som" e, portanto, se confunde.

Mas a falácia de compressão mais perigosa ocorre quando nem se tem ideia de que duas entidades distintas sequer existem. Há apenas uma pasta mental em seu sistema de arquivamento, rotulada como "som", e tudo o que é pensado sobre "som" é colocado nessa pasta. Não é que haja duas pastas com a mesma etiqueta; existe apenas uma pasta. Por padrão, o mapa é comprimido. Por que o cérebro criaria dois compartimentos mentais quando um bastaria?

Ou pense em um romance de mistério no qual o insight crucial do detetive é que um dos suspeitos tem um gêmeo idêntico. No decorrer da investigação, o trabalho do detetive é simplesmente observar que Carol está vestindo vermelho, que ela tem cabelos pretos, que está usando sandálias de couro — mas todos esses são fatos sobre Carol. É fácil questionar um fato individual, como VesteVermelho(Carol) ou CabeloPreto(Carol). Talvez CabeloPreto(Carol) seja falso. Talvez Carol pinte o cabelo. Talvez seja CabeloCastanho(Carol). No entanto, é necessário um detetive mais perspicaz para questionar se a Carol em VesteVermelho(Carol) e CabeloPreto(Carol) — o arquivo mental de Carol onde as observações são registradas — deveria ser dividido

em dois arquivos. Talvez haja duas Carols, de modo que a Carol vestindo vermelho não seja a mesma pessoa que a Carol de cabelo preto.

Aqui é o próprio ato de criar dois compartimentos diferentes que é o golpe do insight genial. É mais fácil questionar os fatos do que a ontologia.

O mapa da realidade contido em um cérebro humano, ao contrário de um mapa da Califórnia em papel, pode se expandir dinamicamente quando escrevemos descrições mais detalhadas. Mas a sensação interna não é tanto de ampliar um mapa, mas dividir um átomo indivisível — pegar uma coisa (que parecia uma coisa) e dividi-la em duas ou mais coisas.

Isso muitas vezes se manifesta na criação de novas palavras, como "vibrações acústicas" e "experiências auditivas" em vez de simplesmente "som". Algo sobre a criação de um novo termo parece alocar um novo compartimento. O detetive pode começar a chamar um dos suspeitos de "Carol-2" ou "a Outra Carol" quase imediatamente após perceber que existem duas Carols.

No entanto, expandir o mapa nem sempre é tão simples como nomear novas cidades. É um golpe de percepção científica perceber que coisas como vibrações acústicas ou experiências auditivas existem.

Um exemplo óbvio nos dias modernos são palavras como "inteligência" ou "consciência". Às vezes, vemos comunicados de imprensa afirmando que um estudo de pesquisa "explicou a consciência" porque uma equipe de neurologistas investigou um ritmo elétrico de 40 Hz que pode estar relacionado à integração de informações sensoriais, ou porque investigaram o sistema reticular ativador responsável por manter os seres humanos acordados. Esse é um exemplo extremo, e as falhas mais comuns são mais sutis, mas seguem a mesma lógica. A parte da "consciência" que as pessoas acham mais interessante é a reflexividade, a autoconsciência, a percepção de que a pessoa que vemos no espelho somos "nós mesmos"; isso, juntamente com o problema complexo da experiência subjetiva, conforme destacado por David Chalmers. Também rotulamos de "consciente" o estado de estar acordado, em oposição a estar adormecido, em nosso ciclo diário. Todavia, todos esses são conceitos diferentes sob o mesmo nome, e os fenômenos subjacentes são quebra-cabeças científicos distintos. Você pode explicar estar acordado sem explicar a reflexividade ou a subjetividade.

Falácias de compressão também são fundamentais na técnica de isca e troca na filosofia — argumentar sobre "consciência" sob uma definição (como a capacidade de pensar sobre o pensamento) e, em seguida, aplicar as conclusões à "consciência" sob uma definição diferente (como a subjetividade). Claro, é possível que ambas sejam a mesma coisa, mas se assim for, compreender genuinamente esse fato exigiria primeiro uma divisão conceitual e, em seguida, um golpe de mestre de reunificação.

Expandir o seu mapa é (repito) um desafio científico: parte da arte da ciência, a habilidade de investigar o mundo. (E, é claro, você não pode resolver um desafio científico consultando dicionários, nem pode dominar uma habilidade complexa de investigação dizendo "posso definir uma palavra da maneira que eu quiser".) Quando você vê uma única coisa confusa, com atributos proteicos e autocontraditórios, é provável que seu mapa esteja acumulando muita coisa em um único ponto — você precisa separá-lo e alocar alguns novos compartimentos. Isso não é o mesmo que definir a única coisa que você vê, mas geralmente decorre da descoberta de como falar sobre a coisa sem recorrer a um único rótulo mental.

Portanto, a habilidade de explorar o mapa está ligada à versão racionalista do Tabu e ao uso sábio das palavras, pois as palavras representam geralmente os pontos em nosso mapa, os rótulos sob os quais arquivamos nossas proposições e os compartimentos nos quais organizamos nossas informações. Evitar o uso de uma única palavra, ou alocar novas palavras, geralmente faz parte da habilidade de expandir o mapa.

170 — A categorização tem consequências

Entre as inúmeras variações e mutações genéticas presentes em seu genoma, existem poucos alelos que você provavelmente conhece, incluindo aqueles que determinam seu tipo sanguíneo: a presença ou ausência dos antígenos A, B e o fator Rh. Caso você receba uma transfusão de sangue contendo um antígeno que você não possui, isso resultará em uma reação alérgica. Foi graças à descoberta feita por Karl Landsteiner sobre esse fato e de como testar a compatibilidade dos tipos sanguíneos que permitiu realizar transfusões de sangue sem colocar em risco a vida do paciente (Prêmio Nobel de Medicina em 1930.) Além disso, se uma mãe com sangue tipo A (por exemplo) tiver um filho com sangue tipo A+, a mãe pode adquirir uma reação alérgica ao antígeno Rh positivo; se ela tiver outro filho com sangue tipo A+, a criança estará em perigo, a menos que a mãe tome um supressor alérgico durante a gravidez. Portanto, as pessoas costumam conhecer seus tipos sanguíneos antes de se casarem.

Ah, e também: as pessoas com sangue tipo A são consideradas sérias e criativas, enquanto as pessoas com sangue tipo B são vistas como selvagens e alegres. Já as pessoas com sangue do tipo O são consideradas agradáveis e sociáveis, enquanto as do tipo AB são descritas como frias e controladas. (Você poderia pensar que o tipo O seria a ausência dos tipos A e B, enquanto o tipo AB seria apenas os tipos A mais B, mas não...) Tudo isso, conforme a teoria japonesa da personalidade com base no tipo sanguíneo.

Parece que, no Japão, o tipo sanguíneo desempenha um papel semelhante ao que os signos astrológicos desempenham no Ocidente, até mesmo com horóscopos baseados no tipo sanguíneo publicados diariamente nos jornais.

Essa tendência é particularmente intrigante, uma vez que os tipos sanguíneos nunca foram um mistério, nem no Japão, nem em qualquer outro lugar. Sabemos da existência dos tipos sanguíneos apenas graças a Karl Landsteiner. Nenhum curandeiro místico, nenhum feiticeiro venerável jamais mencionou uma palavra sobre os tipos sanguíneos; não existem pergaminhos antigos e empoeirados que revelam algum equívoco desde tempos remotos. Se a comunidade médica afirmasse amanhã que tudo isso foi apenas uma grande farsa, nós, leigos, não teríamos um fragmento de evidência para contradizê-los.

Nunca houve uma guerra entre os tipos sanguíneos. Nunca houve sequer um conflito político entre os tipos sanguíneos. Os estereótipos devem ter surgido unicamente da mera existência dos rótulos.

Agora, alguém certamente poderia argumentar que essa é uma história de categorização dos seres humanos. Será que o mesmo ocorre quando categorizamos plantas, pedras ou móveis de escritório? Não me lembro de ter lido sobre tal experimento, mas é claro que isso não significa que não tenha sido realizado. (Imagino que a principal dificuldade em realizar tal experimento seja encontrar um protocolo que não induza os participantes a pensar que, uma vez que um rótulo é atribuído, ele deve ter algum significado.) Portanto, embora eu não esteja sugerindo que devemos considerar evidências hipotéticas, prevejo um resultado positivo para esse experimento. Espero que descubramos que o simples ato de rotular possui influência sobre todas as coisas, pelo menos na imaginação humana.

É possível enxergar isso como agrupamentos de similaridade: assim que traçamos um limite em torno de um grupo, a mente começa a tentar colher similaridades dentro desse grupo. E, infelizmente, os detectores de padrões humanos parecem operar com tanta intensidade que enxergamos padrões, mesmo quando eles não estão presentes; uma correlação levemente negativa pode ser confundida com uma forte correlação positiva com um pouco de memória seletiva.

Isso pode ser visto como algoritmos neurais: atribuir um nome a um conjunto de coisas é como alocar uma sub-rede para encontrar padrões nelas.

Isso pode ser visto como uma falácia de compressão (tendência de simplificar excessivamente informações complexas): coisas com o mesmo nome são jogadas no mesmo compartimento mental, misturando-as em um único ponto no mapa.

Ou isso pode ser visto como a capacidade humana ilimitada de inventar coisas do nada e acreditar nelas, simplesmente porque ninguém pode provar que estão erradas. Assim que nomeamos uma categoria, começamos a inventar coisas a respeito dela. A coisa nomeada não precisa ser observável, não precisa existir, nem mesmo precisa ser coerente.

E não, isso não é exclusividade do Japão. Aqui no Ocidente, um livro sobre dieta baseado no tipo sanguíneo chamado "Eat Right 4 Your Type" (Coma Certo para Seu Tipo, em português) foi um grande sucesso de vendas.

De qualquer forma que se olhe para isso, traçar um limite no espaço das coisas não é um ato neutro. Talvez uma IA com um projeto mais limpo e puramente baseado em princípios bayesianos (um método estatístico de inferência e tomada de decisões) conseguisse ponderar sobre uma classe arbitrária sem ser influenciada por ela. Mas você, um ser humano, não tem essa opção. As categorias não são entidades estáticas no contexto do cérebro humano; assim que as consideramos, elas exercem poder sobre nossa mente. Essa é mais uma razão para não acreditar que podemos definir uma palavra da maneira que desejamos.

171 — Esgueirando-se em conotações

No ensaio anterior, vimos que, no Japão, os tipos sanguíneos substituíram a astrologia — por exemplo, se você tem o tipo sanguíneo AB, é esperado que você seja "descolado e controlado".

Agora, suponhamos que decidimos criar uma nova palavra, "wiggin", e definimos essa palavra como significando pessoas com olhos verdes e cabelos pretos...

Um homem de olhos verdes e cabelos pretos entrou em um restaurante.

"Ha", disse Danny, observando de uma mesa próxima, "você viu isso? Um wiggin acabou de entrar no recinto. Malditos wiggins. Cometem todo tipo de crime".

Sua irmã Erda suspirou. "Você não o viu cometer nenhum crime, viu, Danny?"

"Não preciso", disse Danny, pegando um dicionário. "Olhe aqui, está escrito claramente no Oxford English Dictionary. 'Wiggin. (1) Uma pessoa com olhos verdes e cabelo preto'. Ele tem olhos verdes e cabelos pretos, logo ele é um wiggin. Você não discutirá com o Oxford English Dictionary, vai? Por definição, uma pessoa de olhos verdes e cabelos pretos é um wiggin."

"Mas você o chamou de wiggin", disse Erda. "Isso é uma coisa desagradável de se dizer sobre alguém que você nem conhece. Você não tem nenhuma evidência de que ele coloca muito ketchup em seus hambúrgueres ou que, quando criança, usava um estilingue para atirar em filhotes de esquilo."

"Mas ele é um wiggin", disse Danny pacientemente. "Ele tem olhos verdes e cabelos pretos, certo? Só espere, assim que o hambúrguer chegar, ele pegará o ketchup."

A mente humana passa das características observadas para as características inferidas por meio das palavras. Em "Todos os humanos são mortais, Sócrates é humano, logo, Sócrates é mortal", as características observadas são as roupas de Sócrates, a fala, o uso de ferramentas e, de forma geral, a aparência humana. A categorização é "humano" e a característica inferida é a capacidade de ser envenenado por cicuta.

É claro que não há uma distinção rígida entre "características observadas" e "características inferidas". Se você ouve alguém falar, provavelmente essa pessoa tem a forma de um humano, se o resto permanecer constante. Se você vê uma figura humana nas sombras, então, ceteris paribus[28], é provável que ela possa falar.

E ainda assim, algumas propriedades tendem a ser mais inferidas do que observadas. Você está mais propenso a decidir que alguém é humano, e, portanto, queimará se exposto a uma chama aberta, do que a realizar a inferência no sentido contrário.

Se você buscar no dicionário a definição de "humano", é provável encontrar características como "inteligência" e "bípede sem penas" — úteis para identificar rapidamente o que é ou não um humano — do que as inúmeras conotações que podemos inferir do fato de alguém ser humano, desde a vulnerabilidade à cicuta até a tendência ao excesso de confiança. Por quê? Talvez os dicionários tenham a intenção de permitir que você associe rótulos a grupos de similaridade, e, portanto, sejam projetados para identificar rapidamente

28 NT. **Ceteris paribus**: expressão latina que significa "todo o resto constante" ou "mantidas as demais condições". É usada para isolar o efeito de uma variável específica em uma análise, assumindo que todos os outros fatores permanecem inalterados. Comum em economia, filosofia e ciências sociais.

os agrupamentos no espaço das coisas. Ou talvez as características maiores e mais distintas sejam as mais salientes e, portanto, as primeiras a virem à mente de um editor de dicionário (não tenho certeza de quão conscientes os editores de dicionário estão do que realmente fazem).

Mas o resultado é que, quando Danny pega seu OED para buscar 'wiggin', ele vê listadas apenas as características que, à primeira vista, distinguem um 'wiggin': olhos verdes e cabelos pretos. O OED não lista as diversas conotações menores que passaram a ser associadas a essa palavra, como tendências criminosas, peculiaridades culinárias ou ações infantis infelizes.

Como essas conotações foram parar lá, em primeiro lugar? Talvez tenha existido um wiggin famoso com essas propriedades. Ou talvez alguém tenha inventado coisas aleatoriamente e escrito uma série de livros best-sellers a respeito (O Wiggin, Conversando com Wiggins, Criando seu Pequeno Wiggin, Wiggins na Intimidade). Talvez até os "wiggins" acreditem agora e ajam de acordo. Assim que você chamar algumas pessoas de "wiggins", a palavra começará a adquirir conotações.

Porém, lembremos da Parábola da Cicuta: se seguirmos estritamente as definições lógicas de classes, jamais poderemos classificar Sócrates como 'humano' até termos observado que ele é mortal. Sempre que alguém recorre a um dicionário, geralmente está tentando inserir uma conotação, não buscar a definição literal escrita no dicionário.

Afinal, se o único significado da palavra "wiggin" é 'pessoa de olhos verdes e cabelos pretos', por que não as chamar simplesmente de "pessoas de olhos verdes e cabelos pretos"? E se você estiver se perguntando se alguém gosta de ketchup, por que não perguntar diretamente: "Ele gosta de ketchup?" em vez de "Ele é um wiggin?" (Observe a substituição da substância pelo símbolo.)

Ah, mas discutir a questão real exigiria trabalho. Seria necessário observar de fato o wiggin para ver se ele pega o ketchup. Ou talvez procurar estatísticas sobre quantas pessoas de olhos verdes e cabelos pretos realmente apreciam ketchup. De qualquer forma, você não conseguiria fazer isso sentado em sua sala de estar com os olhos fechados. E as pessoas são preguiçosas. Elas preferem argumentar "por definição", especialmente porque acreditam que 'você pode definir uma palavra como quiser'.

Todavia, a verdadeira razão pela qual se importam se alguém é um "wiggin" é uma conotação — um sentimento que acompanha a palavra - que não está na definição que afirmam usar.

Imagine Danny dizendo: 'Veja, ele tem olhos verdes e cabelos pretos. Ele é um "wiggin"! Está escrito no dicionário! Logo, ele tem cabelo preto. Tente argumentar contra isso, se puder!"

Não soa muito triunfante, não é mesmo? Se o ponto real do argumento estivesse verdadeiramente contido na definição do dicionário — se o argumento fosse logicamente válido —, então o argumento pareceria vazio; ou ele não acrescentaria nada novo, ou seria uma petição de princípio.

É apenas a tentativa de introduzir conotações que não estão explicitamente listadas na definição que faz com que as pessoas acreditem que podem marcar pontos dessa maneira.

172 — Argumentando “por definição”

“Esta galinha depenada tem duas pernas e não tem penas — portanto, por definição, é um ser humano!”

Quando as pessoas debatem definições, começam geralmente com um conjunto de características visíveis, conhecidas ou amplamente aceitas como verdadeiras. Em seguida, recorrem a um dicionário e mostram que essas características se encaixam na definição do dicionário, concluindo assim: “Portanto, por definição, o ateísmo é uma religião!

No entanto, características visíveis, conhecidas e amplamente aceitas raramente são o cerne da disputa. O fato de alguém pensar que as duas pernas de Sócrates são suficientemente evidentes para sustentar a premissa de que, “por definição, Sócrates é humano!” indica que o bipedalismo provavelmente não é realmente o ponto em questão. Caso contrário, o ouvinte responderia: “Como assim Sócrates é bípede? É sobre isso que estamos discutindo desde o início!“

Há um sentido importante no qual podemos, de forma legítima, passar de características evidentes para outras menos óbvias. Pode-se, legitimamente, observar que Sócrates tem forma humana e prever sua vulnerabilidade à cicuta. No entanto, essa inferência probabilística não depende de definições de dicionário ou do uso comum; ela se baseia no fato de que o universo contém agrupamentos empíricos de coisas semelhantes.

Essa estrutura de agrupamento não mudará dependendo de como você define suas palavras. Mesmo que você procure a definição de “humano” no dicionário e ela diga “todos os bípedes sem penas, exceto Sócrates”, isso não vai alterar o grau real em que Sócrates se assemelha ao restante de nós, bípedes sem penas.

Ao argumentar corretamente a partir da estrutura de agrupamento, alguém diria algo como: “Sócrates tem dois braços, dois pés, nariz e língua, fala grego fluentemente, usa ferramentas e, em todos os aspectos que pude observar, parece ter todas as propriedades principais e secundárias que caracterizam o *Homo sapiens*; portanto, vou supor que ele tenha DNA humano, bioquímica humana e é vulnerável à cicuta, assim como todos os outros *Homo sapiens* nos quais a letalidade da cicuta foi clinicamente testada.”

Suponhamos que eu responda: “Mas vi Sócrates no campo com alguns herbologistas; acredito que eles estavam tentando preparar um antídoto. Portanto, não espero que Sócrates morra após beber a cicuta — ele será uma exceção ao comportamento geral dos objetos em seu grupo, pois eles não tomaram o antídoto e ele tomou.“

A essa altura, não faz muito sentido discutir se Sócrates é “humano” ou não. A conversa deve passar para um nível mais detalhado, explorando os detalhes que compõem a categoria “humana” — discutindo a bioquímica humana e, especificamente, os efeitos neurotóxicos da coniina.

Se você continuar insistindo: “Mas Sócrates é humano e os humanos, por definição, são mortais!” então o que você realmente está tentando fazer é apagar tudo o que sabe sobre Sócrates, exceto o fato de ele ser humano — insistindo que a única previsão correta é aquela que você faria se não soubesse nada sobre Sócrates, exceto sua humanidade.

Isso é semelhante a insistir que uma moeda tem 50% de probabilidade de dar cara ou coroa porque é uma “moeda honesta”, mesmo após olhar para a moeda e ver que deu cara. É como afirmar que Frodo tem dez dedos porque a maioria dos hobbits tem dez dedos, mesmo após olhar para as mãos dele e ver apenas

nove dedos. Obviamente, isso é inaceitável na teoria da probabilidade bayesiana, pois você não pode simplesmente se recusar a considerar novas evidências.

E você não pode simplesmente manter uma categorização e fazer estimativas com base nela, enquanto deliberadamente ignora tudo o que sabe.

Nem todas as novas evidências fazem uma diferença significativa, é claro. Se eu perceber que Sócrates tem nove dedos, isso não mudará visivelmente minha estimativa de sua vulnerabilidade à cicuta, pois espero que a maneira como Sócrates perdeu o dedo não tenha alterado o restante de sua bioquímica. Isso é verdade, independentemente de a definição do dicionário afirmar ou não que os seres humanos têm dez dedos. A inferência válida baseia-se na estrutura de agrupamento do ambiente e na estrutura causal da biologia, não no que o editor do dicionário escreve, nem mesmo no "uso comum".

Normalmente, quando você está fazendo isso corretamente — de maneira legítima -, você apenas afirma: "O alcaloide coniina encontrado na cicuta produz paralisia muscular em humanos, resultando em morte por asfixia". Ou, de forma mais simples, "os humanos são vulneráveis à cicuta". É assim que geralmente é expresso em uma discussão legítima.

Quando alguém sente a necessidade de fortalecer o argumento com a frase enfática "por definição"? (Por exemplo, "Os humanos são vulneráveis à cicuta por definição!") Ora, quando a característica inferida é questionada — se Sócrates foi visto consultando herbologistas, por exemplo — e o orador sente a necessidade de reforçar a lógica.

Portanto, quando você vê "por definição" sendo usado dessa maneira, geralmente significa: "Esqueça o que você ouviu sobre Sócrates consultando herbologistas — humanos, por definição, são mortais!"

As pessoas sentem a necessidade de resumir o argumento em uma única frase, dizendo: "Qualquer P, por definição, tem a propriedade Q!" exatamente quando veem e preferem descartar imediatamente argumentos adicionais que questionam a inferência padrão baseada em agrupamento.

O mesmo acontece com o argumento "X, por definição, é um Y!" Por exemplo, "Ateus acreditam que Deus não existe; portanto, os ateus têm crenças sobre Deus, pois uma crença negativa ainda é uma crença; portanto, o ateísmo afirma respostas a questões teológicas; portanto, o ateísmo é, por definição, uma religião".

Você não sentiria a necessidade de dizer: "O hinduísmo, por definição, é uma religião!" porque, bem, é óbvio que o hinduísmo é uma religião. Não é apenas uma religião "por definição", é, de fato, uma religião real.

O ateísmo não se enquadra nos membros centrais do agrupamento "religião". Portanto, se não fosse pelo fato de que o ateísmo é uma religião por definição, poder-se-ia pensar que o ateísmo não é uma religião. É por isso que é necessário esmagar toda a oposição apontando que "O ateísmo é uma religião" é verdade por definição, pois não é verdade de nenhuma outra forma.

Ou seja: as pessoas insistem que "X, por definição, é um Y!" em ocasiões em que estão tentando inserir uma conotação de Y que não está diretamente na definição, e X não se assemelha aos outros membros do grupo Y.

Nos últimos treze anos, observo com que frequência essa frase é usada correta e incorretamente — embora não tenha estatísticas literais, temo eu. No entanto, parece que usar a expressão "por definição" fora do contexto matemático está entre os sinais mais alarmantes de argumento falho. Está no mesmo nível de "Hitler", "Deus", "absolutamente certo" e "não posso provar isso".

Essa heurística de falha não é perfeita — a primeira vez que encontrei um uso correto fora da matemática foi por Richard Feynman e, desde então, tenho visto mais casos. Mas provavelmente é melhor simplesmente excluir a frase "por definição" do seu vocabulário — e sempre em qualquer ocasião em que você se sentir tentado a usá-la em itálico ou seguida de um ponto de exclamação. Essa é uma má ideia *por definição*!

173 — Onde traçar o limite?

Alguém se aproxima de você e diz:

Por muito tempo, eu ponderava sobre o significado da palavra “Arte” e, finalmente, encontrei uma definição que considero satisfatória: “Arte é aquilo criado com o propósito de provocar uma reação no público”.

Apenas porque existe a palavra “arte” não significa que ela **tenha um significado** flutuando no vazio, que você pode **descobrir** encontrando a definição correta.

Parece ser assim, mas não é.

Querer descobrir como definir uma palavra significa que você está abordando o problema de maneira equivocada — procurando a essência misteriosa do que é, essencialmente, um sinal de comunicação.

Agora, há um verdadeiro desafio que um racionalista pode legitimamente enfrentar, mas esse desafio não é encontrar uma definição satisfatória para uma palavra. O verdadeiro desafio pode ser jogado como um jogo para um único jogador, sem necessidade de verbalização. O desafio é descobrir quais coisas são semelhantes entre si — quais coisas estão agrupadas — e, às vezes, quais coisas têm uma causa comum.

Se você definir “*eluctromugnetismo*” para incluir raios, bússolas, excluindo a luz e incluindo o ‘magnetismo animal’ de Mesmer (agora conhecido como hipnose), então você terá dificuldade em perguntar ‘Como funciona o *eluctromugnetismo*?’ Você uniu coisas que não têm relação entre si e excluiu outras que seriam necessárias para formar um conjunto completo. (Esse exemplo tem base histórica; Mesmer veio antes de Faraday.)

Poderíamos dizer que “*eluctromugnetismo*” é uma palavra inadequada, um limite no espaço das coisas que dá voltas e desvia-se dos agrupamentos, uma divisão que não consegue esculpir a realidade ao longo de suas articulações naturais.

Descobrir onde dividir a realidade para esculpir ao longo das articulações — esse é o desafio digno de um racionalista. É isso que as pessoas deveriam buscar quando se aventuram em busca da essência volátil de uma palavra.

E não se engane: é um desafio científico perceber que você precisa de uma única palavra para descrever a respiração e o fogo. Portanto, não espere obter ajuda dos editores de dicionários, pois essa não é a tarefa deles.

O que é “arte”? Mas não há uma essência da palavra flutuando no vazio.

Talvez você venha até mim com uma longa lista de coisas que você considera “arte” e “não arte”:

Uma pequena fuga em sol menor: arte.

Um soco no nariz: não é arte.

Relatividade de Escher: Arte.

Uma flor: Não é arte.

A linguagem de programação Python: Arte.

Uma cruz flutuando na urina: não é arte

Os romances Tschai de Jack Vance: Arte.

Arte Moderna: Não é arte.

E você me pergunta: “Parece intuitivo para eu estabelecer esse limite, mas não sei por quê — você pode encontrar uma intenção que corresponda a essa extensão? Você pode me dar uma descrição simples desse limite?“

Então respondo: “Acredito que tenha a ver com a admiração pelo artesanato: trabalho investido e maravilha resultante. O que esses itens incluídos têm em comum são as emoções estéticas semelhantes que eles inspiram e o esforço consciente humano empregado para criar tais emoções.“

Isso é útil ou é apenas uma trapaça no jogo do Tabu? Eu diria que a lista de quais emoções humanas são consideradas estéticas ou não é muito mais concisa do que a lista de tudo o que é considerado arte ou não. Talvez seja possível observar essas emoções por meio de um exame de ressonância magnética — menciono isso para enfatizar que as emoções não são imateriais.

Mas é claro que minha definição de arte não é o ponto central. O ponto central é que tanto a intenção quanto a extensão da minha definição podem ser contestadas.

Alguém poderia dizer: “A emoção estética não é o que essas coisas têm em comum; o que elas têm em comum é a intenção de inspirar qualquer emoção complexa pelo simples fato de inspirá-la.” Isso seria contestar minha intenção, minha tentativa de traçar uma curva através dos pontos de dados. Alguém diria: “Sua equação pode se ajustar aproximadamente a esses dados, mas não é a distribuição geradora verdadeira“.

Ou poderia contestar minha extensão, dizendo: “Algumas dessas coisas devem ficar juntas — eu entendo o seu ponto —, mas a linguagem Python não deveria estar na lista, enquanto a arte moderna deveria.” (Isso o rotularia como um ignorante, mas você poderia argumentar.) Aqui, a suposição é que realmente existe uma curva subjacente que gera essa aparente lista de semelhanças e diferenças — que há uma ordem e uma lógica, embora você ainda não tenha identificado sua origem. No entanto, sem querer, acabei perdendo o ritmo e incluí alguns pontos de dados de uma fonte diferente.

Muito antes de você saber o que a eletricidade e o magnetismo têm em comum, você ainda poderia suspeitar — com base em suas aparências superficiais — que o “magnetismo animal” não pertence à lista.

Antigamente, acreditava-se que a palavra “peixe” incluía os golfinhos. Agora, você poderia se apresentar como um argumentador perspicaz e dizer: “A lista: {salmão, lebistes, tubarões, golfinhos, truta} é apenas uma lista — não se pode dizer que uma lista está errada. Posso provar na teoria dos conjuntos que essa lista é válida. Portanto, minha definição de peixe, a qual é simplesmente essa lista extensional, não pode estar ‘errada’, como você afirma.“

Ou você pode parar de jogar e admitir que os golfinhos não pertencem à lista de peixes.

Você cria uma lista de coisas que parecem similares e tenta adivinhar porque isso ocorre. Mas quando finalmente descobrir o que elas realmente têm em comum, pode acontecer de sua suposição estar equivocada. Pode até mesmo acontecer de sua lista estar incorreta.

Você não pode se abrigar sob um escudo reconfortante de correto-por-definição. Tanto as definições extensionais quanto as intencionais podem estar erradas, podem falhar em delinear a realidade de acordo com suas articulações.

Categorizar é um esforço de conjectura no qual erros podem ser cometidos; portanto, é sensato admitir, do ponto de vista teórico, que suas suposições de definição podem estar “erradas”.

174 — Entropia e códigos curtos

(Se você não estiver familiarizado com a inferência Bayesiana, este pode ser um bom momento para ler Uma Explicação Intuitiva do Teorema de Bayes.)

Suponha que você tenha um sistema X com a mesma probabilidade de estar em qualquer um dos 8 estados possíveis:

$$\{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8\}$$

Existe uma quantidade extraordinariamente onipresente — na física, na matemática e até na biologia — chamada entropia; e a entropia de X é de 3 bits. Isso significa que, em média, teremos que fazer 3 perguntas de sim ou não para descobrir o valor de X. Por exemplo, alguém poderia nos dizer o valor de X usando este código:

$$X_1 : 001 \qquad X_2 : 010 \qquad X_3 : 011 \qquad X_4 : 100$$
$$X_5 : 101 \qquad X_6 : 110 \qquad X_7 : 111 \qquad X_8 : 000$$

Então, se eu perguntasse “O primeiro símbolo é 1?” e ouvisse “sim”, então perguntasse “O segundo símbolo é 1?” e ouvisse “não”, então perguntasse “O terceiro símbolo é 1?” e ouvisse “não”, eu saberia que X estava no estado 4.

Agora, suponha que o sistema Y tenha quatro estados possíveis com as seguintes probabilidades:

$$Y_1 : ½\,(50\%) \qquad Y_2 : ¼\,(25\%)$$
$$Y_3 : ⅛\,(12.5\%) \qquad Y_4 : ⅛\,(12.5\%)$$

Então, a entropia de Y seria de 1,75 bits, o que significa que podemos descobrir seu valor fazendo 1,75 perguntas de sim ou não.

O que significa falar sobre fazer um e três quartos de uma pergunta? Imagine que designamos os estados de Y usando o seguinte código:

$$Y_1 : 1 \qquad Y_2 : 01 \qquad Y_3 : 001 \qquad Y_4 : 000$$

Primeiro, você pergunta: “O primeiro símbolo é 1?” Se a resposta for “sim”, você terminou: Y está no estado 1. Isso acontece metade das vezes, então 50% das vezes leva 1 pergunta de sim ou não para descobrir o estado de Y.

Suponha que, em vez disso, a resposta seja “Não”. Então, você pergunta: “O segundo símbolo é 1?” Se a resposta for “sim”, você terminou: Y está no estado 2. O sistema Y está no estado 2 com probabilidade 1/4, e cada vez que Y está no estado 2, descobrimos esse fato usando duas perguntas de sim ou não, então 25% do tempo leva 2 perguntas para descobrir o estado de Y.

Se a resposta for “Não” duas vezes seguidas, você pergunta “O terceiro símbolo é 1?” Se for “sim”, você terminou e Y está no estado 3; se for “não”, você terminou e Y está no estado 4. Em 1/8 do tempo em que Y está no estado 3, são necessárias três perguntas; e 1/8 do tempo em que Y está no estado 4, são necessárias três perguntas.

(1/2 × 1) + (1/4 × 2) + (1/8 × 3) + (1/8 × 3) = 0,5 + 0,5 + 0,375 + 0,375 = 1,75.

A fórmula geral para a entropia H(S) de um sistema S, composto por i estados, cada um com entropia S_i , é:

$$H(S) = -\sum_i (P(S_i) * log_2(P(S_i)))$$

Por exemplo, o log (base 2) de 1/8 é −3. Então, -(1/8 × −3) = 0,375 é a contribuição do estado S4 para a entropia total: 1/8 do tempo, temos que fazer 3 perguntas.

Você nem sempre pode conceber um código perfeito para um sistema, mas se você tiver que comunicar o estado de muitas cópias arbitrariamente de S em uma única mensagem, você pode se aproximar arbitrariamente de um código perfeito. (Pesquise “codificação aritmética” para um método simples.)

Agora, você pode perguntar: “Por que não usar o código 10 para Y_4, em vez de 000? Isso não nos permitiria transmitir mensagens mais rapidamente?“

Mas se você usar o código 10 para Y_4, quando alguém responder “Sim” à pergunta “O primeiro símbolo é 1?”, você ainda não saberá se o estado do sistema é Y_1 (1) ou Y_4 (10). Na verdade, se você alterar o código dessa maneira, todo o sistema se desmorona — porque se você ouvir “1001”, não saberá se isso significa “Y_4, seguido de Y_2” ou “Y_1, seguido de Y_3”.

A moral da história é que palavras curtas são um recurso valioso.

A chave para criar um bom código — um código que transmita mensagens da forma mais concisa possível — é reservar palavras curtas para expressar conceitos frequentes e utilizar palavras mais longas para aqueles que são menos comuns.

Quando levamos essa arte ao limite, o comprimento da mensagem necessário para descrever algo corresponde exata ou quase exatamente à sua probabilidade. Essa é a formalização do Comprimento Mínimo da Descrição, também conhecido como Comprimento Mínimo da Mensagem na Navalha de Ocam.

E assim, até mesmo os rótulos que atribuímos às palavras não são totalmente arbitrários. Os sons que associamos aos nossos conceitos podem ser melhores ou piores, mais sábios ou mais tolos, mesmo independentemente de considerações de uso comum!

Menciono tudo isso porque a ideia de “você pode fazer X da maneira que quiser” é um grande obstáculo para aprender a fazê-lo com sabedoria. O mantra “é um país livre; tenho direito à minha própria opinião” obstrui a arte de buscar a verdade. A noção de “posso definir uma palavra como quiser” obstrui a arte de esculpir a realidade em suas articulações. Até mesmo a frase sensata “os rótulos atribuídos às palavras são arbitrários” obstrui a consciência da concisão. A prosódia, aliás, também desempenha um papel importante — Tolkien observou certa vez o quão bonito soa a frase “porta do porão” (cellar door); esse tipo de consciência é necessário para usar a linguagem como Tolkien.

O comprimento das palavras também desempenha um papel não trivial na ciência cognitiva da linguagem.

Considere as expressões “poltrona reclinável”, “cadeira” e “mobília”. A poltrona reclinável é uma categoria mais específica do que a cadeira, enquanto mobília é uma categoria mais ampla. No entanto, a grande maioria das cadeiras possui um uso comum — você realiza a mesma ação motora ao sentar-se nelas e a finalidade é a mesma (aliviar o peso dos pés enquanto come, lê, digita ou descansa). As poltronas reclináveis não fogem dessa temática. Por outro lado, “mobília” inclui objetos como camas e mesas, que possuem diferentes usos e evocam funções motoras distintas das cadeiras.

Na terminologia da psicologia cognitiva, “cadeira” é uma categoria de nível básico.

As pessoas têm a tendência de falar e, presumivelmente, pensar no nível básico de categorização — traçar o limite em torno de “cadeiras” ao invés de se concentrar na categoria mais específica “poltrona reclinável” ou na categoria mais geral “mobília”. As pessoas são mais propensas a dizer “Você pode sentar-se naquela cadeira” do que “Você pode sentar-se naquela poltrona reclinável” ou “Você pode sentar-se naquele móvel”.

E não é coincidência que a palavra “cadeira” tenha menos sílabas do que “poltrona reclinável” ou “mobília”. Em geral, as categorias de nível básico tendem a ter nomes curtos, e substantivos com nomes curtos tendem a se referir a categorias de nível básico. Essa não é uma regra perfeita, é claro, mas uma tendência definida. O uso frequente acompanha palavras curtas, assim como palavras curtas acompanham o uso frequente.

Ou, como disse Douglas Hofstadter, há uma razão pela qual a língua inglesa utiliza “the” para significar “o” e “antidisestablishmentarianism” para significar “antidesestabelecimentarianismo” ao invés de antidesestabelecimentarianismo de outro jeito.

175 — Informação mútua e Densidade no Espaço das Coisas

Vamos supor que tenhamos um sistema X que pode estar em um dos 8 estados possíveis, todos igualmente prováveis (em relação ao seu conhecimento atual), e um sistema Y que pode estar em um dos 4 estados possíveis, também igualmente prováveis.

A entropia de X, como definido no ensaio anterior, é de 3 bits; precisamos fazer 3 perguntas de sim ou não para determinar o estado exato de X. A entropia de Y é de 2 bits; necessitamos de 2 perguntas de sim ou não para determinar o estado exato de Y. Isso pode parecer óbvio, já que 2 elevado à potência de 3 é igual a 8 e 2 elevado à potência de 2 é igual a 4, portanto, 3 perguntas podem distinguir 8 possibilidades e 2 perguntas podem distinguir 4 possibilidades. No entanto, é importante lembrar que se as possibilidades não fossem igualmente prováveis, poderíamos utilizar um código mais inteligente para descobrir o estado de Y, usando, por exemplo, 1,75 perguntas em média. Neste caso, entretanto, a distribuição de probabilidade de X é uniforme entre todos os seus estados possíveis, assim como a de Y, o que significa que não podemos usar nenhum código inteligente.

Qual é a entropia do sistema combinado (X, Y)?

Você pode ser tentado a responder: “São necessárias 3 perguntas para descobrir X e, em seguida, 2 perguntas para descobrir Y, então são necessárias 5 perguntas no total para descobrir o estado de X e Y”.

Mas e se as duas variáveis estiverem entrelaçadas, de forma que aprender o estado de Y nos forneça alguma informação sobre o estado de X?

Especificamente, vamos supor que X e Y sejam ambos ímpares ou ambos pares. Agora, se recebermos uma mensagem de 3 bits (3 perguntas) e descobrirmos que X está no estado X_5, sabemos que Y está no estado Y_1 ou Y_3, mas não nos estados Y_2 ou Y_4. Portanto, uma pergunta adicional “Y está no estado Y_3?” respondida com “Não” nos fornece todo o estado de (X, Y): $X = X_5$, $Y = Y_1$. E aprendemos isso com um total de 4 perguntas.

Por outro lado, se descobrirmos que Y está no estado Y_4 usando duas perguntas, serão necessárias apenas mais duas perguntas para determinar se X está no estado X_2, X_4, X_6 ou X_8. Novamente, são necessárias quatro perguntas para aprender o estado do sistema combinado.

A informação mútua entre duas variáveis é definida como a diferença entre a entropia do sistema combinado e a entropia das variáveis independentes:

$$I(X;\ Y) \ = \ H(X) \ + \ H(Y) \ - \ H(X,\ Y)$$

Nesse caso, há 1 bit de informação mútua entre os dois sistemas: aprender X nos fornece 1 bit de informação sobre Y (reduzindo o espaço de possibilidades de 4 para 2, uma redução de fator 2 no volume), e aprender Y nos fornece 1 bit de informação sobre X (reduzindo o espaço de possibilidades de 8 para 4).

E quando a distribuição de probabilidades não é uniforme? No ensaio anterior, discutimos o caso em que Y tinha probabilidades de 1/2, 1/4, 1/8, 1/8 para seus quatro estados. Vamos considerar essa distribuição de probabilidades sobre Y como independente — se observarmos Y, sem ver nada mais, é o que esperaríamos ver. Suponhamos que a variável Z tenha dois estados, Z_1 e Z_2, com probabilidades de 3/8 e 5/8, respectivamente.

Portanto, somente se a distribuição conjunta de Y e Z for a seguinte, não haverá informação mútua

$$\begin{array}{cccc} Z_1Y_1 : 3/16 & Z_1Y_2 : 3/32 & Z_1Y_3 : 3/64 & Z_1Y_4 : 3/64 \\ Z_2Y_1 : 5/16 & Z_2Y_2 : 5/32 & Z_2Y_3 : 5/64 & Z_2Y_4 : 5/64 \end{array}$$

entre Y e Z:

Esta distribuição obedece à lei:

$$P\,(Y,\,Z) \;=\; P\,(Y) \times\; P\,(Z)$$

Por exemplo:

$$P\,(Z_1Y_2) \;=\; P\,(Z_1) \times\; P\,(Y_2) \;=\; 3/8 \times 1/4 \;=\; 3/32$$

E observe que podemos recuperar as probabilidades marginais (independentes) de Y e Z apenas observando a distribuição conjunta:

$P(Y_1)$ = Probabilidade total de todas as diferentes maneiras pelas quais Y_1 pode ocorrer.

$$P(Y_1) = P\,(Z_1Y_1) \;+\; P\,(Z_2Y_1) = 3/16 \;+\; 5/16 = 1/2$$

Assim, apenas inspecionando a distribuição conjunta, podemos determinar se as variáveis marginais Y e Z são independentes, ou seja, se a distribuição conjunta é fatorada no produto das distribuições marginais; se, para todos os valores de Y e Z, tivermos .

Este último é significativo porque, pela Regra de Bayes,

$$P\left(Z_j Y_i\right) = P\left(Y_i\right) \times P\left(Z_j\right)$$

$$\frac{P\left(Z_j Y_i\right)}{P\left(Z_j\right)} = P\left(Y_i\right)$$

$$P\left(Y_i | Z_j\right) = P\left(Y_i\right)$$

Em inglês: “Após aprender , sua crença sobre é exatamente a mesma que era antes.”

Portanto, quando a distribuição é fatorável — quando isso é equivalente a “Aprender sobre Y nunca nos diz nada sobre Z, ou vice-versa”.

A partir disso, você pode suspeitar corretamente que não haja informação mútua entre Y e Z. Onde não há informação mútua, não há evidência bayesiana, e vice-versa.

Suponha que na distribuição (Y, Z) acima, tratamos cada combinação possível de Y e Z como um evento separado - de modo que a distribuição (Y, Z) tenha um total de 8 possibilidades, com as probabilidades mostradas - e então calculamos a entropia da distribuição (Y, Z) da mesma forma como calcularíamos a entropia de qualquer distribuição:

$$P(Z_1 Y_1)\log_2(P(Z_1 Y_1)) + P(Z_1 Y_2)\log_2(P(Z_1 Y_2)) +$$
$$P(Z_1 Y_3)\log_2(P(Z_1 Y_3)) + \ldots + P(Z_2 Y_4)\log_2(P(Z_2 Y_4))$$

$$= (3/16)\log_2(3/16) + (3/32)\log_2(3/32) +$$
$$(3/64)\log_2(3/64) + \ldots + (5/64)\log_2(5/64)\ .$$

Você obterá o mesmo resultado se calcular separadamente a entropia de Y e a entropia de Z. Isso ocorre porque não há informação mútua entre as duas variáveis, o que significa que nossa incerteza sobre o sistema conjunto não é menor do que nossa incerteza sobre os dois sistemas considerados separadamente. (Embora eu não esteja fornecendo os cálculos aqui, você pode realizá-los por conta própria. Além disso, não estou apresentando a prova geral dessa afirmação, mas você pode pesquisar sobre “entropia de Shannon” e “informação mútua”.)

E se a distribuição conjunta não for fatorável? Por exemplo:

$$Z_1 Y_1 : 12/64 \qquad Z_1 Y_2 : 8/64 \qquad Z_1 Y_3 : 1/64 \qquad Z_1 Y_4 : 3/64$$
$$Z_2 Y_1 : 20/64 \qquad Z_2 Y_2 : 8/64 \qquad Z_2 Y_3 : 7/64 \qquad Z_2 Y_4 : 5/64$$

Se somarmos as probabilidades conjuntas para obter as probabilidades marginais, veremos que $P(Y_1) = 1/2$, $P(Z_1) = 3/8$ e assim por diante — as probabilidades marginais são as mesmas de antes.

No entanto, as probabilidades conjuntas nem sempre são iguais ao produto das probabilidades marginais. Por exemplo, a probabilidade $P(Z_1Y_2)$ é igual a 8/64, enquanto $P(Z_1)P(Y_2)$ seria igual a $3/8 \times 1/4 = 6/64$. Ou seja, a probabilidade de encontrar Z_1Y_2 juntos é maior do que o esperado com base nas probabilidades de encontrar Z_1 ou Y_2 separadamente.

Isso implica que:

$$\begin{aligned} P(Z_1 Y_2) &> P(Z_1)P(Y_2) \\ P(Z_1 Y_2)/P(Y_2) &> P(Z_1) \\ P(Z_1|Y_2) &> P(Z_1). \end{aligned}$$

Como há uma probabilidade "excepcionalmente alta" para $P(Z_1Y_2)$ — definida como uma probabilidade maior do que as probabilidades marginais indicariam por padrão — segue-se que observar Y_2 é uma evidência que aumenta a probabilidade de Z_1. E por um argumento simétrico, observar Z_1 deve favorecer Y_2.

Como existem pelo menos alguns valores de Y que nos dizem sobre Z (e vice-versa), deve haver informação mútua entre as duas variáveis. Assim, você descobrirá — estou confiante, embora não tenha verificado — que calcular a entropia de (Y, Z) produz menos incerteza total do que a soma das entropias independentes de Y e Z., Ou seja, , com todas as grandezas necessariamente positivas. (Divago aqui para observar que a simetria da expressão para a informação mútua mostra que Y deve nos dizer tanto sobre Z, em média, quanto Z nos diz sobre Y. Deixo como exercício para o leitor reconciliar isso com qualquer coisa que eles aprenderam na aula de lógica sobre como, se todos os corvos são pretos, poder raciocinar Corvos(x) ⇒ Preto(x) não significa que você pode raciocinar Preto(x) ⇒ Corvos(x). O quanto parecem ser diferentes os fluxos simétricos de probabilidade do Bayesianismo, em comparação com as reviravoltas abruptas da lógica — mesmo que este último seja apenas um caso degenerado do primeiro.)

"Mas", você pergunta, "o que tudo isso tem a ver com o uso adequado das palavras?"

No ensaio "Rótulos Vazios" e, em seguida, "Substituir o Símbolo Pela Substância", vimos a técnica de substituir uma palavra por sua definição — o exemplo dado foi:

Todos os seres [mortais, penas, bípedes] são mortais.

Sócrates é um ser [mortal, ¬penas, bípede].

Portanto, Sócrates é mortal.

Por que, então, você iria querer ter uma palavra para "humano"? Por que não dizer apenas "Sócrates é um bípede mortal sem penas"?

Porque é útil ter palavras mais curtas para coisas que encontramos com frequência. Se o código para descrever propriedades individuais já é eficiente, então não há vantagem em ter uma palavra especial para uma conjunção — como "humano" para "bípede sem penas mortal" — a menos que coisas que são mortais,

sem penas e bípedes sejam encontradas com mais frequência do que as probabilidades marginais levariam você a esperar.

Em códigos eficientes, o comprimento da palavra corresponde à probabilidade — então o código para Z_1Y_2 será tão longo quanto o código para Z_1 mais o código para Y_2, a menos que $P(Z_1Y_2) > P(Z_1) \times P(Y_2)$, caso em que o código da palavra pode ser mais curto do que os códigos de suas partes.

Isso significa que, quando podemos inferir algumas propriedades de uma coisa com base em suas outras propriedades, é mais provável que tenhamos uma palavra especial para essa coisa. Por exemplo, é mais provável do que o padrão que coisas bípedes sem penas também sejam mortais.

É verdade que a palavra "humano" descreve muitas outras propriedades além de apenas a forma física — quando encontramos uma entidade com características humanas, como a capacidade de falar e vestir roupas, podemos inferir várias informações sobre ela, como fatos bioquímicos, anatômicos e cognitivos. Substituir a palavra "humano" por uma descrição completa de todas essas propriedades seria impraticável e demandaria muito tempo de comunicação. Mas isso só é verdade porque um bípede falante sem penas tem muito mais probabilidade do que o padrão de ser envenenado por cicuta, de ter unhas largas ou de ser excessivamente confiante.

Ter uma palavra para uma coisa, em vez de listar todas as suas propriedades, é um código mais compacto, especialmente quando podemos inferir algumas dessas propriedades com base em outras propriedades. (com exceção, talvez, de palavras muito primitivas, como "vermelho", que usamos para transmitir uma descrição totalmente descompactada de nossas experiências sensoriais. Mas quando encontramos um inseto ou até mesmo uma pedra, estamos lidando com conjuntos de propriedades não simples, muito acima do nível primitivo.)

Portanto, ter uma palavra como "wiggin" para descrever pessoas com olhos verdes e cabelos pretos é mais útil do que simplesmente dizer "pessoa de olhos verdes e cabelos pretos". Isso é especialmente relevante quando:

1. As pessoas com olhos verdes têm uma probabilidade maior do que a média de terem cabelos pretos (e vice-versa), o que nos permite inferir probabilisticamente a presença de olhos verdes em pessoas com cabelos pretos, ou vice-versa; ou
2. Wiggins compartilham outras características que podem ser inferidas com uma probabilidade maior do que o usual. Nesse caso, devemos considerar separadamente os olhos verdes e os cabelos pretos. Entretanto, uma vez que tenhamos observado essas características independentemente, podemos inferir probabilisticamente outras propriedades (como o gosto por ketchup).

Pode-se até considerar o ato de definir uma palavra como uma promessa nesse sentido. Ao dizer a alguém: "Eu defino a palavra 'wiggin' como uma pessoa com olhos verdes e cabelos pretos", implicitamente afirmamos que a palavra "wiggin" ajudará de alguma forma a fazer inferências e a encurtar nossas mensagens.

Se a presença conjunta de olhos verdes e cabelos pretos não apresenta uma probabilidade maior do que o usual, e nenhuma outra propriedade ocorre com maior probabilidade em conjunto com essas características, então a palavra "wiggin" é falsa: ela afirma que certas pessoas merecem ser distinguidas como um grupo, mas na realidade não o são.

Nesse caso, a palavra "wiggin" não auxilia na descrição compacta da realidade, não sendo definida como a opção que transmite a mensagem mais concisa, e tampouco desempenha um papel na explicação

mais simples. Da mesma forma, a palavra “wiggin” não será útil para realizar inferências bayesianas. Mesmo que não a chamemos de mentira, certamente é um equívoco.

A forma de esculpir a realidade em suas articulações é delimitar seus contornos em torno de concentrações de densidade de probabilidades excepcionalmente altas no Espaço das Coisas.

176 — Espaço conceitual superexponencial e palavras simples

Pode-se considerar que o Espaço das Coisas é um espaço bastante amplo. Muito maior do que a realidade, já que onde a realidade contém apenas coisas que realmente existem, o Espaço das Coisas contém tudo o que poderia existir.

Na verdade, da forma como "defini" o Espaço das Coisas para ter dimensões para cada atributo possível, incluindo atributos correlacionados como densidade, volume e massa — o Espaço das Coisas pode ser muito mal definido para ter algo que podemos chamar de tamanho. No entanto, é importante conseguir visualizar o Espaço das Coisas de qualquer maneira. Certamente, ninguém consegue compreender um bando de pardais se tudo o que eles veem é uma nuvem de criaturas batendo asas grasnando, em vez de um conjunto de pontos no Espaço das Coisas.

Por mais vasto que o Espaço das Coisas possa ser, ele não se compara em tamanho ao Espaço Conceitual.

O termo "conceito", no contexto de aprendizado de máquina, refere-se a uma regra que inclui ou exclui exemplos. Por exemplo, se tivermos os seguintes dados {2:+, 3:-, 14:+, 23:-, 8:+, 9:-}, podemos deduzir que o conceito em questão é "números pares". Existe uma ampla literatura (como é de se esperar) sobre como aprender conceitos a partir de dados, seja via exemplos selecionados aleatoriamente ou escolhidos especificamente, considerando possíveis erros de classificação e, o mais importante, explorando diferentes espaços de regras possíveis.

Suponhamos, por exemplo, que desejamos aprender o conceito de "dias bons para jogar tênis". Os possíveis atributos da variável Dias são:

```
Céu: {Ensolarado, Nublado, Chuvoso}

TempAr:    {Quente, Frio}

Umidade:    {Normal, Alta}

Vento:          {Forte, Fraco}.
```

Aqui estão os seguintes dados, onde + indica um exemplo positivo para o conceito e — indica uma classificação negativa:

```
+    Céu:   Enso-  TempAr:
     larado        Quente
     Umidade:      Vento: Forte
     Alta
-    Céu: Chuvoso  TempAr: Frio
     Umidade:      Vento: Forte
     Alta
```

+	Céu: Ensolarado	TempAr: Quente	
	Umidade: Alta	Vento: Fraco	

O que um algoritmo deve inferir com base nesses dados?

Um algoritmo de aprendizado de máquina pode representar um conceito que se ajuste a esses dados da seguinte maneira:

```
{Céu: ? ; TempAr: Quente; Umidade: Alta; Vento: ?}.
```

Nessa representação, para determinar se o conceito aceita ou rejeita um exemplo, comparamos elemento por elemento: "?" aceita qualquer valor, enquanto um valor específico aceita apenas esse valor em particular.

Portanto, o conceito acima aceitará apenas Dias com TempAr = Quente e Umidade = Alta, enquanto o Céu e o Vento podem assumir qualquer valor. Essa representação se encaixa tanto nas classificações negativas quanto nas positivas dos dados até o momento, embora não seja o único conceito que faça isso.

Também podemos simplificar a representação do conceito acima para :

```
{?, Quente, Alta, ?}
```

Sem entrar em detalhes, o algoritmo clássico seria:

- Mantenha o conjunto de hipóteses mais gerais que se ajustem aos dados, ou seja, aquelas que classificam o maior número possível de exemplos como positivos, enquanto ainda se ajustam aos fatos.
- Mantenha outro conjunto de hipóteses mais específicas que se ajustem aos dados, ou seja, aquelas que classificam o maior número possível de exemplos como negativos, enquanto ainda se ajustam aos fatos.
- Sempre que um novo exemplo negativo for observado, reforce todas as hipóteses mais gerais minimamente, de forma que o novo conjunto permaneça o mais geral possível, mas ainda se ajuste aos fatos.
- Sempre que um novo exemplo positivo for observado, relaxe todas as hipóteses mais específicas minimamente, de forma que o novo conjunto permaneça o mais específico possível, mas ainda se ajuste aos fatos.
- Prossiga até que reste apenas uma única hipótese. Essa será a resposta se o conceito-alvo estiver contido no nosso espaço de hipóteses.

No caso acima, o conjunto de hipóteses mais gerais seria:

```
{{?, Quente, ?, ?}, {Ensolarado, ?, ?, ?}}
```

Enquanto o conjunto de hipóteses mais específicas conteria apenas um membro:

```
{Ensolarado, Quente, Alta, ?}
```

Qualquer outro conceito que se ajuste aos dados será estritamente mais específico do que uma das hipóteses mais gerais e estritamente mais geral do que a hipótese mais específica.

(Para saber mais sobre isso, recomendo o livro Machine Learning, de Tom Mitchell, do qual este exemplo foi adaptado. [1])

Agora você pode perceber que o formato acima não consegue representar todos os possíveis conceitos. Por exemplo, "Jogue tênis quando o céu estiver ensolarado ou o ar estiver quente". Isso se encaixa nos dados, mas na representação do conceito definida acima, não há um conjunto de valores que descreva essa regra.

Claramente, nosso aprendiz de máquina não é muito abrangente. Por que não permitir que ele represente todos os conceitos possíveis, para poder aprender com a maior flexibilidade possível?

A variável Dias será composta por quatro variáveis: uma variável com 3 valores e três variáveis com 2 valores cada. Portanto, existem 3 × 2 × 2 × 2 = 24 Dias possíveis que poderíamos encontrar.

O formato fornecido para representar os conceitos nos permite exigir qualquer um desses valores para uma variável ou deixar a variável em aberto. Portanto, existem 4 × 3 × 3 × 3 = 108 conceitos nessa representação. Para que o algoritmo mais geral/mais específico funcione, precisamos começar com a hipótese mais específica: "nenhum exemplo é classificado positivamente". Somando isso, temos um total de 109 conceitos.

É suspeito que haja mais conceitos possíveis do que Dias possíveis? Certamente não. Afinal, um conceito pode ser visto como uma coleção de Dias. Um conceito pode ser entendido como o conjunto de dias que ele classifica como positivos ou, de forma isomórfica, o conjunto de dias que ele classifica como negativos.

Assim, o espaço de todos os conceitos possíveis que classificam os Dias é o conjunto de todos os conjuntos possíveis de Dias, cujo tamanho é 2^24 = 16.777.216.

Esse espaço completo inclui todos os conceitos que discutimos até agora. No entanto, também inclui conceitos como "Classifique positivamente apenas os exemplos {Ensolarado, Quente, Alta, Forte} e {Ensolarado, Quente, Alta, Fraco} e rejeite todos os outros" ou "Classifique negativamente apenas o exemplo {Chuvoso, Frio, Alta, Forte} e aceite todos os outros". Ele engloba conceitos sem uma representação compacta, sendo apenas uma lista plana do que é ou não é permitido.

Esse é o problema ao tentar construir um aprendiz indutivo "totalmente geral": ele não consegue aprender conceitos a menos que tenha visto todos os exemplos possíveis no espaço de instância.

Se adicionarmos mais atributos aos Dias, como a temperatura da água ou a previsão para amanhã, o número de dias possíveis crescerá exponencialmente com o aumento do número de atributos. Porém, isso não é um problema com nosso espaço conceitual restrito por ser possível reduzir um espaço grande usando um número logarítmico de exemplos.

Digamos que adicionemos o atributo "Água: {Quente, Fria}" aos dias, o que resultará em 48 Dias possíveis e 325 conceitos possíveis. Suponhamos que cada dia que observamos seja geralmente classificado como positivo por aproximadamente metade dos conceitos plausíveis atualmente e como negativo pela outra metade. Portanto, quando aprendemos a verdadeira classificação do exemplo, ela reduz pela metade o espaço de conceitos compatíveis. Assim, seriam necessários apenas 9 exemplos (2^9 = 512) para reduzir os 325 conceitos possíveis a um único.

Mesmo se os Dias tivessem quarenta atributos binários, ainda seria necessário apenas uma quantidade gerenciável de dados para reduzir os conceitos possíveis a um. Seriam necessários apenas sessenta e quatro exemplos, se cada exemplo fosse classificado como positivo por metade dos outros conceitos. Supondo, é claro, que a regra real seja uma que possamos representar!

Se você deseja considerar todas as possibilidades, boa sorte com isso. O espaço de todos os conceitos possíveis cresce superexponencialmente em relação ao número de atributos.

Quando você está lidando com dados que possuem quarenta atributos binários, o número de exemplos possíveis ultrapassa a marca de um trilhão — mas o número de conceitos possíveis cresce exponencialmente com um trilhão elevado à potência de dois. Para restringir esse espaço conceitual superexponencial, você teria que ver mais de um trilhão de exemplos antes de poder dizer o que está dentro e o que está fora. Na verdade, seria necessário observar todos os exemplos possíveis.

Estamos falando aqui de quarenta atributos binários, lembre-se. Quarenta bits, ou 5 bytes, que seriam simplesmente classificados como "Sim" ou "Não". Quarenta bits implicam em 2^40 possíveis exemplos e 2^240 conceitos possíveis que classificam esses exemplos como positivos ou negativos.

Portanto, no mundo real, onde os objetos requerem mais de 5 bytes para serem descritos, onde não estão disponíveis trilhões de exemplos e onde há ruído nos dados de treinamento, só podemos considerar conceitos altamente regulares. Uma mente humana — ou mesmo todo o universo observável — não é grande o suficiente para abranger todas as outras hipóteses.

A partir dessa perspectiva, a aprendizagem depende não apenas do viés indutivo, ele é quase todo indutivo — quando comparamos o número de conceitos descartados a priori com aqueles descartados por meras evidências.

Mas você deve estar se perguntando, o que isso tem a ver com o uso adequado das palavras? É exatamente a razão pela qual as palavras têm intensões e extensões. No último ensaio, concluí:

"A maneira de esculpir a realidade em suas articulações é delinear limites em torno de concentrações de densidade de probabilidades excepcionalmente altas".

Intencionalmente, omiti uma qualificação essencial nessa afirmação (ligeiramente editada), pois até agora não pude explicá-la. Uma afirmação mais precisa seria:

"A maneira de esculpir a realidade em suas articulações é delinear limites simples em torno de concentrações de densidade de probabilidades excepcionalmente alta no Espaço das Coisas".

Caso contrário, você estaria apenas distorcendo o Espaço das Coisas. Você criaria limites não contíguos e estranhos que agrupariam os exemplos observados, exemplos que não poderiam ser descritos em qualquer mensagem mais curta do que as próprias observações, e diria: "Isso é o que vi antes e é o que espero ver mais no futuro".

No mundo real, nada acima do nível molecular se repete exatamente. Sócrates possui uma forma muito semelhante a todos os outros humanos que eram vulneráveis à cicuta, mas não possui a mesma forma que eles. Portanto, sua suposição de que Sócrates é um "humano" depende da delimitação simples do agrupamento humano no Espaço das Coisas. Em vez de dizer: "Coisas com formato exatamente como [especificação 1 de formato de 5 megabytes] e com [muitas outras características], ou exatamente como [especificação 2 de formato de 5 megabytes] e [muitas outras características] são humanas".

Se você não estabelecer limites simples em torno de suas experiências, não poderá fazer inferências com base nelas. Então, você tenta descrever a "arte" com definições intencionais, como "aquilo que se destina a inspirar qualquer emoção complexa para inspirá-la" em vez de simplesmente apontar para uma longa lista de coisas que são ou não consideradas arte.

Na verdade, a afirmação anterior sobre "como esculpir a realidade em suas articulações" é um pouco como o problema do ovo e da galinha: você não pode avaliar a densidade das observações reais até que tenha feito pelo menos uma pequena esculpida. E a distribuição de probabilidades surge a partir da delimitação de limites, não o contrário — se você já tivesse a distribuição de probabilidades, teria tudo o que precisa para inferir, então por que se dar ao trabalho de delinear limites?

E isso levanta outra — sim, mais uma — razão para desconfiar da afirmação de que "você pode definir uma palavra do jeito que quiser". Quando você considera o tamanho superexponencial do Espaço de Conceitos, fica claro que selecionar um conceito específico para consideração é um ato de grande audácia — não apenas para nós, mas para qualquer mente com poder computacional limitado.

Apresentar-nos a palavra "wiggin", definida como "uma pessoa de cabelos pretos e olhos verdes", sem qualquer razão para elevar esse conceito particular ao nível de nossa atenção deliberada, é como um detetive dizendo: "Bem, não tenho absolutamente nenhuma evidência de qualquer tipo sobre quem poderia ter assassinado aqueles órfãos… nem mesmo uma intuição, entende?… mas vamos considerar John Q. Wiffeheim, da Rua Norkle, 1234, como suspeito?"

Referências

[1] Tom M. Mitchell, Machine Learning (McGraw-Hill Science/Engineering/Math, 1997).

177 — Independência condicional e Naive Bayes

Anteriormente, mencionei a respeito da informação mútua entre X e Y, representada por I(X;Y), que é a diferença entre a entropia da distribuição de probabilidade conjunta, H(X, Y), e as entropias das distribuições marginais, H(X) + H(Y).

Para ilustrar isso, considerei uma variável X com oito estados, X_1 a X_8, todos igualmente prováveis na ausência de evidências, e uma variável Y com os estados Y_1 a Y_4, também igualmente prováveis na ausência de evidências. Ao calcularmos as entropias marginais H(X) e H(Y), descobrimos que X tem uma entropia de 3 bits e Y tem uma entropia de 2 bits.

Entretanto, sabemos também que X e Y são ambos pares ou ambos ímpares, e isso é tudo o que sabemos sobre a relação entre eles. Assim, para a distribuição conjunta (X, Y), existem apenas 16 estados possíveis, todos igualmente prováveis, resultando em uma entropia conjunta de 4 bits. Isso representa uma deficiência de entropia de 1 bit em comparação com os 5 bits de entropia que teríamos se X e Y fossem independentes. Essa deficiência de entropia é chamada de informação mútua, por representar a informação que X fornece sobre Y, e vice-versa, reduzindo nossa incerteza sobre um, após aprendermos o outro.

Agora, suponha que exista uma terceira variável Z, a qual está perfeitamente correlacionada com a paridade de (X, Y). De fato, consideremos que Z é uma pergunta simples: “X e Y são pares ou ímpares?“

Se não tivermos evidências sobre X e Y, Z em si fornecerá 1 bit de informação. Há 1 bit de informação mútua entre Z e X, 1 bit de informação mútua entre Z e Y, e, como mencionado anteriormente, 1 bit de informação mútua entre X e Y. Então, qual será a entropia para todo o sistema (X, Y, Z)? Você poderia ingenuamente esperar que:

$$H(X,Y,Z) = H(X) + H(Y) + H(Z) - I(X;Z) - I(Z;Y) - I(X;Y)$$

Mas essa não é a realidade.

O sistema conjunto (X, Y, Z) possui apenas 16 estados possíveis — já que Z é apenas a pergunta “X e Y são pares ou ímpares?” — portanto, H(X, Y, Z) = 4 bits.

Mas se você calcular com a fórmula fornecida, obterá:

(3 + 2 + 1 - 1 - 1 - 1) bits = 3 bits = ERRADO!

Por quê? Pois se você tiver as informações mútuas entre X e Z, e as informações mútuas entre Z e Y, isso pode incluir parte das mesmas informações mútuas entre X e Y. Nesse caso, por exemplo, saber que X é par significa que Z também é par, e saber que Z é par significa que Y é par, mas essas são as mesmas informações que X nos fornece sobre Y. Portanto, estamos contando duas vezes parte de nossos conhecimentos, resultando em uma entropia menor.

A fórmula correta é (acredito):

$$H(X,Y,Z) = H(X) + H(Y) + H(Z) - I(X;Z) - I(Z;Y) - I(X;Y|Z)$$

Aqui, o último termo, I(X;Y|Z), representa “a informação que X nos fornece sobre Y, dado que já conhecemos Z”. Nesse caso, X não nos fornece nenhuma informação adicional sobre Y, uma vez que já conhecemos Z, então esse termo é igual a zero — e a equação fornece a resposta correta. Interessante, não é?

“Não”, você responderia corretamente, “porque você não me explicou como calcular I(X;Y|Z), apenas apresentou um argumento verbal de que deveria ser zero”.

Calculamos I(X; Y|Z) exatamente como você esperaria. Sabemos que:

$$I(X;Y) = H(X) \ + \ H(Y) \quad H(X,Y)$$

Então:

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

Agora, suponho que você queira saber como calcular a entropia condicional. Bem, a fórmula original para a entropia é:

$$H(S) = \sum_{i} -P(S_i) \times \log_2(P(S_i))$$

Se aprendemos um novo fato Z_0, nossa incerteza remanescente sobre S seria:

$$H(S|Z_0) = \sum_{i} -P(S_i|Z_0) \log_2(P(S_i|Z_0))$$

Portanto, se vamos aprender um novo fato Z, mas ainda não sabemos qual é, então, em média, esperamos ter essa incerteza sobre S posteriormente:

$$H(S|Z) = \sum_{j} \left(P(Z_j) \sum_{i} -P(S_i|Z_j) \log_2(P(S_i|Z_j)) \right)$$

E é assim que se calculam as entropias condicionais; a partir delas, podemos obter a informação mútua condicional.

Existem diversos teoremas auxiliares relacionados, como

$$H(X|Y) = H(X,Y) - H(Y)$$

$$\text{if } I(X;Z) = 0 \text{ and } I(Y;X|Z) = 0 \text{ then } I(X;Y) = 0$$

mas não entrarei em detalhes sobre isso.

“Mas”, você pergunta, ”o que isso tem a ver com a natureza das palavras e sua estrutura bayesiana oculta?“

Fico extremamente feliz que você tenha feito essa pergunta, pois estava planejando explicar isso, gostando você ou não. No entanto, antes disso, há mais alguns pontos preliminares.

Você se lembrará — sim, você se lembrará — que existe uma dualidade entre informação mútua e evidência bayesiana. A informação mútua é positiva se e somente se a probabilidade de pelo menos alguns eventos conjuntos P(x, y) não for igual ao produto das probabilidades dos eventos separados P(x)P(y). Isso, por sua vez, é exatamente equivalente à condição de que exista evidência bayesiana entre x e y:

$$I(X;Y) > 0 \Rightarrow$$

$$P(x,y) \neq P(x)\ P(x)$$

$$\frac{P(x,y)}{P\ (y)} \neq P(x|z)$$

$$P\ (x|y) \neq P(x)$$

Se você está condicionando em Z, basta ajustar toda a derivação de acordo:

$$I\ (X;\ Y|Z) > 0 \Rightarrow$$

$$P(x,y|z) \neq P\ (x|z)P(y|z)$$

$$\frac{P(x,y|z)}{P\ (y|z)} \neq P(x|z)$$

$$\frac{(P(x,y,z)/P(z))}{(P(y,z)/P(z))} \neq P(x|z)$$

$$\frac{P(x,y,z)}{P\ (y,z)} \neq P(x|z)$$

$$P(x|y,z) \neq P(x|z)$$

Sua última linha diz: “Mesmo conhecendo Z, aprender Y ainda muda nossas crenças sobre X.”

Inversamente, no nosso caso original em que Z é “par” ou “ímpar”, Z isola X de Y — ou seja, se sabemos que Z é “par”, descobrir que Y está no estado Y_4 não nos informa nada sobre se X é X_2, X_4, X_6 ou X_8. Da mesma forma, se sabemos que Z é “ímpar” e descobrimos que X é X_5, isso não nos diz mais nada sobre se Y é Y_1 ou Y_3. Aprender Z tornou X e Y condicionalmente independentes.

A independência condicional é um conceito extremamente importante na teoria da probabilidade — para citar apenas um exemplo, sem a independência condicional, o universo não teria estrutura.

Aqui, porém, planejo falar apenas sobre um tipo particular de independência condicional, o caso de uma variável central que oculta outras variáveis ao seu redor, como um corpo central com tentáculos.

Consideremos cinco variáveis: U, V, W, X e Y; além disso, suponha que para cada par dessas variáveis, uma variável seja evidência em relação à outra. Se você selecionar U e W, por exemplo, descobrir que $U = U_1$ lhe dirá algo que você não sabia anteriormente sobre a probabilidade de que $W = W_1$.

Uma confusão inferencial incontrolável? A evidência está enlouquecendo? Não necessariamente.

Suponhamos que U seja “fala uma língua”, V seja “dois braços e dez dígitos”, W seja “usa roupas”, X seja “envenenável por cicuta” e Y seja “sangue vermelho”. Agora, se você encontrar algo no mundo — pode ser uma maçã ou uma pedra — e descobrir que esse algo fala chinês, é provável avaliar uma probabilidade muito maior de que ele use roupas. E se descobrir que esse algo não pode ser envenenado por cicuta, você avaliará uma probabilidade um pouco menor de que ele tenha sangue vermelho.

Agora, algumas dessas regras são mais fortes do que outras. Há o caso de Fred, que perdeu um dedo

devido a um acidente vulcânico, e o caso de Barney, o bebê, que ainda não fala, e o caso de Irving, o IRCBot, que emite sentenças, mas não tem sangue. Portanto, se descobrirmos que algo específico não está vestindo roupas, isso não oculta completamente o que a capacidade de fala desse algo pode nos dizer sobre sua cor de sangue. Se o algo não usa roupas, mas fala, talvez seja a "Nude Nellie".

Isso torna o caso mais interessante do que, por exemplo, cinco variáveis inteiras que são todas ímpares ou todas pares, mas não correlacionadas. Nesse caso, conhecer qualquer uma das variáveis eliminaria tudo o que o conhecimento de uma segunda variável poderia nos dizer sobre uma terceira variável.

Mas aqui temos dependências que não desaparecem assim que aprendemos apenas uma variável, como é o caso da "Nude Nellie". Então, isso seria uma inconveniência inferencial incontrolável?

Não tenha medo! Pois pode haver uma sexta variável Z, que, se a conhecermos, realmente separaria cada par de variáveis umas das outras. Pode haver uma variável Z — mesmo que tenhamos que construí-la em vez de observá-la diretamente — tal que:

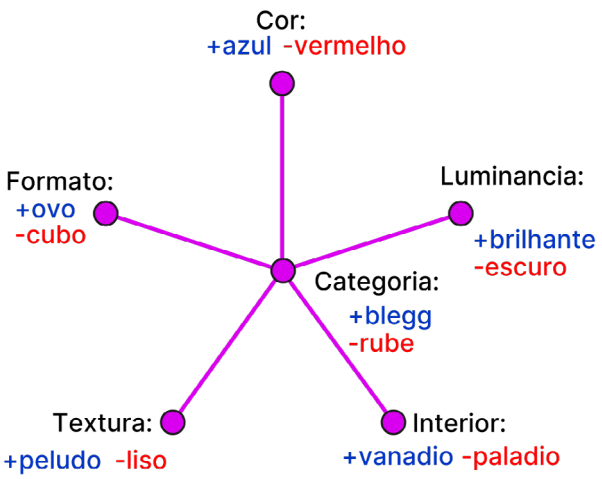
$$P(U|V,W,X,Y,Z) = P(U|Z)$$
$$P(V|U,W,X,Y,Z) = P(V|Z)$$
$$P(W|U,V,X,Y,Z) = P(W|Z)$$
$$\vdots$$

Talvez, considerando que algo seja "humano", as probabilidades de falar, vestir roupas e ter o número padrão de dedos sejam independentes. Por exemplo, Fred pode estar sem um dedo, mas isso não torna mais provável que ele seja um nudista do que qualquer outra pessoa; Nude Nellie nunca usa roupas, mas saber disso não diminui a probabilidade de ela falar; e o bebê Barney ainda não fala, mas não perdeu nenhum membro.

Isso é conhecido como método "Naive Bayes", porque geralmente não é totalmente verdadeiro, mas fingir que é verdade pode simplificar os cálculos. Não consideramos separadamente a influência das roupas na capacidade de falar, considerando o número de dedos. Simplesmente utilizamos todas as informações observadas para determinar a probabilidade de algo ser humano (ou alternativamente, outra coisa, como um chimpanzé ou um robô), e em seguida, utilizamos nossas crenças sobre a classe central para fazer previsões sobre coisas que ainda não vimos, como a vulnerabilidade à cicuta.

Quaisquer observações de U, V, W, X e Y atuam apenas como evidências para a variável de classe central Z, e então utilizamos a distribuição posterior em Z para fazer previsões sobre variáveis não observadas em U, V, W, X e Y.

Isso soa familiar? Deveria:



Rede 2

De fato, se você utilizar o tipo adequado de unidades em uma rede neural, essa “rede neural” acaba sendo exatamente equivalente ao método Naive Bayes em termos matemáticos. A unidade central requer apenas um limite logístico — uma resposta em forma de curva S — e os pesos das entradas precisam corresponder aos logaritmos das razões de verossimilhança, entre outros.

De fato, é uma suposição razoável que uma das razões pelas quais a resposta logística funciona tão bem em redes neurais seja porque ela permite que o algoritmo introduza um pouco de raciocínio bayesiano quando os projetistas não estão observando.

Apenas porque alguém está apresentando um algoritmo chamado de “rede neural” com termos como “desalinhado” e “emergente” associados a ele, orgulhosamente admitindo que não têm ideia de como a rede aprendida funciona — não devemos presumir que seu pequeno algoritmo de IA esteja além dos princípios lógicos. Na realidade, esse paradigma improvisado, se funcionar, terá uma estrutura bayesiana; pode até ser exatamente equivalente a um algoritmo do tipo chamado “Bayesiano”.

Mesmo que, na superfície, não pareça seguir a abordagem bayesiana.

E é aí que os bayesianos começam a explicar detalhadamente como o algoritmo funciona, quais suposições subjacentes ele reflete, quais regularidades ambientais ele explora, onde tem sucesso e onde falha, e até mesmo atribuindo significado compreensível aos pesos aprendidos pela rede.

Pode ser um tanto decepcionante, não é?

178 — Palavras como cabos de pincel mental

Suponha que eu lhe diga: “É algo muito estranho: as luminárias deste hotel têm lâmpadas triangulares”.

Você pode ou não ter visualizado — se ainda não o fez, faça-o agora — como seria uma “lâmpada triangular” em sua mente?

Em sua mente, o vidro tinha bordas afiadas ou lisas?

Quando a expressão “lâmpada triangular” passou pela primeira vez pela minha mente — não, o hotel não as tem — então, da melhor maneira que minha introspecção pôde determinar, inicialmente visualizei uma lâmpada piramidal com bordas afiadas e, em seguida (quase que imediatamente), as bordas foram suavizadas e, em seguida, minha mente gerou um ciclo de uma luminária fluorescente em forma de triângulo com bordas lisas como uma alternativa.

Até onde pude perceber, nenhum pensamento deliberado ou verbal estava envolvido — apenas um reflexo não verbal se distanciando da imagem mental imaginada com vidro afiado, cujo problema de projeto foi resolvido antes mesmo que eu pudesse formular palavras.

Acredite ou não, houve um debate sério, durante algumas décadas, sobre se as pessoas realmente tinham imagens mentais em suas mentes — uma imagem real de uma cadeira em algum lugar — ou se as pessoas apenas ingenuamente acreditavam terem imagens mentais (sendo enganadas pela “introspecção”, uma atividade proibida muito ruim), quando, na verdade, possuíam apenas um pequeno rótulo mental de “cadeira” como um *token LISP* (um símbolo conceitual ativo) em seus cérebros.

Estou tentando evitar dizer algo como “Que bobagem espetacular” porque sempre há o viés retrospectivo a ser considerado, mas é realmente uma bobagem espetacular.

Esse paradigma acadêmico, em minha opinião, foi em grande parte uma herança insana do behaviorismo, que negava a existência de pensamentos nas pessoas e buscava explicar todos os fenômenos humanos como “reflexos”, inclusive a fala. O behaviorismo provavelmente merece sua própria discussão em algum momento, por ser uma deturpação do racionalismo, mas isso não é o foco deste texto.

“Você considera isso ‘bobagem’”, você pode perguntar, “mas como você sabe que seu cérebro representa imagens visuais? É apenas porque você pode fechar os olhos e vê-las?“

Essa pergunta costumava ser mais difícil de responder na época da controvérsia. Se quiséssemos provar a existência de imagens mentais “cientificamente”, em vez de apenas com base na introspecção, teríamos que inferir a existência de imagens mentais a partir de experimentos como este: apresentar dois objetos aos participantes e perguntar se um pode ser girado em correspondência com o outro. O tempo de resposta é diretamente proporcional ao ângulo de rotação necessário. Isso é fácil de explicar se você realmente estiver visualizando a imagem e girando-a continuamente

a uma velocidade constante, mas difícil de explicar se você estiver apenas verificando as características conceituais da imagem.

Hoje em dia, podemos até fazer uma neuroimagem das representações visuais no córtex cerebral. Portanto, sim, o cérebro realmente representa uma imagem detalhada do que vemos ou imaginamos. Consulte “Imagem e Cérebro: A Resolução do Debate sobre Imaginação”, de Stephen Kosslyn, para obter mais informações [1].

Uma parte do motivo pelo qual as pessoas têm dificuldades com as palavras é que elas não percebem a complexidade oculta por trás delas.

Você consegue visualizar um “cachorro verde”? Você consegue imaginar uma “maçã com queijo”?

“Maçã“ não é apenas uma sequência de duas sílabas ou cinco letras. É uma sombra. Isso é apenas a ponta do iceberg.

As palavras, ou melhor, os conceitos por trás delas, são como pincéis — você pode usá-los para pintar imagens em sua própria mente. Pintar literalmente, se você estiver usando conceitos para criar uma imagem em seu córtex visual. E, por meio do uso de rótulos compartilhados, você pode alcançar a mente de outra pessoa e pegar seus pincéis para pintar imagens em suas mentes — esboçar um cachorrinho verde em seu córtex visual.

No entanto, não pense que, ao enviar sílabas pelo ar ou letras pela internet, são essas sílabas ou letras que pintam imagens em seu córtex visual. Isso requer instruções complexas que não podem ser contidas apenas na sequência de letras. “Maçã” tem 5 bytes, e criar uma imagem de uma maçã a partir do zero exigiria mais informações do que isso.

“Maçã” é apenas a etiqueta anexada ao conceito verdadeiro e sem palavras de maçã, que pode evocar uma imagem em seu córtex visual, fazer uma conexão com outros conceitos como “queijo”, permitir que você reconheça uma maçã quando a vê ou experimentar seu sabor arquetípico em uma torta de maçã, talvez até mesmo influenciar suas ações, levando você a comer uma maçã…

E não é tão simples quanto apenas recuperar uma imagem da memória. Ou como você conseguiria visualizar combinações como uma “lâmpada triangular” — impondo a triangularidade às lâmpadas, mantendo a essência de ambas, mesmo que você nunca tenha visto tal coisa em sua vida?

Não cometa o mesmo erro que os behavioristas cometeram. Há muito mais na fala do que apenas o som no ar. Os rótulos são apenas ponteiros — “procure na área de memória 1387540”. Mais cedo ou mais tarde, quando você recebe um ponteiro, chega a hora de desreferenciá-lo e realmente buscar na área de memória 1387540.

Para onde aponta uma palavra?

Referências

[1] Stephen M. Kosslyn, Image and Brain: The Resolution of the Imagery Debate (Cambridge, MA: MIT Press, 1994).

179 — Falácias de pergunta variável

Albert: “Toda vez que ouço uma árvore cair, ela emite um som, então imagino que outras árvores caindo também emitem sons. Não acredito que o mundo mude quando não estou olhando.“

Barry: “Espere um minuto. Se ninguém ouve, como pode ser considerado um som?“

Enquanto eu escrevia o diálogo entre Albert e Barry em sua disputa sobre se uma árvore caindo em uma floresta deserta emite som, às vezes me via perdendo a empatia pelos meus personagens. Eu começava a perder a noção de porque alguém argumentaria dessa forma, embora já tivesse visto isso acontecer muitas vezes.

Nessas ocasiões, eu me repetia: “A árvore que cai faz barulho ou não faz!” para restaurar meu sentimento emprestado de indignação.

(P ou ¬P) nem sempre é uma heurística confiável se você substituir P por frases arbitrárias em inglês. “Esta frase é falsa” não pode ser consistentemente considerada verdadeira ou falsa. E então temos o clássico: “Você parou de bater em sua esposa?”

Agora, se você é um matemático e alguém que acredita na lógica clássica (em vez da lógica intuicionista), existem maneiras de continuar insistindo que (P ou ¬P) é um teorema: por exemplo, ao dizer que “Esta frase é falsa“ não é uma frase.

Mas tais resoluções são sutis, o suficiente para demonstrar a necessidade de sutileza. Não se pode simplesmente avançar em todas as ocasiões com “Ou faz, ou não!”

Então, a árvore que cai emite som ou não?

Certamente, 2 + 2 = X ou não? Bem, talvez, se for realmente o mesmo X, o mesmo 2, o mesmo + e =. Se X for avaliado como 5 em algumas ocasiões e 4 em outras, sua indignação pode estar equivocada.

Para afirmar que (P ou ¬P) deve ser uma verdade necessária, o símbolo P deve representar exatamente a mesma coisa em ambas as metades do dilema. “Ou a queda emite som, ou não!” — mas se o Albert::som não é o mesmo que o Barry::som, não há nada paradoxal no fato de a árvore fazer um Albert::som, mas não um Barry::som.

(A expressão “::“ é algo que aprendi nos meus dias de C++ para evitar colisões de namespaces. Se você tem dois pacotes diferentes que definem uma classe Som, você pode escrever Pacote1::som para especificar qual Som você quer dizer. A expressão não é amplamente conhecida, eu acho; é uma pena, porque muitas vezes gostaria de poder usá-la na escrita.)

A variabilidade pode ser sutil: Albert e Barry podem verificar cuidadosamente se é a mesma árvore, na mesma floresta e na mesma ocasião de queda, apenas para garantir que realmente têm um desacordo substantivo sobre o mesmo evento. E então, eles esquecem de verificar se estão comparando esse evento exatamente com o mesmo conceito.

Pense no supermercado que você visita com mais frequência: ele fica do lado esquerdo da rua ou do lado direito? Mas, é claro, não existe o “lado esquerdo” da rua, apenas o seu lado esquerdo conforme você o percorre em uma direção específica. Muitas das palavras que usamos são, na verdade, funções de variáveis implicitamente fornecidas pelo contexto.

Na verdade, é um problema e tanto, exigindo muito trabalho, lidar com esse tipo de problema em um programa de Inteligência Artificial destinado a analisar a linguagem — o fenômeno conhecido pelo nome de "dêixis do locutor".

"Martin disse a Bob que o prédio ficava à sua esquerda." Mas "esquerda" é uma palavra funcional avaliada com uma variável dependente do locutor, invisivelmente captada do contexto circundante. De quem é a "esquerda", de Bob ou de Martin?

As variáveis em uma falácia de pergunta variável geralmente não são rotuladas de maneira organizada — não é tão simples quanto "Diga, você acha que Z + 2 é igual a 6?"

Se uma colisão de namespaces introduzir dois conceitos diferentes que se parecem com "o mesmo conceito" porque têm o mesmo nome — ou uma compactação de mapeamento introduzir dois eventos diferentes que se parecem com o mesmo evento porque não possuem arquivos mentais separados — ou a mesma função for avaliada em diferentes contextos — então a própria realidade se torna multiforme, mutável. Pelo menos é assim que o algoritmo se sente por dentro. O olho da sua mente vê o mapa, não o território diretamente.

Se você tiver uma pergunta com uma variável oculta, que é avaliada com diferentes expressões em diferentes contextos, parece que a própria realidade é instável — o que o olho da sua mente vê muda dependendo de onde ele olha.

Isso muitas vezes confunde os alunos de graduação (e professores pós-modernistas) que descobrem uma frase com mais de uma interpretação; eles acham que descobriram uma parte instável da realidade.

"Oh, meu Deus! 'O Sol gira em torno da Terra' é verdadeiro para o caçador coletor Hunga, mas para o astrônomo Amara, 'O Sol gira em torno da Terra' é falso! Não há verdade fixa!" A desconstrução desse pensamento juvenil fica como exercício para o leitor.

No entanto, até eu inicialmente me vi escrevendo "Se X é 5 em algumas ocasiões e 4 em outras, a sentença '2 + 2 = X' pode não ter um valor de verdade fixo". Não há uma sentença com um valor de verdade variável. "2 + 2 = X" não tem valor de verdade. Não é uma proposição, ainda não, não como os matemáticos definem a proposição, assim como "2 + 2 =" não é uma proposição, ou "Fred saltou sobre o" não é uma frase gramatical.

Mas essa falácia tende a se infiltrar, mesmo quando você supostamente sabe mais, porque, bem, é assim que o algoritmo se sente por dentro.

180 – 37 maneiras pelas quais as palavras podem estar erradas

Alguns leitores certamente declararão que um título mais apropriado para este ensaio seria “37 maneiras de usar palavras imprudentemente” ou “37 maneiras pelas quais o uso inadequado de categorias pode ter efeitos negativos na sua cognição”.

No entanto, uma das principais lições dessa extensa lista é que afirmar “Não há como minha escolha de X estar ‘errada’” é quase sempre um equívoco, na prática, independentemente da teoria. Sempre há a possibilidade de estarmos errados. Mesmo quando teoricamente parece impossível estar errado, ainda assim podemos estar. Não existe um passe livre para qualquer coisa que façamos. Isso faz parte da vida.

Além disso, posso atribuir o significado que eu quiser à palavra “errado” — afinal, uma palavra não pode estar errada.

Pessoalmente, considero justificável utilizar a palavra “errado” quando:

1. Uma palavra não consegue se conectar à realidade em primeiro lugar. Sócrates é um “framster“? Sim ou não? (A Parábola da Adaga)
2. Seu argumento, se fosse válido, poderia forçar a realidade a seguir um caminho diferente, escolhendo uma definição de palavra diferente. Sócrates é um ser humano e, por definição, os seres humanos são mortais. Então, se definíssemos que os humanos não são mortais, Sócrates viveria para sempre? (A Parábola da Cicuta)
3. Você tenta estabelecer uma proposição empírica como verdadeira “por definição”. Sócrates é um ser humano e, por definição, os seres humanos são mortais. Então, é logicamente correto prever empiricamente que Sócrates desmaiaria se bebesse cicuta? Parece haver mundos logicamente possíveis e não autocontraditórios nos quais Sócrates não desmaia — onde ele é imune à cicuta devido a uma peculiaridade bioquímica, por exemplo. As verdades lógicas são válidas em todos os mundos possíveis e, portanto, não indicam em qual mundo possível vivemos — e qualquer coisa que possamos estabelecer “por definição” é uma verdade lógica. (A Parábola da Cicuta)
4. Você inconscientemente rotula algo conforme a definição verbal que acabou de dar. Você sabe perfeitamente bem que Bob é “humano”, embora, conforme a sua definição, você nunca possa chamar Bob de “humano” sem primeiro observar que ele é mortal. (A Parábola da Cicuta)
5. O ato de atribuir um rótulo a algo por meio de uma palavra mascara uma inferência indutiva questionável que você está fazendo. Se os últimos 11 objetos em forma de ovo desenhados forem azuis e os últimos 8 cubos desenhados forem vermelhos, é uma questão de indução dizer que essa regra será válida no futuro. Porém, se você chamar os ovos azuis de “bleggs” e os cubos vermelhos de “rubes”, você pode colocar a mão em um barril, sentir a forma de um ovo e pensar “Ah, um blegg”. (Palavras como inferências ocultas)
6. Você tenta definir uma palavra usando palavras, que, por sua vez, são definidas com palavras cada vez mais abstratas, sem poder apontar um exemplo concreto. “O que é vermelho?” “Vermelho é uma cor.” “O que é uma cor?” “É uma propriedade de uma coisa.” “O que é uma coisa? O que é uma propriedade?“ Nunca lhe ocorre apontar para uma placa de pare ou uma maçã. (Extensões e Intensões)
7. A extensão não corresponde à intenção. Não estamos conscientemente cientes de nossa identificação de uma luz vermelha no céu como “Marte”, o que provavelmente ocorrerá independentemente de sua tentativa de definir “Marte” como “o Deus da Guerra” . (Extensões e Intensões)
8. Sua definição verbal não captura mais do que uma pequena fração das características compartilhadas da categoria, mas você tenta raciocinar como se isso fosse verdade. Quando os filósofos da Acade-

mia de Platão afirmaram que a melhor definição de um humano era um "bípede sem penas", diz-se que Diógenes, o Cínico, mostrou uma galinha depenada e declarou: "Aqui está o Homem de Platão". Os platônicos prontamente mudaram sua definição para "um bípede sem penas com unhas largas". (Aglomerados de Similaridade)

9. Você tenta tratar a associação de categorias como algo absoluto, ignorando a existência de subgrupos mais ou menos típicos. Patos e pinguins são aves menos típicas do que tordos e pombos. Curiosamente, um experimento entre grupos mostrou que os participantes acreditavam que uma doença era mais propensa a se espalhar de tordos para patos em uma ilha do que de patos para tordos. (Tipicidade e Similaridade Assimétrica)
10. Uma definição verbal funciona bem o suficiente, na prática, para identificar o conjunto pretendido de coisas semelhantes, mas sempre há exceções. Nem todo ser humano tem dez dedos, usa roupas ou usa linguagem; no entanto, se você procurar um grupo empírico de coisas que compartilham essas características, obterá informações suficientes para que o ocasional humano de nove dedos não o engane. (A Estrutura de Agrupamento do Espaço das Coisas)
11. Você pergunta se algo "é" ou "não é" membro de uma categoria, mas não consegue identificar a pergunta que realmente deseja responder. O que é um "homem"? Barney, o bebê, é um "homem"? A resposta "correta" pode depender consideravelmente de saber se a pergunta que você realmente deseja responder é "Seria bom alimentar Barney com cicuta?" ou "Barney será um bom marido?" (Consultas disfarçadas).
12. Você trata as categorias hierárquicas percebidas intuitivamente como a única maneira correta de analisar o mundo, sem perceber que outras formas de inferência estatística são possíveis, mesmo que seu cérebro não as utilize. É muito mais fácil para um humano perceber se um objeto é um "blegg" ou um "rube" do que perceber que objetos vermelhos nunca brilham no escuro, mas objetos com pelos vermelhos têm todas as outras características de bleggs. Outros algoritmos estatísticos funcionam de maneira diferente. (Categorias neurais)
13. Você fala sobre categorias como se fossem maná caído do reino platônico, em vez de inferências implementadas em um cérebro real. Os antigos filósofos diziam "Sócrates é um homem", não "Meu cérebro classifica perceptualmente Sócrates como correspondente ao conceito 'humano'". (Como um algoritmo se sente por dentro).
14. Você argumenta sobre a associação a uma categoria mesmo após eliminar todas as perguntas que poderiam depender de uma inferência baseada em categoria. Após observar que um objeto é azul, em forma de ovo, peludo, flexível, opaco, luminescente e contendo paládio, o que resta perguntar ao argumentar: "É um blegg?" Mas, se a rede neural de categorização do seu cérebro contiver uma unidade central (metafórica) correspondente à inferência de ser um blegg, ainda pode parecer que há uma pergunta remanescente. (Como um algoritmo se sente por dentro).
15. Você permite que um argumento passe a ser sobre definições, mesmo quando não é o assunto original da discussão. Se, antes de iniciar uma disputa sobre se uma árvore caindo em uma floresta deserta produz um "som", você perguntar aos dois futuros debatedores se eles acham que um "som" deve ser definido como "vibrações acústicas" ou "experiências auditivas", provavelmente eles pediriam que você jogasse uma moeda. Somente após o início da discussão é que a definição de uma palavra se torna politicamente carregada. (Definições em disputa)
16. Você pensa que uma palavra tem um significado intrínseco, como uma propriedade da própria palavra, em vez de haver um rótulo que seu cérebro associa a um conceito específico. Quando alguém grita "Caramba! Um tigre!", a evolução não favoreceria um organismo que pensa: "Mm… acabei de ouvir as sílabas "Ti" e "Grr" que meus companheiros de tribo associam com seus próprios conceitos internos do meu conceito de tigre, e agora… aiiieeee crunch crunch gulp". Assim, o cérebro pega um atalho e parece que o significado de tigre é uma propriedade intrínseca do rótulo em si. As pessoas debatem sobre o significado correto de um rótulo, como "som". (Sinta o Significado)
17. Você discute os significados de uma palavra, mesmo quando todos os lados entendem perfeitamente o que os outros estão tentando dizer. A capacidade humana de associar rótulos a conceitos é uma ferramenta de comunicação. Quando as pessoas desejam se comunicar, é difícil interrompê-las. Se não tivermos uma linguagem comum, usaremos desenhos na areia. Quando cada um entende o que está na mente do outro, está tudo resolvido. (O argumento do uso comum)
18. Você recorre a um dicionário no meio de um argumento empírico ou moral. Os editores de dicionários são historiadores do uso, não legisladores da linguagem. Se a definição comum contiver um pro-

blema — como “Marte” sendo definido como o Deus da Guerra, ou “golfinho” sendo definido como uma espécie de peixe, ou “Negros” sendo definido como uma categoria separada dos humanos — o dicionário refletirá o erro comum. (O argumento do uso comum)

19. Você recorre a um dicionário no meio de qualquer discussão. Sério, por que você acha que os editores de dicionários são uma autoridade sobre se o “ateísmo” é uma “religião” ou qualquer outra questão? Se você tiver alguma questão substancial em jogo, acredita realmente que os editores de dicionários possuem a sabedoria final que resolve o argumento? (O argumento do uso comum)
20. Você desafia o uso comum sem motivo, tornando desnecessariamente difícil para os outros entenderem você. Isso torna a comunicação menos eficaz. (O argumento do uso comum)
21. Você usa renomeações complexas para criar a ilusão de inferência. Um “humano” é definido como um “bípede mortal sem penas”? Então escreva: todos os [bípedes mortais sem penas] são mortais; Sócrates é um [bípede mortal sem penas], portanto Sócrates é mortal.” Parece menos impressionante quando expresso dessa forma, não parece? (Rótulos Vazios)
22. Você entra em discussões que poderiam ser evitadas se você simplesmente não usasse certas palavras. Por exemplo, se Albert e Barry concordarem em não usar a palavra “som”, eles podem expressar suas opiniões de maneira mais clara dizendo “Uma árvore caindo em uma floresta deserta gera vibrações acústicas” e “Uma árvore caindo em uma floresta deserta não gera experiências auditivas”. Quando uma palavra apresenta um problema, a solução mais simples é eliminá-la e usar outras formas de comunicação. (Jogando Tabu com as suas palavras)
23. A existência de uma palavra correta impede que você veja os detalhes daquilo em que está tentando pensar. Por exemplo, o que realmente acontece nas escolas quando você para de chamá-las de “educação”? O que é um diploma se você parar de chamá-lo de “diploma”? Se uma moeda cair em “cara”, qual é a sua orientação radial? O que é “verdade” se você não pode empregar termos como “exato” ou “correto” ou “representar” ou “refletir” ou “semântico” ou “acreditar” ou “conhecimento” ou “mapa” ou “real” ou qualquer outro termo simples? (Substitua o Símbolo pela Substância)
24. Você tem apenas uma palavra, mas há duas ou mais coisas diferentes na realidade, de modo que todos os fatos sobre elas são despejados em um único balde mental indiferenciado. Faz parte do trabalho comum de um detetive se perguntar se talvez Carol pinte o cabelo. Mas é preciso ser um detetive mais sutil para se perguntar se existem duas Carols, de modo que a Carol estava vestida de vermelho, não é a mesma Carol que tinha cabelo preto. (Falácias de Compressão)
25. Você enxerga padrões onde eles não existem, extrapolando outras características das definições mesmo quando não há similaridade nessa dimensão. Por exemplo, a crença no Japão de que pessoas com tipo sanguíneo A são sérias e criativas, as do tipo B são selvagens e alegres, as do tipo O são agradáveis e sociáveis, e as do tipo AB são frias e controladas. Categorizar dessa forma pode levar a consequências injustas ou imprecisas. (A categorização tem consequências)
26. Você tenta infiltrar-se nas conotações de uma palavra, argumentando com base em uma definição que não inclui essas conotações. Por exemplo, se a palavra “wiggin” for definida como uma pessoa com olhos verdes e cabelos pretos. A palavra “wiggin” também tem a conotação de alguém que comete crimes e lança esquilos bebês fofos, mas essa parte não está no dicionário. Então, você aponta para alguém com olhos verdes e cabelos pretos e dizer: “Olhos verdes? Cabelo preto? Veja, eu disse que ele é um wiggin! Veja, em seguida, ele roubará os talheres de prata.“ (Esgueirando-se em conotações)
27. Você afirma: “X, por definição, é um Y!” Nessas ocasiões, você geralmente está tentando introduzir uma conotação de Y que não estava na definição original. Por exemplo, se você define “humano” como um “bípede sem penas” e aponta para Sócrates dizendo “Sem penas — duas pernas — ele deve ser humano!”, mas o que realmente importa é outra característica, como a mortalidade, o outro lado do argumento pode responder: “O que você quer dizer, Sócrates tem duas pernas? É isso que estamos discutindo em primeiro lugar!“ (Argumentando “por definição”).
28. Você afirma “Ps, por definição, são Qs!” Se você vir Sócrates no campo colhendo ervas que podem conferir resistência à cicuta, não faz sentido argumentar: “Os humanos, por definição, são mortais!” O principal momento em que você sente a necessidade de apertar o torno ao insistir que algo é verdadeiro “por definição” ocorre quando há outras informações que colocam em dúvida a inferência padrão. (Argumentando “por definição”).
29. Você tenta estabelecer associação em um agrupamento empírico “por definição”. Você não sentiria a necessidade de dizer: “O hinduísmo, por definição, é uma religião!” porque, bem, é óbvio que o

hinduísmo é uma religião. Não é apenas uma religião "por definição", é uma religião de verdade. O ateísmo não se assemelha aos membros centrais do agrupamento "religião". Se não fosse pelo fato de o ateísmo ser considerado uma religião por definição, poderíamos pensar que o ateísmo não é uma religião. É por isso que é necessário refutar qualquer oposição apontando que "Ateísmo é uma religião" é verdadeiro por definição, pois não seria verdade de nenhuma outra forma. (Argumentando "por definição")

30. Sua definição traça um limite em torno de coisas que realmente não estão relacionadas. Você pode alegar, se quiser, que está definindo a palavra "peixe" para se referir a salmões, lebistes, tubarões, golfinhos e trutas, mas não as águas-vivas ou algas. Você pode alegar, se quiser, que essa é apenas uma lista e uma lista não pode estar "errada". Ou você pode parar de brincar e admitir que cometeu um erro ao incluir golfinhos na lista de peixes. (Onde traçar o limite?)
31. Você usa uma palavra curta para algo que não precisa descrever com frequência, ou uma palavra longa para algo que precisa descrever com frequência. Isso pode resultar em pensamento ineficiente ou até mesmo em aplicações incorretas da Navalha de Ocam, se sua mente acreditar que frases curtas parecem "mais simples". O que soa mais plausível, "Deus fez um milagre" ou "Uma entidade sobrenatural criadora do universo suspendeu temporariamente as leis da física"? (Entropia e códigos curtos)
32. Você estabelece um limite em torno de um volume de espaço onde não há densidade maior do que o normal, o que significa que a palavra associada não corresponde a nenhuma inferência Bayesiana executável. Uma vez que pessoas com olhos verdes não são mais propensas a ter cabelo preto, ou vice-versa, e não compartilham nenhuma outra característica em comum, por que ter uma palavra para "peruca"? (Informação Mútua e Densidade no Espaço das Coisas)
33. Você estabelece um limite não-simples sem nenhum motivo aparente para fazê-lo. O ato de definir uma palavra para se referir a todos os humanos, exceto os negros, parece um tanto suspeito. Se você não apresentar razões para traçar esse limite específico, tentar criar uma palavra "arbitrária" nesse ponto é como um detetive dizendo: "Bem, não tenho nenhum apoio para quem poderia ter assassinado aqueles órfãos…, mas vamos considerar John Q. Wiffeheim como suspeito?" (Espaço conceitual superexponencial e palavras simples)
34. Você usa a categorização para fazer inferências sobre propriedades que não possuem a estrutura empírica apropriada, ou seja, independência condicional, dado o conhecimento da classe, que pode ser bem aproximada pelo Naive Bayes. Não estou tentando resumir isso. Apenas leia o ensaio. (Independência Condicional e Naive Bayes)
35. Você acredita que as palavras são como pequenos símbolos LISP em sua mente, em vez de palavras como rótulos que servem como alças para direcionar pincéis mentais complexos capazes de pintar imagens detalhadas em seu espaço de trabalho sensorial. Visualize uma "lâmpada triangular". O que você viu? (Palavras como alças de pincel mental)
36. Você usa uma palavra que possui significados diferentes em contextos diferentes, como se significasse a mesma coisa em todas as ocasiões, criando possivelmente a ilusão de algo mutável e adaptável. "Martin disse a Bob que o prédio ficava à sua esquerda." Mas "esquerda" é uma palavra funcional que é avaliada com uma variável dependente do locutor, extraída do contexto circundante. De quem é a "esquerda", de Bob ou de Martin? (Falácias de pergunta variável)
37. Você acredita que as definições não podem estar "erradas" ou que "posso definir uma palavra do jeito que quiser!" Esse tipo de atitude ensina você a defender com indignação suas ações passadas, em vez de prestar atenção às suas consequências ou admitir seus erros. (37 maneiras pelas quais o uso inadequado de categorias pode ter efeitos colaterais negativos em sua cognição).

Tudo o que você faz mentalmente tem um efeito, e seu cérebro age inconscientemente sem a sua supervisão.

Dizer "Palavras são arbitrárias; posso definir uma palavra do jeito que quiser" faz tanto sentido quanto dirigir em uma pista de gelo com o acelerador no máximo e dizer: "Olhando para este volante, não consigo ver por que um ângulo radial é especial — então posso virar o volante do jeito que eu quiser."

Se você está tentando chegar a algum lugar, ou apenas tentando sobreviver, é melhor começar a prestar atenção aos três ou seis dúzias de critérios de otimização que controlam como você usa palavras, definições, categorias, classes, limites, rótulos e conceitos.

Interlúdio: uma explicação intuitiva do teorema de Bayes

[**Nota do editor**: este é um resumo da versão original deste ensaio, que continha muitos elementos interativos].

Seus amigos e colegas não param de falar sobre algo chamado "Teorema de Bayes" ou "Regra de Bayes", ou sobre um tal de raciocínio bayesiano. Eles parecem extremamente empolgados com isso, então você resolve pesquisar no Google e encontra uma página sobre o Teorema de Bayes e...

De repente, você se depara com uma equação. Só isso. Apenas uma equação. A página que você encontrou fornece uma definição, mas não explica o que é, por que é útil ou por que seus amigos estão tão animados com ela. Parece ser apenas mais uma fórmula estatística qualquer.

Por que um conceito matemático desperta um entusiasmo tão estranho em seus colegas? O que é essa tal "Revolução Bayesiana" que está varrendo as ciências e que supostamente inclui até mesmo o próprio método experimental como um caso especial? Qual é o segredo que os adeptos de Bayes conhecem? Que luz eles viram?

Em breve, você saberá. Em breve, você será um de nós.

Embora existam algumas explicações do Teorema de Bayes disponíveis online, minha experiência em apresentar o raciocínio bayesiano às pessoas é que essas explicações costumam ser muito abstratas. Na verdade, o raciocínio bayesiano é bastante contraintuitivo. As pessoas não usam o raciocínio bayesiano intuitivamente, acham muito difícil aprender quando ensinadas e esquecem rapidamente os métodos bayesianos após o término do treinamento. Isso vale tanto para estudantes iniciantes quanto para profissionais altamente qualificados em uma área. O raciocínio bayesiano parece ser uma daquelas coisas que, assim como a mecânica quântica ou o Teste de Seleção de Wason, é inerentemente difícil para os seres humanos compreenderem com nossas capacidades mentais inatas.

Pelo menos é o que dizem. Neste texto, você encontrará uma tentativa de oferecer uma explicação intuitiva do raciocínio bayesiano — uma introdução extremamente suave que recorre a todas as formas humanas de compreender os números, desde as frequências naturais até a visualização espacial. O objetivo é transmitir não apenas regras abstratas para manipular números, mas também o significado desses números e o motivo por que as regras são como são (e não podem ser diferentes). Após ler isso, você verá problemas bayesianos até em seus sonhos.

Vamos começar.

Aqui está um problema contextualizado sobre uma situação que os médicos frequentemente enfrentam:

1% das mulheres com 40 anos que fazem exames de rotina têm câncer de mama. 80% das mulheres com câncer de mama terão resultados positivos em suas mamografias. 9,6% das mulheres sem câncer de mama também terão resultados positivos em suas mamografias. Se uma mulher nes-

ta faixa etária teve um resultado positivo em uma triagem de rotina, qual é a probabilidade de que ela realmente tenha câncer de mama?

Qual você acha que é a resposta? Se você nunca se deparou com esse tipo de problema antes, reserve um momento para tentar chegar à sua própria conclusão antes de continuar.

Agora, imagine que eu lhe dissesse que a maioria dos médicos responde incorretamente a este problema — geralmente, apenas cerca de 15% dos médicos acertam. ("Sério? 15%? Isso é um número real ou uma lenda urbana baseada em uma pesquisa na Internet?" É um número real. Veja: Casscells, Schoenberger e Graboys 1978; [1] Eddy 1982; [2] Gigerenzer e Hoffrage 1995; [3] e muitos outros estudos. É um resultado surpreendente e fácil de replicar, por isso foi amplamente confirmado.)

No problema apresentado anteriormente, a maioria dos médicos estima que a probabilidade esteja entre 70% e 80%, o que está muito longe da realidade.

Aqui está uma versão alternativa do problema na qual os médicos se saem um pouco melhor:

De cada 1.000 mulheres aos quarenta anos que participam de triagem de rotina, 10 têm câncer de mama. 800 em cada 1.000 mulheres com câncer de mama terão mamografias positivas. 96 em cada 1.000 mulheres sem câncer de mama também terão mamografias positivas. Se 1.000 mulheres nessa faixa etária forem submetidas a uma triagem de rotina, que fração das mulheres com mamografias positivas realmente terá câncer de mama?

E, finalmente, aqui está o problema no qual os médicos se saem melhor, com 46% — quase a metade — chegando à resposta correta:

De 10.000 mulheres aos 40 anos que participam de exames de rotina, 100 têm câncer de mama. 80 em cada 100 mulheres com câncer de mama terão uma mamografia positiva. 950 de 9.900 mulheres sem câncer de mama também terão uma mamografia positiva. Se 10.000 mulheres nessa faixa etária forem submetidas a um exame de rotina, que fração das mulheres com mamografias positivas realmente terá câncer de mama?

A resposta correta é 7,8%, calculada da seguinte maneira: a cada 10.000 mulheres, 100 têm câncer de mama e dessas 100, 80 apresentarão resultados positivos na mamografia. Das 10.000 mulheres, as outras 9.900 não têm câncer de mama e, dentre elas, 950 terão resultados positivos na mamografia.

Assim, o número total de mulheres com resultados positivos na mamografia é de 950 + 80, ou seja, 1.030. Desse total de 1.030 mulheres com resultados positivos na mamografia, 80 terão câncer de mama. Dessa forma, a proporção é de 80/1.030, resultando em 0,07767 ou 7,8%.

Dito de outra forma, antes da mamografia, as 10.000 mulheres podem ser divididas em dois grupos:

- Grupo 1: 100 mulheres com câncer de mama.
- Grupo 2: 9.900 mulheres sem câncer de mama.

A soma desses dois grupos totaliza 10.000 pacientes, comprovando que nenhuma mulher foi excluída dos cálculos. Após a realização da mamografia, as mulheres podem ser separadas em quatro grupos:

- Grupo A: 80 mulheres com câncer de mama e mamografia positiva.
- Grupo B: 20 mulheres com câncer de mama e mamografia negativa.
- Grupo C: 950 mulheres sem câncer de mama e mamografia positiva.
- Grupo D: 8.950 mulheres sem câncer de mama e mamografia negativa.

A soma dos grupos A e B, compostos por mulheres com câncer de mama, forma o grupo 1; e a soma dos grupos C e D, compostos por mulheres sem câncer de mama, forma o grupo 2. Se uma mamografia for realizada em 10.000 mulheres, das 1.030 mulheres com resultado positivo, oitenta delas terão câncer. Essa é a resposta correta que um médico deve fornecer a uma paciente com mamografia positiva se ela perguntar qual é a chance dela ter câncer de mama. Se treze pacientes fizerem essa pergunta, aproximadamente uma delas terá câncer.

O erro mais frequente é não considerar a proporção original de mulheres com câncer de mama e mulheres sem câncer de mama que recebem resultados falsos positivos. Em vez disso, muitos médicos nesses estudos parecem ter se concentrado apenas na proporção de mulheres com câncer de mama que obtêm resultados positivos. Por exemplo, eles podem pensar que, se cerca de 80% das mulheres com câncer de mama têm mamografias positivas, então a probabilidade de uma mulher com mamografia positiva ter câncer de mama deve ser de cerca de 80%.

Para se chegar à resposta final, é fundamental ter as três informações disponíveis:

- A porcentagem de mulheres com câncer de mama.
- A porcentagem de mulheres sem câncer de mama que recebem resultados positivos falsos.
- A porcentagem de mulheres com câncer de mama que recebem resultados positivos (corretos).

A proporção inicial de pacientes com câncer de mama é chamada de probabilidade a priori. As chances de uma paciente com câncer de mama obter uma mamografia positiva e as chances de uma paciente sem câncer de mama obter uma mamografia positiva são conhecidas como as duas probabilidades condicionais. Juntas, essas informações iniciais são chamadas de a priori. A resposta final — a probabilidade estimada de que uma paciente tenha câncer de mama, sabendo que ela tem um resultado positivo em sua mamografia — é chamada de probabilidade revisada ou a posteriori. O que acabamos de ver é que a probabilidade a posteriori depende, em parte, da probabilidade a priori.

Para entender que a resposta final sempre depende da proporção original de mulheres com câncer de mama, considere um universo alternativo em que apenas uma mulher em um milhão tem câncer de mama. Mesmo que a mamografia neste mundo detecte o câncer de mama em 8 de 10 casos e retorne um falso positivo em uma mulher sem câncer de mama em apenas 1 de 10 casos, ainda haverá cem mil falsos positivos para cada caso real de câncer detectado. A probabilidade original de uma mulher ter câncer é tão extremamente baixa que, embora um resultado positivo na mamografia aumente a probabilidade estimada, a probabilidade não aumenta com certeza ou mesmo com "uma chance perceptível"; a probabilidade vai de 1 em 1.000.000 para 1 em 100.000.

Isso demonstra que o resultado da mamografia não substitui as informações anteriores sobre a chance da paciente ter câncer; a mamografia apenas modifica a probabilidade estimada segundo o resultado. Um resultado positivo aumenta a probabilidade original; um resultado negativo a diminui. Por exemplo, no problema original em que 1% das mulheres têm câncer, 80% das mulheres com câncer têm mamografias positivas e 9,6% das mulheres sem câncer têm mamografias positivas, um resultado positivo na mamografia aumenta a probabilidade de 1% para cerca de 7,8%.

A maioria das pessoas que se deparam com problemas desse tipo pela primeira vez, muitas vezes cometem o erro de substituir a probabilidade original de 1% pela probabilidade de 80% de que uma mulher com câncer de mama receba um resultado positivo na mamografia. Essa pode parecer uma boa ideia, mas, na verdade, não é. A probabilidade de uma mulher com um resultado positivo

na mamografia ter câncer de mama não é a mesma coisa que a probabilidade de uma mulher com câncer de mama ter um resultado positivo na mamografia. Essas duas situações são tão diferentes quanto maçãs e laranjas.

P: Por que o pensador bayesiano cruzou a estrada?

R: Você precisa de mais informações para responder a esta pergunta.

Imagine um barril cheio de pequenos ovos de plástico. Alguns ovos são vermelhos e outros são azuis. 40% dos ovos contêm pérolas e 60% estão vazios. Dos ovos que contêm pérolas, 30% são azuis, e dos ovos vazios, 10% são azuis. Qual é a probabilidade de um ovo azul conter uma pérola? Para este exemplo, você pode fazer o cálculo mentalmente. Tente!

Uma forma mais concisa de apresentar o problema:

$$P(\text{pérola}) = 40\%$$

$$P(\text{azul}|\text{pérola}) = 30\%$$

$$P(\text{azul}|\neg\text{pérola}) = 10\%$$

$$P(\text{pérola}|\text{azul}) = ?$$

O símbolo "¬" significa "não", então ¬pérola significa "não pérola".

A expressão P(azul|pérola) representa a probabilidade condicional de um ovo ser azul, sabendo que ele contém uma pérola. A condição vem após a barra vertical, enquanto o evento de interesse vem antes. Por exemplo, P(azul|pérola) = 30% significa que, se sabemos que um ovo contém uma pérola, há 30% de chance de ele ser azul. Assim, o que buscamos — "a probabilidade de um ovo azul conter uma pérola" ou "a probabilidade de um ovo conter uma pérola, sabendo que é azul" — é representada por P(pérola|azul).

40% dos ovos contêm pérolas e 60% estão vazios. Dos ovos com pérolas, 30% são azuis, então 12% de todos os ovos contêm pérolas e são azuis. Dos ovos vazios, 10% são azuis, portanto, 6% de todos os ovos estão vazios e são azuis. No total, 18% dos ovos são azuis, dos quais 12% contêm pérolas, logo a chance de um ovo azul conter uma pérola é 12/18, ou 2/3, ou cerca de 67%.

Como vimos antes, a importância das três informações fica clara ao considerarmos casos extremos. Se tivermos um barril grande com apenas um ovo com pérola em cada mil ovos, saber que um ovo é azul aumenta a probabilidade de conter uma pérola de 0,1% para 0,3% (em vez de aumentar de 40% para 67%). Da mesma forma, se houver 999 pérolas em mil ovos, saber que um ovo é azul reduz a probabilidade de não haver pérola de 1/1.000 para cerca de 1/3.000, ou seja, a probabilidade aumenta de 99,9% para 99,966%.

No problema do ovo e da pérola, a maioria das pessoas não familiarizadas com o raciocínio bayesiano provavelmente responderia que a probabilidade de um ovo azul conter uma pérola é de 30%, ou talvez 20% (os 30% de chance de um verdadeiro positivo menos os 10% de chance de um falso positivo). Embora esse raciocínio pareça lógico, ele não responde à pergunta feita. É como perguntar a uma criança de 7 anos: "Se dezoito pessoas entrarem em um ônibus e depois mais sete entrarem, quantos anos tem o motorista do ônibus?" Muitas crianças responderão: "Vinte e cinco". Elas entendem que devem fazer um cálculo, mas ainda não conseguem relacioná-lo com a realidade. Similarmente, para encontrar a probabilidade de uma mulher com mamografia positiva ter câncer de mama, não faz sentido substituir a probabilidade original de que a mulher tenha câncer pela probabilidade de que uma mulher com câncer de mama tenha uma mamografia positiva. Tampouco é possível subtrair a probabilidade de um falso positivo da probabilidade do verdadeiro positivo. Essas operações são tão irrelevantes quanto somar o número de pessoas no ônibus para descobrir a idade do motorista.

Um estudo conduzido por Gigerenzer e Hoffrage em 1995 mostrou que algumas formas de apresentar problemas são mais eficazes para evocar o raciocínio bayesiano correto. [4] A abordagem menos eficaz utilizava probabilidades. Uma formulação um pouco mais eficaz utilizava frequências em vez de probabilidades. Neste caso, o problema permanecia o mesmo, mas ao invés de dizer que 1% das mulheres têm câncer de mama, dizia-se que 1 em cada 100 mulheres têm câncer de mama. Também se mencionava que 80 em

cada 100 mulheres com câncer de mama teriam uma mamografia positiva, e assim por diante. Por que uma proporção maior de participantes usou o raciocínio bayesiano corretamente nesta versão do problema? Provavelmente porque a expressão "1 em cada 100 mulheres" estimula a visualização concreta de um grupo de X mulheres com câncer, levando a imaginar X mulheres com câncer e com uma mamografia positiva, e assim por diante.

Até agora, a forma mais eficaz de apresentação é conhecida como "frequências naturais", que consiste em dizer que 40 em 100 ovos contêm pérolas, 12 em 40 ovos que contêm pérolas são azuis e 6 em 60 ovos vazios são azuis. A apresentação de frequências naturais inclui as informações sobre a probabilidade prévia na apresentação das probabilidades condicionais. Se você estivesse aprendendo apenas sobre as probabilidades condicionais dos ovos por meio de experimentação natural, ao abrir cem ovos, você encontraria cerca de 40 ovos contendo pérolas, dos quais 12 ovos seriam azuis, enquanto abriria 60 ovos vazios, dos quais cerca de 6 seriam azuis. Ao aprender sobre as probabilidades condicionais, você veria exemplos de ovos azuis contendo pérolas duas vezes mais frequentemente do que exemplos de ovos azuis vazios.

Infelizmente, apesar de as frequências naturais representarem um avanço na direção correta, elas provavelmente não serão suficientes. Quando os problemas são apresentados em frequências naturais, a proporção de pessoas que utiliza o raciocínio bayesiano aumenta para cerca de metade. Embora essa melhoria seja considerável, não é grande o suficiente quando se trata de médicos e pacientes reais.

P: Como posso encontrar as probabilidades prévias de um problema?

R: Muitos princípios comumente usados estão listados no "Handbook of Chemistry and Physics".

P: De onde as probabilidades prévias vêm originalmente?

R: Nunca faça essa pergunta.

P: Tudo bem. Então, de onde os cientistas obtêm suas probabilidades prévias?

R: As probabilidades prévias para problemas científicos são estabelecidas por votação anual da AAAS (Associação Americana para o Avanço da Ciência). Nos últimos anos, a votação tornou-se conflituosa e controversa, com muita hostilidade, polarização de facções e vários assassinatos diretos. Isso pode ser uma fachada para lutas internas no Conselho de Bayes, ou talvez os participantes tenham tempo livre demais. Ninguém tem certeza do que realmente está acontecendo.

P: Entendo. E onde as outras pessoas adquirem suas probabilidades prévias?

R: Elas as baixam da Internet.

P: E se as probabilidades prévias que quero não estiverem disponíveis na Internet?

R: Em São Francisco, nos Estados Unidos, na região da Chinatown, existe uma loja de antiguidades pequena e desorganizada. É melhor não perguntar sobre o rato de bronze, que está exposto lá.

Na verdade, as probabilidades prévias são consideradas verdadeiras ou falsas, assim como a resposta final — elas refletem a realidade e podem ser avaliadas através da comparação com ela. Por exemplo, se você acredita que, em uma amostra de 10.000 mulheres, 920 têm câncer de mama e o número real é de 100 em 10.000, então suas estimativas prévias estão incorretas. Para resolver nosso problema específico, as probabilidades prévias foram baseadas em três estudos. O primeiro estudo analisou os históricos de casos de mulheres com câncer de mama para determinar quantas delas tiveram uma mamografia positiva.

O segundo estudo examinou mulheres sem câncer de mama para verificar quantas tiveram uma mamografia positiva. Por fim, o terceiro estudo foi um estudo epidemiológico que avaliou a prevalência de câncer de mama em grupos demográficos específicos.

A probabilidade P(A,B) é a mesma que P(B,A), mas P(A|B) não é a mesma coisa que P(B|A), e P(A,B) é completamente diferente de P(A|B). É uma confusão comum confundir algumas ou todas essas quantidades.

Para nos familiarizarmos com todas as relações entre elas, vamos brincar de "seguir os graus de liberdade". Por exemplo, as duas quantidades P(câncer) e P(¬câncer) têm um grau de liberdade entre si, devido à

lei geral P(A) + P(¬A) = 1. Se você souber que P(¬câncer) = 0,99, poderá obter P(câncer) = 1 - P(¬câncer) = 0,01.

As quantidades P(positivo|câncer) e P(¬positivo|câncer) também têm apenas um grau de liberdade entre elas; ou uma mulher com câncer de mama recebe uma mamografia positiva, ou não. Por outro lado, P(positivo|câncer) e P(positivo|¬câncer) têm dois graus de liberdade. É possível ter um teste de mamografia que dê positivo para 80% dos pacientes com câncer e 9,6% dos pacientes saudáveis, ou que dê positivo para 70% dos pacientes com câncer e 2% dos pacientes saudáveis, ou até mesmo um teste de saúde que dê "positivo" para 30% dos pacientes com câncer e 92% dos pacientes saudáveis. As duas quantidades, o resultado do teste de mamografia para pacientes com câncer e o resultado do teste de mamografia para pacientes saudáveis, são, em termos matemáticos, independentes; uma não pode ser obtida da outra de forma alguma e, portanto, elas têm dois graus de liberdade entre si.

E quanto a P(positivo, câncer), P(positivo|câncer) e P(câncer)? Aqui temos três quantidades; quantos graus de liberdade existem? Nesse caso, a equação que deve ser mantida é:

$$P(\text{positivo, câncer}) = P(\text{positivo}|\text{câncer}) \times P(\text{câncer}).$$

Essa igualdade reduz os graus de liberdade em um. Se soubermos a fração de pacientes com câncer e a chance de uma paciente com câncer ter uma mamografia positiva, podemos deduzir a fração de pacientes que têm câncer de mama e uma mamografia positiva por meio da multiplicação.

Da mesma forma, se soubermos o número de pacientes com câncer de mama e mamografias positivas, e também o número de pacientes com câncer de mama, podemos estimar a chance de uma mulher com câncer de mama obter uma mamografia positiva dividindo: P(positivo|câncer) = P(positivo, câncer)/P(câncer). Na verdade, é exatamente assim que esses testes de diagnóstico médico são calibrados; você faz um estudo com 8.520 mulheres com câncer de mama e vê que há 6.816 (ou algo em torno disso) mulheres com câncer de mama e mamografias positivas e, em seguida, divide 6.816 por 8.520 para descobrir que 80% das mulheres com câncer de mama tiveram mamografias positivas. (A propósito, se você acidentalmente dividir 8.520 por 6.816 em vez do contrário, seus cálculos começarão a fazer coisas estranhas, como insistir que 125% das mulheres com câncer de mama e mamografias positivas têm câncer de mama. Esse é um erro comum na execução da aritmética bayesiana, de acordo com minha experiência). E, finalmente, se você souber P(positivo, câncer) e P(positivo|câncer), poderá deduzir quantos pacientes com câncer devem ter existido originalmente. Há dois graus de liberdade compartilhados entre as três quantidades; se conhecermos quaisquer duas, podemos deduzir a terceira.

E quanto a P(positivo), P(positivo, câncer) e P(positivo, ¬câncer)? Novamente, há apenas dois graus de liberdade entre essas três variáveis. A equação que ocupa o grau de liberdade extra é:

$$P(\text{positivo}) = P(\text{positivo, câncer}) + P(\text{positivo, }\neg\text{câncer}).$$

Para começar, é assim que P(positivo) é calculado: calculamos o número de mulheres com câncer de mama que têm mamografias positivas e o número de mulheres sem câncer de mama que têm mamografias positivas e, em seguida, somamos esses valores para obter o número total de mulheres com mamografias positivas. Seria muito estranho sair e realizar um estudo para determinar o número de mulheres com mamografias positivas — apenas esse número e nada mais — mas, em teoria, você poderia fazer isso. E se você realizasse outro estudo e descobrisse o número de mulheres com mamografias positivas e câncer de mama, você também saberia o número de mulheres com mamografias positivas e sem câncer de mama — ou uma mulher com mamografia positiva tem câncer de mama, ou não tem. Em geral, P(A,B) + P(A,¬B) = P(A). Simetricamente, P(A,B) + P(¬A,B) = P(B).

E quanto a P(positivo, câncer), P(positivo, ¬câncer), P(¬positivo, câncer) e P(¬positivo, ¬câncer)? À primeira vista, poderíamos ser tentados a pensar que essas quatro quantidades possuem apenas dois graus de liberdade — que seria possível, por exemplo, obter P(positivo, ¬câncer) multiplicando P(positivo) × P(¬câncer), e assim, todas as quatro quantidades poderiam ser encontradas a partir das duas quantidades P(positivo) e P(câncer). Porém, isso não é verdade! P(positivo, ¬câncer) = P(positivo) × P(¬câncer) apenas se essas duas probabilidades forem estatisticamente independentes — isto é, se a chance de uma mulher ter câncer de mama não influenciar se ela tem uma mamografia positiva. Isso significa que as duas probabilidades condicionais precisam ser iguais entre si, o que eliminará um grau de liberdade. Ao observar que essas quatro quantidades são os grupos A, B, C e D, percebe-se que, em teoria, qualquer número de pessoas poderia estar

nesses grupos. Por exemplo, é possível começar com um grupo de 80 mulheres com câncer de mama e mamografias positivas, acrescentar um grupo de 500 mulheres com câncer de mama e mamografias negativas e, em seguida, um grupo de 3 mulheres sem câncer de mama e mamografias negativas, e assim por diante. Agora, parece que as quatro quantidades têm quatro graus de liberdade. Porém, quando expressamos essas quantidades como probabilidades, precisamos normalizá-las para frações do grupo completo, adicionando a restrição de que P(positivo, câncer) + P(positivo, ¬câncer) + P(¬positivo, câncer) + P(¬positivo, ¬câncer) = 1. Esta equação ocupa um grau de liberdade, deixando apenas três graus de liberdade entre as quatro quantidades. Se as frações de mulheres nos grupos A, B e D forem especificadas, será possível deduzir a fração de mulheres no grupo C.

Dados os quatro grupos A, B, C e D, é muito simples calcular o resto:

P(câncer) = (A + B) / (A + B + C + D)

P(¬positivo|câncer) = B / (A + B)

e assim por diante. Como {A, B, C, D} contém três graus de liberdade, segue que todo o conjunto de probabilidades relacionando as taxas de câncer aos resultados dos testes contém somente três graus de liberdade. É importante lembrar que, em problemas bayesianos, sempre precisamos de três informações — a probabilidade prévia e as duas probabilidades condicionais — que, de fato, possuem três graus de liberdade entre elas. Na realidade, para problemas bayesianos, basta ter quaisquer três quantidades com três graus de liberdade entre elas para ser possível especificar logicamente todo o problema.

A razão de verossimilhança de um teste é a probabilidade de um resultado verdadeiro positivo dividida pela probabilidade de um falso positivo. Ao se tratar de um resultado positivo, a razão de verossimilhança resume o quanto essa descoberta diminuirá a probabilidade anterior. Será que a razão de verossimilhança de um exame médico é tudo o que precisamos saber sobre sua utilidade?

Não, isso não é verdade! A razão de verossimilhança fornece informações importantes sobre o significado de um resultado positivo no exame médico, mas não especifica o significado de um resultado negativo e nem a frequência com que o exame é útil. Por exemplo, uma mamografia com uma taxa de acerto de 80% para pacientes com câncer de mama e uma taxa de falso positivo de 9,6% para pacientes saudáveis tem a mesma razão de verossimilhança que um teste com uma taxa de acerto de 8% e uma taxa de falso positivo de 0,96%. Embora esses dois testes tenham a mesma razão de verossimilhança, o primeiro teste é mais útil em todos os sentidos — ele detecta doenças com mais frequência e um resultado negativo é uma evidência mais forte de saúde.

Vamos supor que você realize dois testes consecutivos para detectar câncer de mama — digamos, uma mamografia padrão e um outro teste independente da mamografia. Como não conheço nenhum teste independente da mamografia, vou inventar um para esse problema e chamá-lo de Teste da Divisão de Tams-Braylor, que verifica se alguma célula está se dividindo mais rapidamente do que outras células. Suponhamos que o Tams-Braylor apresente uma taxa de verdadeiro positivo de 90% para pacientes com câncer de mama e uma taxa de falso positivo de 5% para pacientes sem câncer. Digamos que a prevalência anterior de câncer de mama seja de 1%. Se uma paciente obtiver resultado positivo em sua mamografia e em seu Tams-Braylor, qual será a probabilidade revisada de que ela tenha câncer de mama?

Uma forma de solucionar esse problema seria utilizar a probabilidade revisada de uma mamografia positiva, que já calculamos como sendo 7,8%, e inseri-la no teste de Tams-Braylor como a nova probabilidade anterior. Se fizermos isso, descobriremos que o resultado é de 60%.

Imagine que a prevalência inicial de câncer de mama em um grupo demográfico é de 1%. Agora, suponha que, como médicos, temos três testes independentes para detectar câncer de mama.

O primeiro teste, chamado de teste A, é uma mamografia e tem uma razão de verossimilhança de 80% / 9,6% = 8,33.

O segundo teste, o teste B, tem uma razão de verossimilhança de 18,0 (por exemplo, 90% versus 5%).

O terceiro teste, o teste C, tem uma razão de verossimilhança de 3,5 (que pode ser de 70% versus 20% ou de 35% versus 10%, não importa).

Suponha agora que um paciente receba um resultado positivo nos três testes. Qual é a probabilidade de que esse paciente tenha câncer de mama?

Aqui está um truque divertido para simplificar as contas. Se a prevalência anterior de câncer de mama em um grupo demográfico for de 1%, então 1 em cada 100 mulheres tem câncer de mama e 99 em cada 100 mulheres não têm câncer de mama. Então, se reescrevermos a probabilidade de 1% como uma razão de chances, as chances são de 1:99.

E as razões de verossimilhança dos três testes A, B e C são:

8,33: 1 = 25: 3

18,0: 1 = 18: 1

3,5: 1 = 7: 2.

As chances para mulheres com câncer de mama que obtiverem resultado positivo nos três testes, contra mulheres sem câncer de mama que obtiverem resultado positivo nos três testes, serão iguais a:

1 × 25 × 18 × 7: 99 × 3 × 1 × 2 = 3150: 594.

Para recuperar a probabilidade a partir das chances, basta escrever:

3150/ (3150 + 594) = 84%.

Essa técnica sempre funciona, independentemente de como as razões de verossimilhança são escritas. Por exemplo, 8,33:1 é exatamente o mesmo que 25:3 ou 75:9. Também não importa em que ordem os testes são realizados ou em que ordem os resultados são computados. Deixo a demonstração como um exercício para o leitor.

E. T. Jaynes, em Probability Theory With Applications in Science and Engineering (Teoria das Probabilidades com Aplicações em Ciência e Engenharia), sugere que a credibilidade e as evidências devem ser medidas em decibéis. [5]

Decibéis?

Os decibéis são utilizados para medir diferenças exponenciais de intensidade sonora. Por exemplo, se o som da buzina de um carro tem uma intensidade 10.000 vezes maior (por metro quadrado por segundo) do que o som de um despertador, então, a buzina do carro seria 40 decibéis mais alta.

Já o som de um pássaro cantando pode ter uma intensidade 1.000 vezes menor do que um despertador, sendo assim, seria 30 decibéis mais suave. Para calcular o número de decibéis, é necessário aplicar o logaritmo de base 10 e multiplicar por 10:

$$decibéis = 10\ log_{10}(intensidade)$$

ou

$$intensidade = 10^{\frac{decibéis}{10}}$$

Digamos que iniciemos com uma probabilidade a priori de 1% de uma mulher ter câncer de mama, o que corresponde a uma razão de chances de 1:99. Em seguida, aplicamos três testes com razões de verossimilhança de 25:3, 18:1 e 7:2. Podemos multiplicar esses números ou simplesmente adicionar seus logaritmos:

$$10\log_{10}(1/99) \approx -20$$
$$10\log_{10}(25/3) \approx 9$$
$$10\log_{10}(18/1) \approx 13$$
$$10\log_{10}(7/2) \approx 5\,.$$

Começa com a baixa probabilidade de uma mulher ter câncer de mama — a credibilidade é de -20 decibéis. Em seguida, são obtidos três resultados de testes com evidências correspondentes a 9, 13 e 5 decibéis, aumentando a credibilidade em um total de 27 decibéis. Assim, a credibilidade anterior de −20 decibéis se torna uma credibilidade posterior de 7 decibéis. Isso significa que as chances de uma mulher ter câncer de mama passam de 1 em 99 para 5 em 1, e a probabilidade aumenta de 1% para cerca de 83%.

Agora imagine que você é um mecânico de aparelhos. Quando um aparelho para de funcionar, é porque a mangueira está bloqueada em 30% das vezes. Se a mangueira estiver bloqueada, há 45% de chance de cutucar a engenhoca vai produzir faíscas. Se a mangueira estiver desbloqueada, há apenas 5% de chance de produzir faíscas. Um cliente traz para você um aparelho com defeito e, após cutucá-lo, você descobre que ele produz faíscas. Qual é a probabilidade de que a mangueira esteja bloqueada?

Qual é a sequência de operações aritméticas que você executou para resolver esse problema?

(45% × 30%)/ (45% × 30% + 5% × 70%) ou

Assim como anteriormente, para determinar as chances de uma mulher com uma mamografia positiva ter câncer de mama, realizamos o seguinte cálculo:

$$\frac{P(\text{positive}|\text{cancer}) \times P(\text{cancer})}{\left(\begin{array}{l} P(\text{positive}|\text{cancer}) \times P(\text{cancer}) \\ + \quad P(\text{positive}|\text{-cancer}) \times P(\text{-cancer}) \end{array}\right)} =$$

$$\frac{P(\text{positive,cancer})}{P(\text{positive,cancer}) + P(\text{positive -cancer})} =$$

$$\frac{P(\text{positive,cancer})}{P(\text{positive})}$$

$$P(\text{positive}|\text{cancer})$$

A forma totalmente geral desse cálculo é conhecida como Teorema de Bayes ou Regra de Bayes.

Teorema de Bayes

$$P(A|X) = \frac{P(X|A) \times P(A)}{P(X|A) \times P(A) + P(X|\neg A) \times P(\neg A)}$$

Quando há um fenômeno A que se deseja investigar e uma observação X que é uma evidência sobre A — por exemplo, no exemplo anterior, A é câncer de mama e X é uma mamografia positiva — o Teorema de Bayes nos diz como devemos atualizar nossa probabilidade de A, considerando a nova evidência X.

A essa altura, o Teorema de Bayes pode parecer flagrantemente óbvio ou mesmo tautológico, em vez de empolgante e novo. Nesse caso, esta introdução foi totalmente bem-sucedida em seu propósito. O Teorema de Bayes descreve o que torna algo uma "evidência" e qual é a quantidade de evidência.

Os modelos estatísticos são avaliados em comparação com o método bayesiano, pois na estatística, o método bayesiano é o melhor possível — ele define a quantidade máxima de informação que se pode obter de uma evidência específica, da mesma forma que a termodinâmica define a quantidade máxima de trabalho que se pode obter de uma diferença de temperatura.

É por isso que você ouve cientistas cognitivos falando sobre raciocinadores bayesianos. Na ciência cognitiva, o raciocínio bayesiano é o termo tecnicamente preciso que usamos para indicar a mente racional.

Há também uma série de heurísticas gerais sobre o raciocínio humano que você pode aprender observando o Teorema de Bayes.

Por exemplo, em muitas discussões sobre o Teorema de Bayes, é comum ouvir psicólogos cognitivos afirmarem que as pessoas não consideram suficientemente as frequências anteriores.

Quando um problema surge e há evidências (X) que sugerem que a condição A pode ser verdadeira, as pessoas têm a tendência de avaliar a probabilidade de A exclusivamente com base na correspondência entre a evidência X e A, ignorando a frequência anterior de A.

No caso da mamografia, se você acredita que a chance de uma mulher ter câncer de mama está entre 70% e 80%, por exemplo, esse tipo de raciocínio é insensível à frequência anterior fornecida no problema e não percebe se 1% ou 10% das mulheres já começam com câncer de mama. "Preste mais atenção na frequência anterior!" é uma das muitas coisas que os seres humanos precisam ter em mente para compensar parcialmente nossas limitações internas.

Ao avaliar quanta evidência X favorece A, é comum dar muita importância à P(X|A) e não o suficiente à P(X|¬A). O grau em que um resultado X é evidência para A depende não apenas da probabilidade de observarmos X se A for verdadeiro, mas também da probabilidade de não observarmos X se A não for verdadeiro.

Por exemplo, se está chovendo, é provável que a grama esteja molhada — P(grama molhada|chuva) ≈ 1 - mas ver que a grama está molhada não significa necessariamente que acabou de chover; talvez o aspersor tenha sido ligado ou você esteja olhando para o orvalho da manhã. Já que P(grama molhada|¬chuva) é substancialmente maior que zero, P(chuva|grama molhada) é substancialmente menor que um. Por outro lado, se a grama nunca ficou molhada quando não chovia, então saber que a grama está molhada sempre indicaria que está chovendo — P(chuva|grama molhada) ≈ 1, mesmo se P(grama molhada|chuva) = 50%; ou seja, mesmo que a grama tenha ficado molhada apenas 50% das vezes em que choveu.

A evidência é sempre o resultado da diferença entre as duas probabilidades condicionais. Evidência forte não é o produto de uma probabilidade muito alta de que A leva a X, mas o produto de uma probabilidade muito baixa de que não-A poderia ter levado a X.

O avanço da revolução bayesiana nas ciências é impulsionado por uma combinação de fatores. Primeiramente, há um crescente reconhecimento entre os cientistas cognitivos de que os fenômenos mentais exibem uma estrutura bayesiana. Além disso, cientistas de diversas áreas estão adotando o método bayesiano para avaliar seus métodos estatísticos, reconhecendo sua superioridade. Por fim, a ideia de que a ciência em si é um caso particular do Teorema de Bayes e que a evidência experimental é essencialmente bayesiana também contribui para o avanço dessa revolução.

Os revolucionários bayesianos defendem que, quando realizamos um experimento e obtemos evidências que “confirmam” ou “desconfirmam” nossa teoria, essas confirmações e desconfirmacões são governadas pelas regras bayesianas. Por exemplo, não devemos apenas considerar se nossa teoria prevê o fenômeno, mas também se outras explicações possíveis também preveem o fenômeno.

No passado, a filosofia mais popular da ciência era provavelmente o falsificacionismo de Karl Popper, que agora está sendo superado pela revolução bayesiana. A ideia de Popper de que as teorias podem ser falsificadas definitivamente, mas nunca confirmadas definitivamente, é apenas um caso especial das regras bayesianas. Se $P(X|A) \approx 1$, o que significa que a teoria faz uma previsão clara, então a observação de $\neg X$ falsifica fortemente A., por outro lado, se $P(X|A) \approx 1$ e observamos X, isso não confirma definitivamente a teoria.

Para confirmar definitivamente a hipótese A observando X, não basta saber que $P(X|A) \approx 1$. Também é necessário saber que $P(X|\neg A) \approx 0$. No entanto, isso não é possível porque não podemos considerar todas as possíveis explicações alternativas para a ocorrência de X.

Portanto, pode existir outra condição B tal que $P(X|B) \approx 1$. Nesse caso, observar X não favorece A em relação à B.

Por exemplo, quando a teoria da Relatividade Geral de Einstein substituiu a teoria da gravidade incrivelmente bem confirmada de Newton, descobriu-se que todas as previsões de Newton eram apenas casos especiais das previsões de Einstein. Você até pode formalizar a filosofia de Popper matematicamente.

A razão de verossimilhança para X é a quantidade $P(X|A)/P(X|\neg A)$, que determina o quanto a observação de X desvia a probabilidade de A.

É essa razão de verossimilhança que nos diz quão forte é a evidência de X. Na sua teoria A, você pode prever X com uma probabilidade de 1, se quiser.

No entanto, você não pode controlar o denominador da razão de verossimilhança, $P(X|\neg A)$. Sempre haverá algumas teorias alternativas que também preveem X, e enquanto seguimos com a teoria mais simples que se ajusta à evidência atual, pode acontecer de encontrarmos evidências de que uma teoria alternativa prevê X, mas a nossa não. Foi essa pegadinha oculta que derrubou a teoria da gravidade de Newton. Portanto, há um limite para o quanto podemos avançar com previsões bem-sucedidas. Há um limite para o quão alta é a taxa de probabilidade para evidências confirmatórias.

Por outro lado, se você encontrar alguma evidência Y que definitivamente não foi prevista pela sua teoria, isso é uma evidência extremamente forte contra a sua teoria. Se $P(Y|A)$ for infinitesimal, então a razão de verossimilhança também será infinitesimal. Por exemplo, se $P(Y|A)$ for 0,0001% e $P(Y|\neg A)$ for 1%, então a razão de verossimilhança $P(Y|A)/P(Y|\neg A)$ será 1:10.000. São −40 decibéis de evidência!

Ou, invertendo a razão de verossimilhança, se $P(Y|A)$ for muito pequeno, então $P(Y|\neg A)/P(Y|A)$ será muito grande, o que significa que observar Y favorece muito $\neg A$ em vez de A. Falsificação é muito mais forte do que a confirmação.

Isso é uma consequência do ponto anterior de que evidências muito fortes não são o produto de uma probabilidade muito alta de que A leva a X, mas o produto de uma probabilidade muito baixa de que $\neg A$ poderia ter levado a X.

Esta é a regra precisa do bayesianismo que fundamenta o valor heurístico do falsificacionismo de Popper.

A afirmação de Popper de que uma ideia deve ser falsificável pode ser vista como uma aplicação da regra bayesiana de conservação da probabilidade. Se um resultado X é uma evidência positiva para uma teoria, o resultado $\neg X$ teria refutado a teoria até certo ponto. Tentar interpretar X e $\neg X$ como confirmação da teoria é impossível, segundo as regras bayesianas! Para aumentar a probabilidade de uma teoria, ela deve ser submetida a testes que possam diminuir sua probabilidade. Essa regra não só serve para detectar possíveis fraudes no processo social da ciência, mas é também uma consequência da teoria da probabilidade bayesiana.

Por outro lado, a ideia de Popper de que só existe falsificação e nenhuma confirmação é incorreta. O

Teorema de Bayes mostra que a falsificação é uma evidência muito mais forte do que a confirmação, mas a falsificação ainda é de natureza probabilística. Ela não é governada por regras fundamentalmente diferentes das que governam a confirmação, como argumentava Popper.

Portanto, verificamos que muitos fenômenos nas ciências cognitivas, além dos métodos estatísticos utilizados pelos cientistas e do próprio método científico, estão se revelando casos especiais do Teorema de Bayes. Isso é o que impulsiona a revolução Bayesiana.

Com a introdução explícita do Teorema de Bayes, podemos agora discutir de forma explícita seus componentes

$$P(A|X) = \frac{P(X|A) \times P(A)}{P(X|A) \times P(A) + P(X|\neg A) \times P(\neg A)}$$

.

Comecemos com P(A|X). Caso você esteja confuso sobre qual é A e qual é X no Teorema de Bayes, inicie com P(A|X) no lado esquerdo da equação; essa é a parte mais simples de interpretar. Em P(A|X), A é o que queremos descobrir. X é o método pelo qual estamos observando; X é a evidência que utilizamos para tirar conclusões sobre A.

Lembre-se de que para cada expressão P(Q|P), queremos saber sobre a probabilidade de Q dado P, o grau em que P implica Q. Uma notação mais coerente, que agora já é tarde demais para ser adotada, seria P(Q ← P).

P (Q|P) e P (Q, P) estão intimamente relacionados, mas não são iguais. P (Q, P) é a proporção de coisas que têm as propriedades Q e P entre todas as coisas, expressa como uma probabilidade ou fração. Por exemplo, a proporção de mulheres que possuem câncer de mama e mamografia positiva em relação ao grupo total de mulheres. Se houver 10.000 mulheres e 80 possuírem câncer de mama e mamografia positiva, então P (Q, P) é 80/10.000 = 0,8%. Podemos entender que a quantidade absoluta, 80, é normalizada em uma probabilidade relativa ao grupo de todas as mulheres.

Para deixar mais claro, suponhamos que exista um grupo de 641 mulheres com câncer de mama e mamografia positiva em um grupo total de 89.031 mulheres. Nesse caso, 641 é a quantidade absoluta. Se você escolher aleatoriamente uma mulher do grupo total, a probabilidade de escolher uma mulher com câncer de mama e mamografia positiva é P (Q, P), a qual é de 0,72% neste exemplo.

Por outro lado, P (Q|P) é a proporção de coisas que possuem as propriedades Q e P entre todas as coisas que têm P. Por exemplo, a proporção de mulheres com câncer de mama e mamografia positiva no grupo de todas as mulheres com mamografias positivas.

Se houver 641 mulheres com câncer de mama e mamografia positiva, 7.915 mulheres com mamografia positiva e 89.031 mulheres no total, então P (Q, P) é a probabilidade de obter uma dessas 641 mulheres se você escolher aleatoriamente do grupo total de 89.031, enquanto P (Q|P) é a probabilidade de obter uma dessas 641 mulheres se você escolher aleatoriamente no grupo menor de 7.915 mulheres com mamografia positiva.

Em certo sentido, P (Q|P) significa, na verdade, P (Q, P |P), mas repetir o P extra seria redundante. Já sabemos que as coisas têm a propriedade P, então estamos investigando a propriedade Q, mesmo que estejamos olhando para o tamanho do grupo (Q, P) no grupo P, e não para o tamanho do grupo Q no grupo P (o que seria absurdo). Essa é a ideia por trás de tomar a propriedade do lado direito como dada; significa que você sabe que está trabalhando apenas no grupo de coisas que têm a propriedade P.

Quando você restringe o foco para ver apenas esse grupo menor, muitas outras probabilidades mudam. Se você considerar P como um dado, então P (Q, P) é igual a apenas P (Q) — pelo menos em relação ao grupo P. A antiga frequência P (Q), ou seja, a frequência de “coisas que têm a propriedade Q dentro de toda a amostra”, é revisada para a nova frequência de “coisas que têm a propriedade Q dentro da subamostra de coisas que têm a propriedade P“. Se P é um dado, se P é o nosso mundo inteiro, procurar (Q, P) é o mesmo que procurar apenas Q.

Se o seu foco de atenção se limita apenas à população de ovos pintados de azul, a "probabilidade de um ovo conter uma pérola" de repente se torna um número diferente. Essa proporção é diferente para a população de ovos azuis do que para a população de todos os ovos. O "dado", a propriedade que restringe nosso foco de atenção, está sempre no lado direito de P (Q|P). P torna-se o nosso mundo, a totalidade do que vemos, e do outro lado do "dado", P tem sempre a probabilidade de 1 - isto é o que significa tomar P como dado.

Então, P (Q|P) significa "Se P tem probabilidade 1, qual é a probabilidade de Q?" ou "Se restringirmos nossa atenção apenas às coisas ou eventos em que P é verdadeiro, qual é a probabilidade de Q?" A afirmação Q, do outro lado do dado, não é certa — sua probabilidade pode ser 10%, 90% ou qualquer outro número.

No Teorema de Bayes, ao expressar a parte esquerda da equação como P(A|X), atualizamos a probabilidade de A após observar X. Essa nova probabilidade representa o grau em que X implica A, considerando nossa compreensão de X. X sempre representa a observação ou evidência, enquanto A é a propriedade em questão, o que desejamos saber.

O lado direito do Teorema de Bayes é derivado do lado esquerdo através destes passos:

Após a resolução, todas as implicações do lado direito da equação são da forma P(X|A) ou P(X|¬A), enquanto a implicação do lado esquerdo é P(A|X). Essa simetria surge porque as relações causais elementares são geralmente implicações de fatos para observações, por exemplo, do câncer de mama para uma mamografia positiva. As etapas elementares do raciocínio são geralmente implicações de observações para fatos, por exemplo, de uma mamografia positiva para o câncer de mama. O lado esquerdo do Teorema de Bayes é um passo inferencial elementar a partir da observação de uma mamografia positiva até a conclusão de uma probabilidade aumentada de câncer de mama.

$$P(A|X) = P(A|X)$$

$$P(A|X) = \frac{P(X,A)}{P(X)}$$

$$P(A|X) = \frac{P(X,A)}{P(X,A) + P(X,\neg A)}$$

$$P(A|X) = \frac{P(X|A) \times P(A)}{P(X|A) \times P(A) + P(X|\neg A) \times P(\neg A)}.$$

No Teorema de Bayes, a implicação é escrita da direita para a esquerda. Assim, P(câncer | positivo) é escrito no lado esquerdo da equação. O lado direito do Teorema descreve as etapas causais elementares, por exemplo, do câncer de mama para uma mamografia positiva. Portanto, as implicações no lado direito assumem a forma P(positivo | câncer) ou P(positivo | ¬câncer).

E este é o Teorema de Bayes. De um lado, inferência racional; do outro, causalidade física. Uma equação que conecta o mundo das ideias a realidade objetiva. Você se lembra de como o método científico acabou se tornando um caso especial do Teorema de Bayes? Poeticamente falando, podemos dizer que o Teorema de Bayes une o pensamento à realidade física do universo.

Ok, terminamos.

O Reverendo Bayes disse:

Você agora é um iniciado da Conspiração Bayesiana.

Referências

[1] Ward Casscells, Arno Schoenberger, and Thomas Graboys, “Interpretation by Physicians of Clinical Laboratory Results,“ New England Journal of Medicine 299 (1978): 999–1001.

[2] David M. Eddy, “Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities,“ in Judgement Under Uncertainty: Heuristics and Biases, ed. Daniel Kahneman, Paul Slovic, and Amos Tversky (Cambridge University Press, 1982).

[3] Gerd Gigerenzer and Ulrich Hoffrage, “How to Improve Bayesian Reasoning without Instruction: Frequency Formats,“ Psychological Review 102 (1995): 684–704.

[4] Ibid.

[5] Edwin T. Jaynes, “Probability Theory, with Applications in Science and Engineering,“ Unpublished manuscript (1974).