LIVRO 6

TORNANDO-SE MAIS FORTE

RACIONALIDADE

De A a Z

ELIEZER YUDKOWSKY

RACIONALIDADE DE A a Z

TORNANDO-SE MAIS FORTE

LIVRO 6

por **ELIEZER YUDKOWSKY**

Tradução de Mariana Hungria

Revisão de Enéas Canavezzi Versehgi

Brasil, 2024

Sumário

Inícios: uma introdução

por Rob Bensinger

Este livro, o último de "Racionalidade: da IA aos Zumbis", é menos uma conclusão do que um apelo à ação. Ao manter a função de "Tornando-se mais forte" como ponto de partida para futuras investigações, encerro citando recursos que o leitor pode explorar para ir além destas sequências e alcançar uma compreensão mais abrangente do Bayesianismo.

A definição de racionalidade normativa neste texto, em termos da teoria bayesiana da probabilidade e da teoria da decisão, é padrão na ciência cognitiva. Para uma introdução à abordagem heurística e de preconceitos, consulte *Thinking and Deciding* (Pensando e Decidindo) de Baron [1]. Para uma visão geral do campo, consulte o *Oxford Handbook of Thinking and Reasoning* (Manual de Pensamento e Raciocínio de Oxford) [2].

Os argumentos apresentados nestas páginas sobre a filosofia da racionalidade são mais controversos. Yudkowsky argumenta, por exemplo, que um agente racional deveria adotar a caixa no Problema de Newcomb — uma posição minoritária entre os teóricos da decisão. [3] (Consulte Holt para obter uma descrição não técnica do Problema de Newcomb. [4]) O livro *Good and Real* (Bom e Real), de Gary Drescher, chega independentemente a muitas das mesmas conclusões de Yudkowsky sobre filosofia da ciência e teoria da decisão [5]. Sendo assim, este livro serve como um excelente tratamento do conteúdo filosófico central de "Racionalidade: de IA aos Zumbis".

Talbott distingue vários pontos de vista na epistemologia bayesiana, incluindo a posição de E. T. Jaynes de que nem todas as hipóteses anteriores são igualmente razoáveis. [6], [7] Assim como Jaynes, Yudkowsky está interessado em complementar o critério de otimalidade bayesiano para a revisão de crenças com um critério de otimalidade para hipóteses anteriores. Isso alinha Yudkowsky com pesquisadores que buscam compreender melhor a IA de propósito geral por meio de uma teoria aprimorada do raciocínio ideal, como Marcus Hutter. [8] Para uma discussão mais ampla dos esforços filosóficos para naturalizar teorias do conhecimento, consulte Feldman. [9]

O "bayesianismo" é frequentemente contrastado com o "frequentismo". Alguns frequentistas criticam os bayesianos por tratarem as probabilidades como estados subjetivos de crença, em vez de frequências objetivas de eventos. Kruschke e Yudkowsky responderam que o frequentismo é ainda mais 'subjetivo' do que o bayesianismo, pois as atribuições de probabilidade do frequentismo dependem das intenções do experimentador [10].

É importante destacar que essa discordância filosófica não deve ser confundida com a distinção entre métodos de análise de dados bayesianos e frequentistas, os quais podem ser úteis quando utilizados corretamente. As ferramentas estatísticas bayesianas tornaram-se mais acessíveis desde a década de 1980, e sua informatividade, intuição e generalidade têm sido mais amplamente reconhecidas, resultando em 'revoluções Bayesianas' em muitas ciências. No entanto, os métodos frequentistas tradicionais ainda são mais populares e, em alguns contextos, continuam claramente superiores às abordagens bayesianas. *Doing Bayesian Data Analysis* (Fazendo Análise de Dados Bayesiana), de Kruschke, é uma introdução envolvente e acessível ao tema. [11]

Diante das evidências de que a formação em estatística — e em alguns outros campos, como a psicologia — melhora as habilidades de raciocínio fora da sala de aula, a literacia estatística é diretamente relevante para o projeto de superação de preconceitos. (Aulas de lógica formal e falácias informais não se mostraram

igualmente úteis) [12] [13].

Uma arte em sua infância

Concluímos com três sequências sobre aprimoramento individual e coletivo. "O Amadurecimento de Yudkowsky" oferece uma última análise aprofundada da dinâmica da crença irracional, destacando desta vez a história intelectual do próprio autor. "Desafiando o Difícil" questiona o que é necessário para resolver um problema verdadeiramente difícil — incluindo exigências que transcendem a racionalidade epistêmica. Por fim, "O Ofício e a Comunidade" aborda grupos de racionalidade e racionalidade em grupo, levantando questões cruciais:

- A racionalidade pode ser aprendida e ensinada?
- Em caso afirmativo, até que ponto é possível aprimorá-la?
- Como podemos ter certeza de que estamos vendo um efeito real em uma intervenção de racionalidade, e escolhendo a causa correta?
- Quais normas comunitárias facilitariam o processo de aprimoramento pessoal?
- Podemos colaborar eficazmente em problemas de grande escala sem comprometer nossa liberdade de pensamento e conduta?

Acima de tudo: o que está faltando? O que deveria ser incluído na próxima geração de guias sobre racionalidade — aquelas que substituirão este texto, aprimorarão seu estilo, testarão suas prescrições, complementarão seu conteúdo e se expandirão em direções totalmente novas?

Embora Yudkowsky tenha escrito esses ensaios devido a seus próprios erros filosóficos e desafios profissionais na teoria da IA, o material resultante mostrou-se útil para um público muito mais amplo. As postagens originais do blog inspiraram o crescimento da *Less Wrong*, uma comunidade de intelectuais e hackers com interesses comuns em ciência cognitiva, ciência da computação e filosofia. Yudkowsky e outros escritores da *Less Wrong* ajudaram a semear o movimento de altruísmo eficaz, um esforço vibrante e audacioso para identificar as instituições de caridade e causas humanitárias de maior impacto. Esses escritos também foram a base para o estabelecimento do *Center for Applied Rationality* (Centro de Racionalidade Aplicada), uma organização sem fins lucrativos que visa traduzir os resultados da ciência da racionalidade em técnicas utilizáveis para o aprimoramento pessoal.

Não sei o que o futuro reserva — que outros projetos ou ideias não convencionais podem se inspirar nessas páginas. Certamente, não enfrentamos escassez de desafios globais, e a arte da racionalidade aplicada é algo novo e em desenvolvimento. Não são muitos os racionalistas, e há muitas coisas deixadas por fazer.

Mas, para onde quer que você vá, leitor, pode cumprir bem o seu propósito.

Referências

[1] Jonathan Baron, Thinking and Deciding (Cambridge University Press, 2007).

[2] Keith J. Holyoak and Robert G. Morrison, The Oxford Handbook of Thinking and Reasoning (Oxford University Press, 2013).

[3] Bourget and Chalmers, “What Do Philosophers Believe?”

[4] Holt, “Thinking Inside the Boxes.”

[5] Gary L. Drescher, Good and Real: Demystifying Paradoxes from Physics to Ethics (Cambridge, MA: MIT Press, 2006).

[6] William Talbott, “Bayesian Epistemology,” in The Stanford Encyclopedia of Philosophy, Fall 2013, ed. Edward N. Zalta.

[7] Jaynes, Probability Theory.

[8] Marcus Hutter, Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability (Berlin: Springer, 2005), doi:10.1007/b138233.

[9] Richard Feldman, “Naturalized Epistemology,” in The Stanford Encyclopedia of Philosophy, Summer 2012, ed. Edward N. Zalta.

[10] John K. Kruschke, “What to Believe: Bayesian Methods for Data Analysis,” Trends in Cognitive Sciences 14, no. 7 (2010): 293–300.

[11] John K. Kruschke, Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan (Academic Press, 2014).

[12] Geoffrey T. Fong, David H. Krantz, and Richard E. Nisbett, “The Effects of Statistical Training on Thinking about Everyday Problems,” Cognitive Psychology 18, no. 3 (1986): 253–292, doi:10.1016/0010-0285(86)90001-0.

[13] Paul J. H. Schoemaker, “The Role of Statistical Knowledge in Gambling Decisions: Moment vs. Risk Dimension Approaches,” Organizational Behavior and Human Performance 24, no. 1 (1979): 1–17.

Parte X — O amadurecimento de Yudkowsky

292 — Meu declínio mortal na infância

Meus pais sempre menosprezaram o valor da inteligência. E dar ênfase ao valor do esforço, como recomendado pelas pesquisas mais recentes? Não, não é sobre esforço. É sobre experiência. Uma ferramenta inatingível para aplicar em uma criança inteligente, com certeza. Foi o que meus pais me disseram quando questionei a religião judaica, por exemplo. Tentei apresentar um argumento e me disseram algo como: 'A lógica tem limites; você entenderá quando for mais velho que a experiência é o importante, e então compreenderá a verdade do Judaísmo.' Não tentei novamente. Ao questionar o Judaísmo na escola, levei um tapa e não repeti a tentativa. Nunca fui um aprendiz lento.

Sempre que meus pais faziam algo imprudente, diziam: "Sabemos melhor porque temos mais experiência. Você entenderá quando for mais velho: maturidade e sabedoria são mais importantes do que inteligência."

Se isso foi uma tentativa de focar o jovem Eliezer na inteligência acima de tudo, foi o exemplo mais bem-sucedido de psicologia reversa que já ouvi falar.

Mas meus pais não são tão astutos, e os resultados não foram exatamente positivos.

Durante muito tempo, pensei que a moral desta história era que a experiência não era páreo para a pura inteligência inata. Só muito mais tarde, aos vinte anos, olhei para trás e percebi que não poderia ter sido mais inteligente do que meus pais antes da puberdade, com meu cérebro nem mesmo totalmente desenvolvido. Aos onze anos, quando já era quase um ateu declarado, não poderia ter vencido meus pais em uma disputa mental justa. Minhas pontuações no SAT[1] eram altas para uma criança de 11 anos, mas não teriam superado as pontuações de meus pais no SAT na idade adulta. Em uma luta justa, a inteligência e a experiência de meus pais poderiam ter esmagado qualquer criança pré-adolescente. Foi a irracionalidade que os derrotou; eles usaram sua inteligência somente para se autodestruir.

Mas essa compreensão veio muito mais tarde, quando minha inteligência já havia processado e destilado muitos anos de experiência.

A lição que tirei quando era jovem era que qualquer pessoa que subestimasse o valor da inteligência não entendia a inteligência de forma alguma. Minha própria inteligência influenciou todos os aspectos da minha vida, minha mente e minha personalidade; isso era extremamente óbvio, visto de trás para frente. 'Inteligência não tem nada a ver com sabedoria ou ser uma boa pessoa' — ah, e a autoconsciência não tem nada a ver com sabedoria ou ser uma boa pessoa? Modelar a si mesmo exige inteligência. Por um lado, é preciso inteligência suficiente para aprender psicologia evolucionista.

Somos as cartas que recebemos, e a inteligência é a mais injusta de todas essas cartas. Mais injusta do que a riqueza, a saúde ou o país de origem, mais injusta do que o seu ponto de referência de felicidade. As pessoas têm dificuldade em aceitar que a vida pode ser tão injusta; não é um pensamento feliz. 'A inteligência não é tão importante quanto X' é uma forma de se afastar da injustiça, recusando-se a lidar com ela, preferindo, em vez disso, um pensamento mais feliz. É uma tentação, tanto para aqueles que recebem cartas ruins quanto para aqueles que recebem cartas boas. Da mesma forma, subestimar a importância do dinheiro

1 NT. **SAT**: Teste padronizado para admissão em universidades dos EUA, administrado pelo *College Board*. Avalia leitura, escrita e matemática, com pontuação de 400 a 1600. Originalmente chamado *Scholastic Aptitude Test*, hoje é conhecido apenas pela sigla. Embora historicamente essencial, muitas instituições adotaram políticas *test-optional* recentemente.

é uma tentação tanto para os pobres quanto para os ricos.

Contudo, o jovem Eliezer era um entusiasta do transumanismo. Distribuir pontos de QI seria mais desafiador do que se tivesse nascido com recursos financeiros extras. No entanto, era um desafio passível de solução, algo a ser encarado e resolvido, mesmo que isso exigisse toda a minha vida. "Os fortes existem para servir aos fracos", escreveu o jovem Eliezer, "e só podem cumprir esse dever tornando os outros igualmente fortes". Eu discordava das tendências randianas e niilistas na ficção científica, e, como você deve ter percebido, o jovem Eliezer tinha uma propensão a levar as coisas ao extremo na direção oposta. Ninguém existe unicamente para servir. No entanto, eu tentei, e não me arrependo. Se chamar isso de uma extravagância adolescente, é raro ver a sabedoria adulta aprimorando-se.

Todos precisavam de mais inteligência, inclusive eu, como fiz questão de destacar. Longe de mim declarar uma nova ordem mundial comigo mesmo no topo; isso seria algo que um vilão estereotipado de ficção científica faria, ou pior, um adolescente comum, e eu nunca permitiria ser tão clichê. Não, todos precisavam ser mais inteligentes. Estávamos todos no mesmo barco, um pensamento excelente e edificante.

Eliezer$_{1995}$ devorou sua ficção científica. Ele possuía uma base moral e ética e conseguia identificar as armadilhas mais óbvias. Nenhuma discussão sobre o *Homo novis* para ele. Nenhuma linha divisória entre ele e os outros. Nenhuma filosofia elaborada para se colocar no topo da hierarquia. Era um modo de falha muito evidente. Sim, ele teve o cuidado de considerar a própria estupidez e nunca reivindicar superioridade moral. Bem, eu não vejo as coisas de maneira tão diferente agora, embora já não transforme minha ética em um espetáculo tão dramático. (Ou talvez seja mais correto dizer que sou mais contido quando me dou um momento de autocomplacência.)

Digo tudo isso para enfatizar que Eliezer$_{1995}$ não era tão digno de censura a ponto de falhar de maneira óbvia.

E então, em 1996, Eliezer encontrou o conceito de explosão de inteligência. Foi um lampejo de revelação? Pulei da cadeira e gritei "Eurisko!"? Não. Eu não era uma rainha do drama. Em retrospectiva, era extremamente óbvio que uma inteligência superior à humana mudaria o futuro de maneira mais fundamental do que qualquer avanço científico isolado. E imediatamente soube que isso seria minha missão pelo resto da vida: criar a explosão de inteligência. Não a nanotecnologia, como pensei aos onze anos; a nanotecnologia seria somente uma ferramenta gerada pela inteligência. Ora, a inteligência era ainda mais poderosa, uma bênção ainda maior, do que eu havia percebido antes.

Isso foi uma espiral de morte feliz? Como se descobriu mais tarde, sim: levou à aceitação de crenças ilusórias sobre a inteligência. Talvez o ponto de inflexão tenha sido quando comecei a acreditar que certamente o limite da velocidade da luz não seria uma barreira para a superinteligência.

(Como minha visão sobre inteligência mudou desde então… vejamos: quando penso nos problemas humanos, hoje em dia, penso primeiro na morte e no envelhecimento. Todos devem ter algum nível de inteligência, e o importante, do ponto de vista da teoria da diversão, é que ela deveria aumentar com o tempo, não diminuir como ocorre atualmente. Não é uma maneira inteligente de me sentir melhor? Mas não me esforço tanto agora para minimizar minha própria inteligência, porque isso seria apenas outra forma de chamar a atenção para ela. Sou inteligente para um ser humano, se o assunto surgir, e como me sinto sobre isso é uma questão pessoal.

A ideia de que a inteligência é a alavanca que ergue os mundos permanece a mesma. Exceto que a inteligência se tornou menos misteriosa para mim, de modo que agora vejo mais claramente a inteligência como algo intrínseco à física. As superinteligências podem tornar-se FTL[2] se permitido pelas verdadeiras leis físicas, e se não, então não. Não é nada impossível, mas eu não apostaria nisso.)

Mas o verdadeiro desvio ocorreu mais tarde, quando alguém perguntou: "Ei, como você sabe que a superinteligência será moral? Inteligência não tem nada a ver com ser uma boa pessoa, você sabe — isso é o que chamamos de sabedoria, jovem prodígio."

---

2 NT. ***Faster Than Light*** Mais rápido que a luz, em português. Ou seja, o autor está especulando sobre a possibilidade de superinteligências serem capazes de viajar ou se comunicar acima da velocidade da luz, caso as leis fundamentais da física o permitam.

E para o jovem Eliezer, isso parecia uma negação flagrante. Certamente, seu próprio código ético cuidadosamente elaborado foi concebido usando sua inteligência e baseado nela. Qualquer tolo poderia perceber que a inteligência está intrinsecamente ligada à ética, moralidade e sabedoria; tente explicar o Dilema do Prisioneiro a um chimpanzé, por exemplo?

Assim, para o jovem Eliezer, a superinteligência necessariamente implicaria em supermoralidade.

Ele argumentava: “Os pais fazem tudo o que dizem aos filhos para não fazer, e é assim que sabem que não devem fazê-lo”.

293 — Meu melhor e pior erro

No último capítulo, eu abordei a espiral de desencanto afetivo do jovem Eliezer em relação a algo que ele denominou como "inteligência". Eliezer$_{1996}$, ou mesmo Eliezer$_{1999}$, neste caso, teriam se recusado a tentar estabelecer uma definição matemática — recusado consciente e deliberadamente. Na verdade, ele teria hesitado em dar qualquer definição de "inteligência".

Por quê? Porque há um problema comum de isca e troca na IA, onde você define "inteligência" como significando algo como "raciocínio lógico" ou "a capacidade de tirar conclusões quando apropriado", e então você constrói um provador de teoremas barato ou um raciocinador não monotônico ad-hoc e diz: "Eis que implementei a inteligência!" As pessoas criaram definições pobres de inteligência — focando em correlatos em vez de essências — e depois perseguiram a definição superficial que escreveram, esquecendo-se, você sabe, da inteligência real. Não é como se Eliezer$_{1996}$ quisesse construir uma carreira em Inteligência Artificial. Ele apenas desejava uma mente capaz de criar nanotecnologia. Portanto, ele não foi tentado a redefinir a inteligência apenas para preencher um artigo.

Olhando para trás, parece-me que muitos dos meus erros podem ser definidos em termos de ter ido longe demais na outra direção ao ver a estupidez alheia. Tendo presenciado tentativas frequentes de definir "inteligência" sendo abusadas, recusei-me a defini-la. E se eu dissesse que a inteligência era X e, na verdade, não fosse X? Eu sabia intuitivamente o que procurava — algo poderoso o suficiente para desmontar estrelas e transformá-las em matéria-prima — e não queria cair na armadilha de ser distraído por definições.

Da mesma forma, tendo visto tantos projetos de IA serem prejudicados pela inveja da física — tentando manter a matemática simples e elegante e, como resultado, sendo limitados a sistemas de brinquedo — generalizei que qualquer matemática simples o suficiente para ser formalizada em uma equação organizada provavelmente não iria funcionar para, você sabe, inteligências reais. "Exceto o Teorema de Bayes", acrescentou Eliezer$_{2000}$, o que, dependendo do ponto de vista, pode mitigar completamente a ofensa ou mostrar que ele deveria ter desconfiado de generalizações em vez de tentar adicionar uma única exceção.

Se você está se perguntando por que Eliezer$_{2000}$ pensava assim — não acreditava na matemática da inteligência — bem, é difícil para mim lembrar disso há muito tempo. Certamente não foi porque eu não gostava de matemática. Se eu tivesse que apontar uma causa raiz, seria a leitura de poucos livros, muito populares e os livros errados sobre Inteligência Artificial.

Mas na época, eu não acreditava que as respostas viriam da Inteligência Artificial; eu geralmente considerava isso um campo problemático e estagnado. Portanto, não é surpreendente que eu tenha dedicado pouco tempo para investigá-lo. Eu acreditava nos clichês sobre as promessas exageradas da Inteligência Artificial. Você pode enquadrar isso no padrão de "muito longe na direção oposta" — o campo não cumpriu suas promessas, então eu estava pronto para descartá-lo. Como resultado, não investiguei o suficiente para descobrir que a matemática não era falsa.

Minha descrença juvenil na matemática da inteligência geral foi simultaneamente um dos meus piores erros de todos os tempos e um dos meus melhores erros de todos os tempos.

Como não acreditava que poderia haver respostas simples para a inteligência, eu me dediquei à leitura sobre psicologia cognitiva, neuroanatomia funcional, neuroanatomia computacional, psicologia evolutiva, biologia evolutiva e mais de um ramo da Inteligência Artificial. Quando surgiam ideias que pareciam simples e brilhantes, eu não parava por aí, nem corria para tentar implementá-las, porque sabia que mesmo

que fossem verdadeiras, mesmo que fossem necessárias, não seriam suficientes: inteligência não deveria ser simples, não deveria ter uma resposta que coubesse em uma camiseta. Era para ser um grande quebra-cabeça com muitas peças; e quando você encontrava uma peça, você não saía correndo segurando-a no alto em triunfo, você continuava procurando. Tentar construir uma mente com uma única peça faltando poderia resultar em nada de interessante acontecendo.

Enganei-me ao pensar que a Inteligência Artificial, o campo acadêmico, era um terreno baldio desolado; e ainda mais equivocado ao imaginar que não poderia haver uma matemática específica para a inteligência. No entanto, não lamento ter estudado, por exemplo, neuroanatomia funcional, embora hoje acredite que uma Inteligência Artificial não deva se assemelhar em nada a um cérebro humano. Estudar neuroanatomia significou que adotei a ideia de que, se dividíssemos uma mente em partes, essas partes seriam coisas como "córtex visual" e "cerebelo" – em vez de "módulo de negociação no mercado de ações" ou "módulo de raciocínio de senso comum", que é um caminho errado padrão em IA.

Explorar campos como neuroanatomia funcional e psicologia cognitiva me proporcionou uma perspectiva muito diferente sobre como as mentes deveriam ser, comparada àquela obtida apenas por meio da leitura de livros sobre IA – mesmo os bons livros sobre IA.

Quando eliminamos todas as conclusões e justificativas equivocadas e nos perguntamos apenas o que essa crença levou o jovem Eliezer a realmente fazer...

Aí a crença de que a Inteligência Artificial estava em crise e que as verdadeiras respostas teriam que vir de campos externos mais saudáveis o levou a estudar diversas ciências cognitivas;

A convicção de que a IA não poderia oferecer respostas simples o impediu de interromper prematuramente uma ideia brilhante, incentivando-o a acumular muitas informações;

A compreensão de que não se deveria definir inteligência resultou em um cenário em que ele dedicou muito tempo ao estudo do problema antes de, anos depois, começar a propor sistematizações.

É a isso que me refiro quando digo que este é um dos meus maiores erros de todos os tempos.

Ao olhar para trás, anos depois, tirei uma lição valiosa disso:

O que você realmente acaba fazendo reflete a verdadeira razão inteligente pela qual está fazendo isso.

Compare o raciocínio inteligente incrível que o leva a estudar muitas ciências com o raciocínio inteligente incrível que sugere que você não precisa ler todos esses livros. Posteriormente, quando seu raciocínio inteligente incrível se revelar inadequado, você estará em uma posição muito melhor se seu raciocínio inteligente incrível foi do primeiro tipo.

Ao revisitar meu passado, fico impressionado com o número de sucessos semi-acidentais, às vezes em que fiz algo certo, pelos motivos errados. Do seu ponto de vista, você poderia atribuir isso ao princípio antrópico: se eu tivesse seguido por um verdadeiro beco sem saída, provavelmente você não teria notícias minhas neste livro. Do meu ponto de vista, ainda é algo constrangedor. Minha educação Racionalista Tradicional proporcionou muitos vieses direcionais para esses "sucessos acidentais" – me levou a racionalizar razões para estudar em vez de não estudar, evitou que eu me perdesse completamente, ajudou-me a corrigir erros. Ainda assim, nada disso foi a ação certa pelos motivos certos, e é assustador olhar para trás e revisitar minha história de juventude. Um dos meus principais objetivos ao escrever sobre *Overcoming Bias* é deixar um rastro até onde acabei por acidente – para evitar o papel que a sorte desempenhou em minha própria formação como racionalista.

Então, por que considero este um dos meus piores erros de todos os tempos? Porque às vezes, o termo "informal" é apenas outra maneira de dizer "manter padrões baixos". Eu tinha razões surpreendentemente inteligentes para não definir com precisão o termo "inteligência" e alguns dos meus outros conceitos, como o fato de que outras pessoas se perderam ao tentar defini-los. Este era um portal pelo qual o raciocínio descuidado poderia se infiltrar.

Então, eu deveria ter adiantado e tentado criar uma definição precisa imediatamente? Não, todas

as razões pelas quais eu sabia que isso era a coisa errada a fazer estavam corretas; você não pode conjurar a definição correta do nada se seu conhecimento não for adequado.

Você não pode chegar à definição de fogo se não conhecer átomos e moléculas; é melhor dizer "aquela coisa laranja brilhante". E você precisa ser capaz de discutir essa coisa alaranjada, mesmo que não consiga dizer exatamente o que é, para investigar o incêndio. Mas hoje em dia eu diria que todo raciocínio nesse nível é algo em que não se pode confiar – na verdade, é algo que você faz no caminho para conhecer melhor, mas não confia nele, não coloca seu peso sobre isso. Disso, você não tira conclusões firmes, não importa quão inevitável o raciocínio informal possa parecer.

O jovem Eliezer colocou seu peso no lugar errado – pisou em uma armadilha carregada.

294 — Criado em tecnofilia

Meu pai costumava dizer que se o sistema atual existisse há cem anos, os automóveis teriam sido proibidos para proteger a indústria de selas.

Uma das minhas principais influências na infância foi a leitura de *A Step Farther Out* (Um passo mais além), de Jerry Pournelle, aos nove anos. Foi a resposta de Pournelle a Paul Ehrlich e ao Clube de Roma, que afirmavam nas décadas de 1960 e 1970 que a Terra estava ficando sem recursos e que a fome massiva estava a apenas alguns anos de distância. Foi uma resposta à chamada quarta lei da termodinâmica de Jeremy Rifhin; foi uma resposta a todas as pessoas que temiam a energia nuclear e tentavam regulá-la até o esquecimento.

Cresci em um mundo onde as linhas de demarcação entre os mocinhos e os bandidos eram bastante claras; não uma batalha final apocalíptica, mas uma batalha que teve de ser travada repetidamente, uma batalha onde se podiam ver os ecos históricos que remontavam à Revolução Industrial, e onde se podiam reunir as provas históricas sobre os resultados reais.

De um lado estavam os cientistas e engenheiros que impulsionaram todos os aumentos do nível de vida desde a Idade das Trevas, cujo trabalho apoiou luxos como a democracia, uma população instruída, uma classe média, a proibição da escravatura.

Do outro lado, aqueles que outrora se opuseram à vacinação contra a varíola, aos anestésicos durante o parto, às máquinas a vapor e ao heliocentrismo: os teólogos apelando ao regresso a uma idade perfeita que nunca existiu, os políticos homens brancos idosos que impuseram os seus caminhos, os grupos de interesse especial que ficaram para perder e muitos para quem a ciência era um livro fechado, temendo o que não conseguiam compreender.

E tentando fazer o papel do meio, os pretendentes à Sabedoria Profunda, expressando pensamentos escondidos sobre como a tecnologia beneficia a humanidade, mas apenas quando é devidamente regulamentada – alegando, desafiando o fato histórico bruto, que a ciência em si não era nem boa, nem má – estabelecendo comitês burocráticos de aparência solene para fazerem uma demonstração ostensiva da sua cautela – e à espera dos seus aplausos. Como se a verdade fosse sempre um compromisso. E como se alguém pudesse realmente ver tão longe. Será que a humanidade teria feito melhor se houvesse um debate público sincero e preocupado sobre a adoção do fogo e se tivessem sido criados comitês para supervisionar a sua utilização?

Quando entrei no problema, comecei com alergia a qualquer coisa que correspondesse a um padrão "Ah, mas a tecnologia tem riscos e benefícios, pequenino". A presunção de culpa era que você estava tentando receber aplausos baratos ou tentando regular secretamente a tecnologia até o esquecimento. E de qualquer forma, ignorando imensamente o registo histórico em favor de tecnologias com as quais as pessoas outrora se preocuparam.

Robin Hanson levantou o tema da lenta aprovação pela FDA de medicamentos aprovados em outros países. Alguém nos comentários apontou que a talidomida foi vendida em 50 países sob 40 nomes, mas que apenas uma pequena quantidade foi doada nos EUA, de modo que nasceram 10.000 crianças malformadas em todo o mundo, mas apenas 17 crianças nos EUA.

Mas quantas pessoas morreram devido à aprovação lenta nos EUA, de medicamentos aprovados mais rapidamente em outros países – todos os medicamentos que não deram errado? E faço esta pergunta

porque é sobre isso que se pode tentar recolher estatísticas – isto não diz nada sobre todos os medicamentos que nunca foram desenvolvidos porque o processo de aprovação é demasiado longo e dispendioso. De acordo com esta fonte, o processo de aprovação mais longo da FDA evita 5.000 vítimas por ano através do rastreio de medicamentos considerados prejudiciais, e causa pelo menos 20.000 a 120.000 vítimas por ano apenas por atrasar a aprovação dos medicamentos benéficos que ainda são desenvolvidos e eventualmente aprovados.

Portanto, há realmente uma razão para ter cautela ao lidar com pessoas que afirmam: "Ah, mas a tecnologia possui tanto riscos quanto benefícios." Existe um registro histórico que evidencia um excesso de conservadorismo, com muitas mortes silenciosas causadas pela regulamentação compensadas por algumas mortes visíveis resultantes da falta de regulamentação. Se você está verdadeiramente no meio-termo, por que não afirmar: "Ah, mas a tecnologia tem benefícios bem como riscos"?

Bem, isso não é uma descrição tão equivocada dos adversários. (Exceto que é crucial enfatizar um pouco mais que eles não são mutantes malignos, mas seres humanos convencionais agindo sob uma visão de mundo diferente, que os coloca na razão; alguns deles serão inevitavelmente mais competentes do que outros, e a competência faz diferença, muito.) Mesmo olhando para trás, não creio que minha afinidade pela tecnologia na infância estivesse tão equivocada sobre quem eram os adversários e qual foi o principal erro. Contudo, é sempre mais fácil dizer o que não fazer do que acertar. E uma das minhas falhas fundamentais, naquela época, era acreditar que evitar ao máximo tudo o que os adversários estavam fazendo faria de você um mocinho.

Particularmente prejudicial, creio eu, foi o mau exemplo dado pelos aspirantes a Sabedoria Profunda, tentando traçar um caminho intermediário; sorrindo condescendentemente tanto para tecnófilos quanto para tecnófobos, rotulando-os como imaturos. Na verdade, este é um caminho equivocado; e, de fato, a ideia de tentar estabelecer um meio-termo em geral está geralmente errada. O Caminho Correto não é um compromisso com nada; é a expressão clara dos seus próprios critérios.

Isso, no entanto, tornou mais difícil para o jovem Eliezer se afastar do veredicto de ataque direto, pois qualquer afastamento parecia alinhar-se com os aspirantes a Sabedoria Profunda.

A primeira fissura na minha afinidade pela tecnologia na infância surgiu, acredito eu, em 1997 ou 1998, quando percebi meus colegas tecnófilos dizendo coisas tolas sobre como a nanotecnologia molecular seria fácil de gerenciar. (Como você deve estar notando novamente, o jovem Eliezer era intensamente motivado pela capacidade de encontrar falhas – eu até tinha uma filosofia pessoal sobre por que esse tipo de coisa era uma boa ideia.)

Houve um debate sobre a nanotecnologia molecular e se o ataque seria mais fácil de realizar do que a defesa. E havia pessoas argumentando que a defesa seria fácil. No domínio da nanotecnologia, pelo amor de Ghu, matéria programável, quando parece que nem conseguimos resolver o problema de segurança das redes de computadores, onde podemos observar e controlar tudo. As pessoas estavam falando sobre paredes diamantinas inexpugnáveis. Observei que o diamante não se compara a uma arma nuclear, que o ataque se derrotou na defesa desde 1945 e que a nanotecnologia não parecia capaz de mudar isso.

E quando o debate terminou, parece que o jovem Eliezer – envolvido no calor da discussão – conseguiu perceber, pela primeira vez, que a sobrevivência da vida inteligente originada na Terra estava em risco.

Parece tão estranho, ao olhar para trás, pensar que houve um momento em que acreditava que apenas vidas individuais estavam em jogo no futuro. Como o mundo parecia mais acolhedor para se viver... embora não estivesse exatamente pensando assim naquela época. Eu não descartava a possibilidade, mas simplesmente não conseguia visualizá-la. Quando o assunto finalmente surgiu, a compreendi. Não me recordo exatamente como essa compreensão se desdobrou. Há uma razão pela qual me refiro ao meu passado na terceira pessoa.

Pode parecer que Eliezer_{1998} foi um completo tolo, mas isso seria uma explicação reconfortante, de certa forma; a verdade é mais assustadora. Elieze_{r1998} era um Racionalista Tradicional perspicaz, no que diz respeito a algumas coisas. Eu compreendia que as hipóteses precisavam ser testáveis, sabia que a racionalização não era uma operação mental permitida, conhecia as regras do Tabu do Racionalista e estava obcecado pela autoconsciência... Não entendia muito bem o conceito de "respostas misteriosas"... e nada de Bayes ou

Kahneman. Mas um Racionalista Tradicional perspicaz, muito acima da média... E daí? A natureza não nos classifica em uma curva. Um desvio do Caminho, uma influência inadequada em seus processos de pensamento, pode anular todas as outras proteções.

Uma das principais lições que tiro ao revisitar minha história pessoal é que não é surpreendente muitas pessoas acreditarem que "inteligência não é tudo" no mundo real, ou que os racionalistas não se saem melhor, na prática. Um pouco de racionalidade, ou mesmo muita racionalidade, não ultrapassa a barreira astronomicamente alta necessária para as coisas começarem realmente a funcionar.

Que minha interpretação equivocada do Caminho Certo não seja atribuída a Jerry Pournelle, ao meu pai, ou à ficção científica, em geral. Acredito que a personalidade do jovem Eliezer influenciou a seletividade das partes de seus ensinamentos que foram transmitidas. Não é como se Pournelle não tivesse dito: as regras mudam quando você deixa a Terra, o berço; se você for descuidado ao vedar sua roupa pressurizada apenas uma vez, você morre. Ele disse isso muitas vezes. Mas as palavras não pareciam realmente importantes, porque isso era algo que acontecia com personagens secundários nos romances – o personagem principal geralmente não morria no meio, por algum motivo.

Qual foi a lente através da qual filtrei esses ensinamentos? Ter esperança. Otimismo. Anseio por um futuro melhor. Esse foi o significado fundamental de "A Step Farther Out" para mim, a lição que aprendi em contraste com a desgraça e tristeza do Sierra Club. De um lado estava a racionalidade e a esperança; do outro, a ignorância e o desespero.

Alguns adolescentes acham que são imortais e andam de motocicleta. Eu não tinha essa ilusão e estava bastante relutante em aprender a dirigir, considerando o quão inseguros pareciam aqueles pedaços de metal em movimento. Mas havia algo mais importante para mim do que minha própria vida: O Futuro. Eu agia como se fosse imortal. Vidas poderiam ser perdidas, mas não o Futuro.

E quando percebi que a nanotecnologia realmente representaria um desafio potencialmente ao nível da extinção?

O jovem Eliezer refletiu, de maneira explícita: "Meu Deus, como pude não perceber algo que deveria ser evidente? Talvez tenha me envolvido emocionalmente demais com os benefícios que esperava da tecnologia; possivelmente recuei diante da ideia da extinção humana."

E então...

Não proclamei um "Pare, Derreta e Pegue Fogo". Não reavaliei todas as conclusões que havia tirado com minha postura anterior. Consegui integrá-lo à minha visão de mundo de alguma forma, com poucas mudanças propagadas. Velhas ideias e planos foram desafiados, mas minha mente encontrou razões para mantê-los. Infelizmente, não ocorreu nenhum colapso sistêmico.

De maneira mais marcante, decidi que precisávamos avançar o mais rápido possível na IA, desenvolvê-la antes da nanotecnologia. Exatamente como planejava fazer inicialmente, mas agora por uma razão diferente.

Acredito que isso reflita a natureza da maioria dos seres humanos, não é mesmo? A Racionalidade Tradicional não foi suficiente para mudar isso.

Contudo, chegou um momento em que percebi completamente meu erro. Foi necessário apenas um golpe mais forte na cabeça.

295 — Um prodígio de refutação

"Meu declínio mortal na infância" descreveu o impulso central que me levou ao meu erro, uma espiral de declínio emocional em torno daquilo que Eliezer$_{1996}$ chamou de "inteligência". Eu também era um tecnófílo, imune ao receio do futuro. Li muita ficção científica centrada na ética da personalidade - na qual o medo do desconhecido coloca a humanidade na posição dos vilões, maltratando alienígenas ou IAs conscientes porque eles "não são humanos".

Isso faz parte da mentalidade que você desenvolve com a ficção científica - definindo seu grupo, sua tribo, de maneira adequada e ampla. Daí o meu endereço de e-mail, sentience@pobox.com.

Portanto, Eliezer$_{1996}$ planeja criar superinteligência para o bem da humanidade e de toda a vida consciente.

Inicialmente, penso eu, a questão de saber se uma superinteligência seria ou poderia ser boa, ou má não me ocorreu realmente como um tópico de discussão separado. Apenas a intuição padrão de: "Certamente nenhuma mente tão avançada seria tola o suficiente para transformar a galáxia em clipes de papel; certamente, sendo tão inteligente, saberá o que é certo, muito melhor do que um ser humano poderia."

Até que eu me apresentei e compartilhei minha busca em uma lista de discussão transumanista, recebendo respostas nos moldes de (de memória):

A moralidade é arbitrária - se você diz que algo é bom ou ruim, não pode estar certo ou errado sobre isso. Uma superinteligência formaria sua própria moralidade.

No final, todos cuidam de seus próprios interesses. Uma superinteligência não seria diferente; apenas aproveitaria todos os recursos.

Pessoalmente, sou humano, então sou a favor dos humanos, não das inteligências artificiais. Não acho que devamos desenvolver essa tecnologia. Em vez disso, deveríamos desenvolver a tecnologia para fazer o upload de humanos primeiro.

Ninguém deveria desenvolver uma IA sem um sistema de controle que a monitore e garanta que ela não possa fazer nada de ruim.

Bem, tudo isso está obviamente equivocado, pensa Eliezer1996, e ele começou a desmontar os argumentos de seus oponentes. (Fiz isso principalmente em outros ensaios, e o restante é deixado como exercício para o leitor.)

Não é que Eliezer1996 tenha raciocinado explicitamente: "O homem mais estúpido do mundo diz que o Sol está brilhando; portanto, está escuro." Mas Eliezer1996 era um Racionalista Tradicional; ele foi inculcado com a metáfora da ciência como uma luta justa entre lados que assumem posições diferentes, despojada de mera violência e outros exercícios de força política, para que, idealmente, o lado com os melhores argumentos possa vencer.

É mais fácil apontar onde o argumento alheio está equivocado do que acertar a questão; e Eliezer1996 era muito habilidoso em identificar falhas. (Eu também. Não é como se você pudesse resolver o perigo desse poder recusando-se a se preocupar com falhas.) Do ponto de vista de Eliezer1996, parecia-lhe que o lado escolhido estava vencendo a luta – que ele estava formulando argumentos melhores do que seus oponentes

- então por que ele mudaria de lado?

Portanto, está escrito: “Como este mundo contém muitos cuja compreensão da racionalidade é péssima, os estudantes iniciantes da racionalidade ganham argumentos e adquirem uma visão exagerada de suas próprias capacidades. Mas é inútil ser superior: a vida não se gradua numa curva. O melhor físico da Grécia antiga não conseguiu calcular a trajetória de uma maçã caindo. Não há garantia de que a adequação seja possível, dado o seu maior esforço; portanto, não pense se os outros estão fazendo pior.”

Você não pode confiar em ninguém para dissuadi-lo de seus erros; você não pode confiar em mais ninguém para salvá-lo; você e somente você é obrigado a encontrar as falhas em suas posições; se você deixar esse fardo de lado, não espere que mais ninguém o carregue. E pergunto-me se esse conselho acabará por não ajudar a maioria das pessoas, até que elas tenham perdido o próprio pé, dizendo o tempo todo para si mesmas, corretamente: “É evidente que estou ganhando esta discussão”.

Hoje procuro não considerar nenhum ser humano como meu adversário. Isso só leva ao excesso de confiança. É a Natureza que estou enfrentando, que não alinha seus desafios com sua habilidade, que não é obrigada a lhe oferecer uma chance justa de vencer em troca de um esforço diligente, que não se importa se você é o melhor que já viveu, se você não for bom o suficiente.

Mas voltemos a 1996. Eliezer$_{1996}$ segue a intuição básica de “Certamente uma superinteligência saberá melhor do que nós o que é certo”, e derruba imediatamente vários argumentos apresentados contra a sua posição. Ele era habilidoso nesse sentido, como você pode ver. Ele até tinha uma filosofia pessoal sobre por que era sensato procurar falhas nas coisas e assim por diante.

Não quero usar isso como desculpa, afirmando que ninguém que contestou Eliezer$_{1996}$ apresentou realmente a dissolução do mistério – a redução completa da moralidade que analisa todos os seus processos cognitivos, desmembrando a ‘moralidade’ em um passo a passo detalhado dos algoritmos que tornam a moralidade tangível para ele. Considere isso mais como uma acusação, uma medida do nível de Eliezer$_{1996}$, indicando que ele precisaria da solução completa que lhe foi fornecida para apresentar um argumento que ele não pudesse refutar.

Os poucos filósofos presentes não o livraram de suas dificuldades. Não foi como se um filósofo dissesse: ‘Desculpe, a moralidade está compreendida, é uma questão resolvida na ciência cognitiva e na filosofia, e sua perspectiva está simplesmente errada’. A natureza da moralidade ainda é uma questão em aberto na filosofia; o debate continua. Um filósofo sentir-se-ia compelido a oferecer-lhe uma lista de argumentos clássicos de todas as perspectivas – a maioria dos quais Eliezer$_{1996}$ é suficientemente inteligente para refutar, levando-o a concluir que a filosofia é um terreno estéril.

Mas espere. A situação piora.

Não recordo exatamente quando – pode ter sido em 1997 –, mas meu eu mais jovem, vamos chamá-lo de Eliezer$_{1997}$, começou a argumentar de maneira inevitável que criar superinteligência é a escolha certa a fazer.

296 — A pura loucura da juventude inexperiente

> Aqui se expressa a pura ousadia da juventude inexperiente; a temeridade de uma ignorância tão profunda que só seria possível para alguém de sua efêmera extirpe..[3].
>
> —Gharlane of Eddore [1]

Houve um tempo, anos atrás, em que propus uma resposta enigmática para uma pergunta igualmente misteriosa – como sugeri em várias ocasiões. A questão misteriosa para a qual propus uma resposta misteriosa não era, contudo, a consciência – ou melhor, não apenas a consciência. Não, o erro mais embaraçoso foi ter uma visão misteriosa da moralidade.

Posterguei a discussão sobre isso até agora, após a série sobre metaética, porque queria deixar claro que Eliezer_{1997} havia entendido erroneamente.

Da última vez que paramos, Eliezer_{1997}, insatisfeito em argumentar intuitivamente que a superinteligência seria moral, estava propondo argumentar de forma inevitável que criar superinteligência era a coisa certa a fazer.

"Bem" (disse Eliezer_{1997}), "vamos começar fazendo a pergunta: a vida tem, de fato, algum sentido?

"Não sei", respondeu Eliezer_{1997} imediatamente, com uma certa nota de autocongratulação por admitir sua própria ignorância sobre este tema onde tantos outros pareciam certos.

"Mas," ele continuou—

(Tenha sempre cuidado quando uma admissão de ignorância é seguida por "Mas".)

"Mas, se supormos que a vida não tenha sentido – que a utilidade de todos os resultados é igual a zero – essa possibilidade anula qualquer cálculo de utilidade esperada. Pode-se, portanto, sempre agir como se a vida fosse conhecida por ter significado, mesmo que não saibamos qual é esse significado. Como podemos descobrir tal significado? Considerando que os humanos continuam debatendo isso, provavelmente é um problema muito difícil para os humanos resolverem. Assim, precisamos de uma superinteligência para resolver esse problema para nós. Quanto à possibilidade de não haver justificação lógica para uma preferência em detrimento de outra, então, neste caso, não é mais certo ou mais errado construir uma superinteligência, do que fazer qualquer outra coisa. Esta é uma possibilidade real, mas não se encaixa em qualquer tentativa de calcular a utilidade esperada – devemos simplesmente ignorá-la. Enquanto alguém afirma que uma superinteligência exterminaria a humanidade, ou está argumentando que exterminar a humanidade é, de fato, a coisa certa a fazer (mesmo que não vejamos razão para que isso aconteça) ou está argumentando não haver nada certo a fazer (nesse caso, o argumento de que não devemos construir inteligência se derrota).

Ugh... Esse foi um parágrafo realmente difícil de escrever. Meu "eu" do passado é sempre minha kriptonita mais concentrada, porque meu "eu" do passado é exatamente todas aquelas coisas das quais o eu moderno instalou alergias para bloquear. É verdade que se diz que os pais fazem todas as coisas que dizem

3 NT. Texto original em inglês. *There speaks the sheer folly of callow youth; the rashness of an ignorance so abysmal as to be possible only to one of your ephemeral race…*

aos filhos para não fazerem, e é assim que sabem que não devem fazê-las; aplica-se também entre eus passados e futuros.

Quão falho é o argumento de Eliezer$_{1997}$? Eu nem conseguia contar os caminhos. Sei que a memória é falível, reconstruída cada vez que nos lembramos, e por isso não confio na montagem dessas peças antigas usando minha mente moderna. Não me peça para reler meus escritos antigos; isso é muita dor.

Mas parece claro que eu estava pensando na utilidade como uma espécie de coisa, uma propriedade inerente. Portanto, "a vida não tem sentido" correspondia à utilidade = 0. Mas é claro que o argumento funciona igualmente bem com utilidade = 100, de modo que se tudo tiver sentido, mas for igualmente significativo, isso também deverá ocorrer... Certamente eu não estava pensando em uma função de utilidade como uma estrutura afim nas preferências. Eu estava pensando em "utilidade" como um nível absoluto de valor inerente.

Eu estava pensando no dever como uma espécie de essência puramente abstrata de competência, aquilo que faz você fazer alguma coisa; de modo que claramente qualquer mente que derivasse um dever estaria vinculada a ele. Daí a suposição, que Eliezer$_{1997}$ nem sequer pensou em observar explicitamente, de que uma lógica que obriga uma mente arbitrária a fazer algo é exatamente a mesma que os seres humanos querem dizer e a que se referem quando pronunciam a palavra "certo"...

Mas agora estou tentando contar as maneiras e, se você estiver acompanhando, deverá ser capaz de lidar com isso sozinho.

Um aspecto importante de todo esse fracasso foi que, ao provar que não valia a pena considerar o caso "a vida não tem sentido", não achei necessário definir rigorosamente "inteligência" ou "significado". Eu já havia inventado uma razão inteligente para não tentar ser totalmente formal e rigoroso ao tentar definir "inteligência" (ou "moralidade") – ou seja, todas as armadilhas e truques que o passado da IA, filósofos e moralistas usaram com definições que perderam o foco.

Tiro a seguinte lição: não importa quão inteligente seja a justificativa para relaxar seus padrões ou fugir de alguma exigência de rigor, ela irá explodir você do mesmo jeito.

E outra lição: eu era hábil em refutação. Se eu tivesse aplicado à minha própria posição o mesmo nível de rejeição baseada em qualquer falha que usei para derrotar os argumentos apresentados contra mim, então eu teria me concentrado na lacuna lógica e rejeitado a posição - se eu quisesse. Se eu tivesse o mesmo nível de preconceito contra isso que tive contra outras posições no debate.

Mas isso foi antes de eu tomar conhecimento de Kahneman, antes de ouvir falar do termo 'ceticismo motivado', antes de integrar o conceito de um estado de incerteza precisamente correto que resume todas as evidências, e antes de compreender a letalidade de perguntar 'Posso acreditar?' para posições preferidas e 'Sou obrigado a acreditar?' para posições não apreciadas. Eu era apenas um racionalista tradicional que via o processo científico como um árbitro entre pessoas que assumiam posições e as debatiam, onde o melhor lado venceria.

Minha falha final não foi um apreço pela 'inteligência', nem de qualquer quantidade de tecnofilia e ficção científica que exaltasse a irmandade da senciência. Certamente não foi minha habilidade em identificar falhas. Nenhuma dessas coisas teria me desencaminhado se eu tivesse mantido sempre um padrão elevado de rigor e não adotado nenhuma posição de maneira diferente. Ou mesmo se eu tivesse examinado minha posição vaga preferida com a mesma exigência de rigor que aplicava aos contra-argumentos.

Mas eu não estava muito interessado em tentar refutar a minha crença de que a vida tinha sentido, uma vez que meu raciocínio seria sempre dominado por casos em que a vida tinha sentido.

E com a explosão de inteligência em jogo, pensei que só tinha que prosseguir a toda velocidade usando os melhores conceitos que pudesse empregar no momento, sem pausar e desligar tudo enquanto buscava uma definição perfeita que tantos outros haviam estragado...

Não.

Não, você não utiliza os melhores conceitos que pode empregar no momento.

É a Natureza que o julga, e a Natureza não aceita nem mesmo as desculpas mais justas. Se você não atender ao padrão, falhará. É simples assim. Não há nenhum argumento inteligente para explicar por que você deve se contentar com o que tem, porque a Natureza não ouvirá esse argumento, não o perdoará só porque havia tantas justificativas excelentes para a velocidade.

Todos sabemos o que aconteceu com Donald Rumsfeld quando entrou em guerra com o exército que tinha, em vez do exército de que precisava.

Talvez Eliezer_{1997} não pudesse simplesmente criar o modelo correto do nada. (Embora quem sabe o que teria acontecido se ele tivesse realmente tentado...) E não teria sido prudente da parte dele parar completamente de refletir, até que a precisão surgisse de forma inesperada.

Mas também não foi apropriado Eliezer_{1997} contentar-se apenas com seu "melhor palpite" na ausência de precisão. É aceitável utilizar conceitos vagos em processos de pensamento provisórios, enquanto se busca uma resposta mais precisa, mantendo-se insatisfeito com as orientações vagas atuais e sem depositar plenamente confiança nelas. A construção de uma superinteligência não se sustenta em um entendimento provisório, nem mesmo no "melhor" entendimento vago disponível. Esse foi meu erro – acreditar que usar a expressão "melhor palpite" justificaria qualquer coisa. Havia apenas o padrão que eu não fui capaz de atender.

Claro que Eliezer_{1997} não desejava desacelerar na corrida em direção à explosão da inteligência, com tantas vidas em jogo, e a própria sobrevivência da vida inteligente originada na Terra, caso chegássemos à era das nano-armas antes da era da superinteligência—

A natureza não se importa com essas razões justas. Existe apenas o padrão astronomicamente alto necessário para o sucesso. Ou você está à altura, ou fracassa. É só isso.

O apocalipse não precisa ser justo com você.

O apocalipse não precisa oferecer a você uma oportunidade de sucesso

Em troca do que você já trouxe para a mesa.

A dificuldade do apocalipse não se ajusta às suas habilidades.

O preço do apocalipse não se alinha aos seus recursos.

Se o apocalipse exigir algo irracional e você tentar negociar um pouco

(Porque todo mundo tem que ceder ocasionalmente)

O apocalipse não tentará negociar de volta.

E, ah, sim, piora.

Como Eliezer_{1997} enfrentou o argumento óbvio de que não seria possível derivar um "dever" da lógica pura, já que declarações de "dever" só poderiam ser derivadas de outras declarações de "dever"?

Bem (observou Eliezer_{1997}), este problema se assemelha à estrutura do argumento de que uma causa só procede de outra causa, ou que uma coisa real só pode surgir de outra coisa real, o que levaria à conclusão de que nada existe.

Assim, como ele destacou, enfrentamos três "problemas difíceis": o problema difícil da experiência consciente, no qual percebemos que os qualia não podem surgir de processos computáveis; o problema difícil da existência, que questiona como qualquer existência aparentemente surge do nada; e o problema difícil da moralidade, que envolve chegar a um "dever".

Esses problemas estão provavelmente interligados. Por exemplo, os qualia do prazer são fortes candidatos a algo intrinsecamente desejável. Portanto, talvez não possamos compreender o problema difícil da moralidade sem desvendar o problema difícil da consciência. É evidente que esses problemas são demasiadamente difíceis para os humanos resolverem - caso contrário, alguém os teria solucionado nos últimos 2.500 anos desde a invenção da filosofia.

Não é como se esses problemas pudessem ter soluções complicadas - são simples demais para isso. O problema deve estar fora do espaço-conceito humano. Ao percebermos que a consciência não pode surgir de nenhum processo computável, podemos concluir que ela deve envolver uma nova física - uma física que nosso cérebro utiliza, mas que não conseguimos compreender completamente. É por isso que precisamos da superinteligência para solucionar esse problema. Possivelmente, isso está relacionado à mecânica quântica, talvez com uma pitada de pequenas curvas fechadas semelhantes ao tempo, inspiradas na Relatividade Geral; os paradoxos temporais podem possuir algumas das mesmas propriedades de irredutibilidade que a consciência parece exigir...

E assim por diante, até que você possa começar a perceber, ao longo das minhas postagens de *Overcoming Bias*, a carta que eu gostaria de ter escrito para mim mesmo.

Dessa experiência, aprendo uma lição: você não pode manipular a confusão. Não é possível criar planos inteligentes para contornar as lacunas em seu entendimento. Nem mesmo é possível fazer "melhores suposições" sobre coisas que, fundamentalmente, o confundem e relacioná-las a outras coisas igualmente confusas. Bem, pode-se tentar, mas não terá sucesso até que a confusão se dissolva. A confusão reside na mente, não na realidade, e tentar tratá-la como algo tangível só resultará em comédia involuntária.

Da mesma forma, não é possível apresentar razões inteligentes pelas quais as lacunas em seu modelo não importam. Não é possível traçar uma fronteira em torno do mistério, colocar alças elegantes que permitam usar a Coisa Misteriosa sem realmente entendê-la - como minha tentativa de fazer com que a possibilidade de a vida não ter sentido seja anulada por uma fórmula de utilidade esperada. Não se pode agarrar a lacuna e manipulá-la.

Se o espaço em branco no seu mapa esconde uma mina terrestre, então colocar o seu peso nesse ponto será fatal, não importa quão boa seja a sua desculpa para não saber. Qualquer caixa preta pode conter uma armadilha, e não há como saber a não ser abrindo a caixa preta e olhando dentro. Se você apresentar alguma justificativa para explicar por que precisa prosseguir com o melhor entendimento que possui, a armadilha dispara.

Somente quando você conhece as regras,

Que você percebe por que precisava aprender;

O que teria acontecido de outra forma,

O quanto você precisava saber.

Somente o conhecimento pode prever o custo da ignorância. Os antigos alquimistas não tinham uma maneira lógica de entender as razões exatas pelas quais era difícil para eles transformar chumbo em ouro. Então, eles se envenenaram e morreram. A natureza não se importa.

Mas chegou um momento em que a compreensão começou a surgir em mim.

Referências

[1] Edward Elmer Smith, Second Stage Lensmen (Old Earth Books, 1998).

297 — Aquela pequena nota de discórdia

Quando nos despedimos de Eliezer_{1997} pela última vez, ele acreditava que qualquer superinteligência agiria automaticamente de forma “correta” e, de fato, compreenderia isso melhor do que nós - embora, modestamente confessasse, não compreendesse a natureza última da moralidade. Ou melhor, após algum debate, Eliezer_{1997} desenvolveu um argumento elaborado, que ele chamava carinhosamente de “formal”, argumentando que poderíamos condicionar a crença de que a vida tem sentido; assim, os casos em que as superinteligências não se sentissem compelidas a fazer algo em particular seriam desconsiderados. (A falha está na equação não considerada e injustificada entre “argumento universalmente convincente” e “correto”).

Até então, o jovem Eliezer está trilhando o caminho para se juntar ao clube das “pessoas inteligentes que são estúpidas por serem habilidosas em defender crenças adotadas sem razões qualificadas”. Toda sua dedicação à “racionalidade” não o salvou desse erro, e pode-se ser tentado a concluir que é inútil lutar pela racionalidade.

Mas, embora muitos cavem seus próprios buracos, nem todos conseguem sair.

Com isso, aprendo minha lição: tudo começou—

—com uma pequena, insignificante pergunta; uma única nota discordante; um pensamento minúsculo e solitário...

À medida que nossa história avança, pulamos três anos até Eliezer_{2000}, que, em muitos aspectos, se assemelha ao que era em 1997. Atualmente, ele acredita ter provado que construir uma superinteligência é a coisa certa a fazer, se é que existe algo certo. Daí resulta não haver conflito de interesses justificável sobre a explosão de inteligência entre os povos e as pessoas da Terra.

Essa é uma conclusão significativa para Eliezer_{2000}, pois ele considera a ideia de lutar pela explosão da inteligência insuportavelmente estúpida. (Mais ou menos como a noção de Deus intervindo em lutas entre tribos de bárbaros briguentos, só que ao contrário.) O autoconceito de Eliezer_{2000} não permite que ele - ele nem mesmo quer - dar de ombros e dizer: «Bem, meu lado chegou primeiro, então vamos aproveitar a banana antes que alguém a pegue.” É um pensamento muito doloroso de se considerar.

E ainda assim, a ideia surge para ele:

Talvez algumas pessoas prefiram que uma IA faça coisas específicas, como não matá-las, mesmo que a vida não tenha sentido?

O pensamento imediatamente seguinte é o óbvio, dada suas premissas:

Caso a vida não tenha sentido, nada é a coisa “certa” a fazer; portanto, não seria particularmente correto respeitar as preferências das pessoas nesse caso.

Esta é a esquiva óbvia. O problema reside no fato de que Eliezer_{2000} não se enxerga como um vilão. Ele não sai por aí proclamando: ‹De quais balas devo desviar hoje?› Ele se vê como um racionalista zeloso que segue tenazmente linhas de investigação. Posteriormente, ao refletir, perceberá muitas questões que sua mente, de alguma forma, conseguiu não acompanhar – mas este não é o seu autoconceito atual.

Portanto, Eliezer_{2000} não apenas captura o óbvio. Ele continua a ponderar.

Contudo, se as pessoas acreditam ter preferências no caso de a vida não ter sentido, então têm razões para contestar meu projeto de explosão de inteligência e seguir um projeto que respeite seus desejos no caso de que a vida não tenha sentido. Isso gera um conflito de interesses atual em relação à explosão da inteligência, impedindo que as coisas certas sejam feitas no evento principal para que a vida tenha sentido.

Agora, há várias justificativas que Elieze_{r2000} poderia ter utilizado para negligenciar este problema. Eu sei, pois já ouvi diversas desculpas para descartar a IA amigável. 'O problema é muito difícil de resolver' é uma afirmação que recebo de aspirantes a AGI que se imaginam inteligentes o suficiente para criar uma verdadeira Inteligência Artificial, mas não suficientemente para resolver um problema realmente difícil, como a IA Amigável. Ou 'preocupar-se com esta possibilidade seria um mau uso de recursos, dada a urgência incrível de criar IA antes que a humanidade se extinga – você tem que seguir com o que tem', sendo dito por pessoas que, basicamente, não se interessam pelo problema.

Entretanto, Eliezer_{2000} é um perfeccionista. Ele não é perfeito, obviamente, e não atribui tanta importância quanto eu à virtude da precisão, mas certamente é um perfeccionista. A ideia de metaética que Eliezer_{2000} defende, na qual as superinteligências conhecem o que é certo, melhor do que nós, anteriormente parecia envolver todos os problemas de justiça e moralidade em um invólucro hermético.

A nova objeção parece abrir um pequeno buraco na embalagem hermética. Vale a pena corrigir isso. Se você tem algo que é perfeito, realmente deixará que uma pequena possibilidade o comprometa?

Portanto, Eliezer_{2000} não deseja simplesmente abandonar o assunto; ele quer corrigir o problema e restaurar a perfeição. Como ele pode justificar gastar tempo? Refletindo sobre pensamentos como:

E Brian Atkins? [Brian Atkins sendo o financiador inicial do Machine Intelligence Research Institute, então chamado de Singularity Institute.] Ele provavelmente preferiria não morrer, mesmo que a vida não tivesse sentido. Ele está financiando o MIRI agora; não quero manchar a ética de nossa cooperação.

O sentimento de Elieze_{r2000} não se traduz muito bem – o inglês não tem uma descrição simples para ele, ou para qualquer outra cultura que conheço. Talvez a passagem do Antigo Testamento: "Não ferverás um cabrito no leite de sua mãe". Alguém que o ajuda por altruísmo não deveria se arrepender de tê-lo ajudado; você deve a eles, não tanto fidelidade, mas sim, que eles estão realmente fazendo o que pensam que estão fazendo ao ajudá-lo.

Bem, mas como Brian Atkins descobriria se eu não contasse a ele? Eliezer_{2000} nem pensa isso a não ser entre aspas, como o pensamento óbvio que um vilão teria na mesma situação. E Eliezer_{2000} também tem um contra-pensamento padrão pronto, uma proteção contra as tentações da desonestidade – um argumento que justifica a honestidade em termos de utilidade esperada, e não apenas um amor pessoal pela virtude:

Os seres humanos não são enganadores perfeitos; é provável que eu seja descoberto. E se detectores de mentiras genuínos forem inventados antes da Singularidade, em algum momento nos próximos trinta anos? Eu seria incapaz de passar em um teste de detector de mentiras.

Eliezer_{2000} segue a regra de que você deve estar sempre pronto para ter seus pensamentos transmitidos para o mundo inteiro a qualquer momento, sem constrangimentos. Caso contrário, claramente, você caiu em desgraça: ou você está pensando algo que não deveria estar pensando, ou está envergonhado por algo que não deveria envergonhá-lo.

(Hoje em dia, não defendo um ponto de vista tão extremo, principalmente por razões de Teoria da Diversão. Vejo um papel para a competição social contínua entre formas de vida inteligentes, pelo menos no que diz respeito à minha visão de curto prazo. Admito, hoje em dia, que pode ser correto para os seres humanos terem um eu; como disse John McCarthy: "Se todos vivessem para os outros o tempo todo, a vida seria como uma procissão de formigas seguindo umas às outras em um círculo". Se você tem um eu, você pode muito bem ter segredos, e talvez até conspirações. Mas eu ainda tento respeitar o princípio de ser capaz de passar em um futuro teste de detector de mentiras, com qualquer outra pessoa que esteja também disposta a passar pelo detector de mentiras, se o assunto for profissional. A Teoria da Diversão precisa de uma exceção de bom senso para o gerenciamento global de riscos catastróficos.)

Mesmo considerando a honestidade como garantida, existem outras desculpas que Eliezer_{2000} pode-

ria usar para jogar a questão no vaso sanitário. “O mundo não tem tempo” ou “É insolúvel” ainda funcionariam. Mas Eliezer$_{2000}$ não sabe que este problema, o problema da moralidade “reserva”, será particularmente difícil ou demorado. Ele só agora pensou em toda a questão.

E assim Eliezer$_{2000}$ começa a considerar realmente a questão: supondo que “a vida não tem sentido” (que as superinteligências não produzem as suas próprias motivações a partir da lógica pura), então como você especificaria uma moralidade alternativa? Sintetizando-a, incorporando-a na IA?

Há muita coisa que Eliezer$_{2000}$ não sabe neste momento. No entanto, ele dedica os últimos três anos a aprimorar a inteligência artificial e é um Racionalista Tradicional há ainda mais tempo. Existem técnicas de racionalidade que ele praticou e salvaguardas metodológicas que ele já desenvolveu. Ele compreende que não se deve pensar que tudo o que uma IA precisa é de Um Grande Princípio Moral. Elieze$_{r2000}$ já entende que é mais sensato abordar a questão de maneira tecnológica em vez de política. Ele está ciente do ditado que afirma que os programadores de IA devem pensar em código, utilizando conceitos que podem ser incorporados a um computador. Eliezer$_{2000}$ já possui a noção de algo chamado ‹pensamento técnico›, embora ainda não tenha formulado uma visão bayesiana sobre isso. Há muito tempo, ele percebeu que tokens LISP com nomes sugestivos não têm significado, entre outras coisas. Essas precauções impedem que ele caia em algumas das armadilhas iniciais, as quais observou consumirem outros novatos em seus primeiros passos no problema da IA Amigável... mesmo que tecnicamente este seja meu segundo passo; realmente falhei no primeiro.

Mas no final, o que importa é o seguinte: pela primeira vez, Eliezer$_{2000}$ está tentando abordar tecnicamente a incorporação da moralidade em uma IA, sem a saída de emergência da misteriosa essência da retidão.

Isso é o que realmente importa, no final. Sua filosofia anterior não foi suficiente para forçar seu cérebro a lidar com os detalhes. Este novo padrão é rigoroso o bastante para exigir trabalho real. A moralidade está gradualmente se tornando menos misteriosa para ele – Eliezer$_{2000}$ está começando a pensar dentro da caixa preta.

Suas razões para seguir esse curso de ação não têm importância alguma.

Há uma lição em ser perfeccionista. Há uma lição na parte sobre como Eliezer$_{2000}$ inicialmente pensou que isso era uma pequena falha e poderia tê-la descartado se esse fosse seu impulso.

No final das contas, a cadeia de causa e efeito é assim: Eliezer$_{2000}$ investigou mais detalhadamente, portanto melhorou com a prática. As ações desativam as justificativas. Se acontecer de seus argumentos justificarem não resolver as coisas em detalhes, como aconteceu com Eliezer$_{1996}$, então você não será bom em pensar sobre o problema. Se seus argumentos exigirem que você resolva as coisas detalhadamente, você terá a oportunidade de começar a acumular experiência.

No final das contas, essa era a única escolha que importava — e não as razões para fazê-la.

Digo tudo isso, como você pode imaginar, devido aos aspirantes a IA que encontro às vezes e que têm suas próprias razões inteligentes para não pensar no problema da IA Amigável. Nossas razões inteligentes para fazer o que fazemos tendem a ter muito menos importância para a Natureza do que para nós mesmos e nossos amigos. Se suas ações não parecem boas quando são despojadas de todas as suas justificativas e apresentadas como meros fatos brutos... então talvez você deva reexaminá-las.

Um esforço diligente nem sempre salvará uma pessoa. Existe algo chamado falta de habilidade. Mesmo assim, se você não tentar, ou não se esforçar o suficiente, você não terá a chance de se sentar em uma mesa de apostas altas – muito menos a habilidade ante. Isso é causa e efeito para você.

Além disso, o perfeccionismo realmente importa. O fim do mundo nem sempre vem com trombetas e trovões e a mais alta prioridade na sua caixa de entrada. Às vezes, a verdade devastadora primeiro se apresenta a você como uma pequena pergunta; uma única nota discordante; um pequeno pensamento solitário, que você poderia descartar com um toque fácil e sem esforço...

… e assim, ao longo dos anos seguintes, a compreensão começa a surgir lentamente naquele Eliezer do passado. Aquele Sol nasceu mais devagar do que poderia ter nascido.

298 — Lutando contra uma ação de retaguarda contra a verdade

Quando nos despedimos do Eliezer$_{2000}$ pela última vez, ele estava apenas começando a investigar como incorporar uma moralidade em uma IA. Suas razões para fazê-lo não importam realmente, exceto enquanto historicamente demonstram a importância do perfeccionismo. Se você praticar algo, pode melhorar; se investigar algo, pode descobrir; a única coisa que realmente importa é que o Eliezer$_{2000}$ está, de fato, concentrando suas energias em tempo integral para refletir tecnicamente sobre a moralidade da IA, em vez de, como fazia anteriormente, encontrar justificativas para não dedicar seu tempo dessa maneira. No final, é isso que importa.

Contudo, quando nossa história começa, à medida que o céu se torna cinzento e a ponta do sol surge no horizonte, o Eliezer$_{2001}$ ainda não reconheceu que o Eliezer$_{1997}$ estava equivocado em qualquer sentido significativo. Ele está aprimorando a estratégia de Eliezer$_{1997}$ ao incluir um plano de contingência para "o evento improvável de que a vida se revele sem sentido"...

… o que significa que o Eliezer$_{2001}$ agora possui uma saída para se afastar de seu erro.

Não estou dizendo apenas que o Eliezer$_{2001}$ pode afirmar "IA amigável é um plano de contingência" em vez de exclamar "Oops!"

Estou querendo dizer que o Eliezer$_{2001}$ agora possui um plano de contingência. Se ele começar a duvidar de sua metaética de 1997, a explosão da inteligência tem uma estratégia de backup, especificamente a IA Amigável. O Eliezer$_{2001}$ pode questionar sua metaética sem que isso sinalize o fim do mundo.

Além disso, seu gradiente foi suavizado; ele pode admitir 10% de chance de ter errado anteriormente, depois 20% de chance. Ele não precisa reconhecer todo o seu erro de uma só vez.

Se você acha que o Eliezer$_{2001}$ está sendo muito hesitante, concordo plenamente.

As estratégias de Eliezer 1996–2000 foram formadas na total ausência da consideração de "IA Amigável". A ideia era obter uma superinteligência, qualquer superinteligência, o mais rápido possível – sopa de codelet, heurística ad-hoc, programação evolucionária, código aberto, qualquer coisa que parecesse funcionar – de preferência todas as abordagens simultaneamente em um Projeto Manhattan. ("Todos os pais fazem coisas que dizem aos filhos para não fazerem. É assim que sabem dizer-lhes para não fazerem." [1]) Não é como se adicionar mais uma abordagem pudesse fazer mal.

Suas atitudes em relação ao progresso tecnológico foram formadas – ou mais precisamente, preservadas da tecnofilia absorvida durante a infância – em torno da suposição de que todo/qualquer movimento em direção à superinteligência é um bem puro, sem qualquer indício de perigo.

Ao olhar para trás, o que o Eliezer$_{2001}$ precisava fazer neste momento era declarar um evento *HMC* – *Halt, Melt and Catch Fire* (Pare, Derreta e Pegue Fogo). Um dos pressupostos fundamentais sobre os quais o resto foi construído revelou-se falho. Isso exige uma pausa mental até o ponto final: retirar o peso de todas as crenças construídas sobre suposições erradas, fazer o possível para repensar tudo do zero. Esta é uma arte sobre a qual preciso escrever mais – é semelhante ao esforço convulsivo necessário para limpar seriamente a casa, depois que um adulto religioso percebe pela primeira vez que Deus não existe.

Mas o que Eliezer$_{2001}$ realmente realizou foi ensaiar seus argumentos tecnofílicos anteriores sobre a dificuldade de proibir ou controlar novas tecnologias pelo governo – argumentos padrão contra a "renúncia".

A meu ver atual, parece que todas aquelas terríveis consequências que os tecnófilos argumentam resultarem de vários tipos de regulamentação governamental estão mais ou menos corretas – é muito mais fácil apontar o que alguém está fazendo errado do que dizer o que é certo. Meu ponto de vista atual não mudou para discordar dos tecnófilos sobre as desvantagens da tecnofobia; no entanto, tendo a ser mais solidário com o que os tecnofóbicos dizem sobre as desvantagens da tecnofilia. O que os Eliezers anteriores mencionaram sobre as dificuldades, por exemplo, do governo agir de maneira sensata em relação à IA Amigável ainda parece bastante verdadeiro. Acontece que muitas de suas esperanças para a ciência, a indústria privada, etc., parecem agora igualmente equivocadas.

Não vamos nos aprofundar no ponto de vista tecnovolátil. Eliezer_{2001} simplesmente jogou pela janela uma suposição fundamental importante – que a IA não pode ser perigosa, ao contrário de outras tecnologias. Intuitivamente, você suspeitaria que isso deveria ter algum grande efeito em sua estratégia.

Bem, Eliezer_{2001} pelo menos abandonou sua ideia de 1999 de um Projeto Manhattan de IA de código aberto usando sopa heurística auto-modificável, mas em geral...

No geral, ele queria atacar com armas em punho, utilizando imediatamente sua melhor ideia na época; e depois, ele ainda queria atacar com armas em punho. Ele não disse: 'Não sei como fazer isso'. Ele não disse: 'Preciso de um conhecimento melhor'. Ele não disse: 'Este projeto ainda não está pronto para começar a codificação'. Ainda era tudo: "O tempo está se esgotando, preciso agir agora! O MIRI começará a programar assim que tiver dinheiro suficiente!"

Anteriormente, ele queria concentrar o máximo de esforço científico possível no compartilhamento total de informações, e mesmo depois disso, ele continuava pensando nesses termos. Sigilo científico = vilão, abertura = herói. (Eliezer_{2001} não tinha lido sobre o Projeto Manhattan e não estava familiarizado com a discussão semelhante que Leó Szilárd teve com Enrico Fermi.)

Esse é o problema ao converter um grande 'Oops!' em um gradiente de probabilidade variável. Isso significa que não existe um único divisor de águas – um enorme impacto visível – que indicaria que mudanças igualmente enormes podem ser necessárias.

Em vez disso, ocorrem todas essas pequenas mudanças de opinião... que dão a ele a chance de ajustar os argumentos de suas estratégias; mudar um pouco a justificativa, mas manter a 'ideia básica' no lugar. Pequenos solavancos que o sistema consegue absorver sem se romper, pois a cada vez ele tem a chance de voltar e se corrigir. Acontece que no domínio da racionalidade, ruptura = positivo, correção = negativo. Na arte da racionalidade, é muito mais eficiente admitir um grande erro do que admitir muitos pequenos erros.

Acredito que existe algum tipo de instinto que os humanos têm de preservar suas estratégias e planos anteriores, para não ficarem constantemente se debatendo e desperdiçando recursos; e, claro, um instinto para preservar qualquer posição que tenhamos defendido publicamente, para não sofrermos a humilhação de estarmos errados. E embora o jovem Eliezer tenha lutado pela racionalidade durante muitos anos, ele não está imune a esses impulsos; eles exercem influências suaves sobre seus pensamentos, e isso, infelizmente, é um dano mais que suficiente.

Mesmo em 2002, o Eliezer anterior ainda não tinha certeza de que o plano de Eliezer_{1997} não poderia ter funcionado. Pode ter dado certo. Você nunca sabe, certo?

Mas chegou um momento em que tudo desabou.

Referências

[1] John Moore, Slay and Rescue (Xlibris Corp, 2000).

299 — Meu despertar naturalista

No episódio anterior, Eliezer$_{2001}$ engaja-se em uma manobra defensiva contra a verdade. Alterando gradativamente suas convicções, admitindo uma probabilidade crescente em um cenário diferente, mas nunca declarando abertamente: “Eu estava equivocado antes”. Ele ajusta suas estratégias conforme são desafiadas, encontrando novas justificativas para o mesmo plano que buscava anteriormente.

(Daí a advertência: “Cuidado para não realizar uma retirada defensiva diante das evidências, cedendo relutantemente a cada centímetro de terreno apenas quando forçado, sentindo-se ludibriado. Entregue-se à verdade tão prontamente quanto possível. Faça isso no momento em que perceber o que está resistindo; no momento em que puder discernir de que lado os ventos da evidência estão soprando contra você.”)

A memória se desvanece, e mal consigo suportar olhar para trás, para aqueles tempos - sério, não suporto revisar meus escritos antigos. Já fui corrigido uma vez em minhas recordações por alguém que estava presente. Assim, embora me recorde dos eventos cruciais, não tenho certeza da ordem em que ocorreram, muito menos do ano.

No entanto, se eu tivesse que eleger um momento em que minha insanidade cedeu, seria quando compreendi, de maneira abrangente, a noção de um processo de otimização. Foi nesse instante que, ao olhar para trás pela primeira vez, declarei: “Fui um tolo”.

Anteriormente, em 2002, escrevi sobre a psicologia evolutiva da inteligência humana - embora naquela época acreditasse que estava abordando a IA; àquela altura, pensava estar contra a inteligência antropomórfica, mas ainda buscava inspiração no cérebro humano. (O artigo em questão é “Níveis de Organização na Inteligência Geral”, um capítulo solicitado para o volume Inteligência Artificial Geral, [1] que foi finalmente publicado em 2007.)

Assim, eu contemplava (e escrevia) sobre como a seleção natural havia liberado a inteligência humana de maneira natural; vislumbrei uma dicotomia entre elas, a cegueira da seleção natural, e a antecipação da previsão inteligente, o raciocínio por simulação contra a representação de tudo na realidade, o pensamento abstrato contra o concreto. Entretanto, foi a seleção natural que deu origem à inteligência humana, de modo que nossos cérebros, embora não nossos pensamentos, são inteiramente moldados pela assinatura da seleção natural.

Até hoje, essa visão ainda me parece extraordinariamente impactante, e por isso me intriga quando as pessoas rotulam a seleção natural e os processos orientados pela inteligência como “evolutivos”. Na verdade, eles são praticamente distintos em vários aspectos cruciais - embora compartilhem conceitos que podem ser empregados para descrevê-los, como consequencialismo e generalidade entre domínios.

Porém, o fato de Eliezer$_{2002}$ pensar em termos de uma dicotomia entre evolução e inteligência revela algo sobre as limitações de sua perspectiva - assim como alguém que enxerga a política como uma dicotomia entre posturas conservadoras e liberais, ou alguém que vê as frutas como uma dicotomia entre maçãs e morangos.

Após a publicação online do esboço de “Níveis de Organização”, Emil Gilliam observou que minha visão da IA assemelhava-se bastante à minha visão da inteligência. Claro, Eliezer$_{2002}$ não advoga pela construção de uma IA à imagem da mente humana; Eliezer$_{2002}$ entende que a mente humana é apenas uma solução improvisada da seleção natural. No entanto, Eliezer$_{2002}$ descreveu esses níveis de organização no pensamento

humano e não propôs a implementação de diferentes níveis de organização na IA. Emil Gilliam pergunta se acho que estou me aproximando demais da linha humana. Chamo a alternativa de "Design de Mente Totalmente Alienígena" e respondo que um DMTA é provavelmente muito difícil para engenheiros humanos criarem, mesmo que seja teoricamente possível, porque não seríamos capazes de compreender algo tão alienígena enquanto o estivéssemos montando.

Não sei se Eliezer$_{2002}$ inventou essa resposta por conta própria ou se a encontrou em outro lugar. Desnecessário dizer, que ouvi essa desculpa muitas vezes desde então. Na realidade, aquilo que você compreende de verdadeiramente pode ser reconfigurado de diversas maneiras, preservando alguma essência estrutural. Entretanto, quando não se entende o voo, presume-se que uma máquina voadora necessite de penas, pois é difícil se desvincular da analogia com um pássaro.

Assim, Eliezer$_{2002}$ continua, em certo sentido, ligado a designs mentais humanistas. Ele busca aprimorá-los, mas a arquitetura humana continua sendo, de certa forma, seu ponto de partida.

O que rompe finalmente com essa ligação?

É uma confissão embaraçosa: originou-se de uma história de ficção científica que eu estava tentando escrever. (Não, você não pode ver; não está concluída.) A trama envolvia um processo de otimização não-cognitivo e não-evolutivo, algo semelhante a uma Bomba de Resultados. Não era inteligência, mas um efeito físico intertemporal – pelo menos, eu o imaginava como tal – que restringia estreitamente o espaço de resultados possíveis. (Não posso revelar mais do que isso; seria um spoiler, caso eu um dia terminasse a história. Apenas consulte o ensaio sobre Bombas de Resultados.) Era «apenas uma história», então eu estava livre para explorar a ideia e desenvolvê-la logicamente: C foi compelido a acontecer, portanto B (no passado) foi compelido a acontecer, portanto A (que levou a B) foi compelido a acontecer.

Traçar uma linha através de um ponto é geralmente considerado perigoso. Dois pontos constituem uma dicotomia; você os imagina opostos reciprocamente. Mas quando você tem três pontos diferentes, é nesse momento que é preciso despertar e generalizar.

Agora eu tinha três pontos: inteligência humana, seleção natural e minha trama fictícia.

Foi nesse ponto que generalizei a ideia de um processo de otimização, de algo que comprime o futuro em uma região estreita do possível.

Isso pode parecer um ponto óbvio se você tem acompanhado *Overcoming Bias* o tempo todo; mas se observarmos a coleção de 71 definições de inteligência de Shane Legg, perceberemos que "restringir o futuro a uma região limitada" é uma resposta menos evidente do que parece.

Muitas das definições de "inteligência" feitas por pesquisadores de IA falam sobre "resolver problemas" ou "atingir metas". Mas do ponto de vista dos Eliezers do passado, pelo menos, é apenas a retrospectiva que torna isso equivalente a "restringir o futuro".

Uma meta é um constructo mental; elétrons não têm objetivos nem resolvem problemas. Quando um ser humano concebe uma meta, imagina um agente imbuído de desejo – ainda é uma linguagem empática.

Pode-se sustentar a ideia de que inteligência está relacionada a "alcançar objetivos" – e depois discutir se alguns "objetivos" são superiores a outros – ou falar sobre a sabedoria necessária para julgar entre os próprios objetivos – ou abordar um sistema modificando deliberadamente seus objetivos — ou discutir o livre-arbítrio necessário para escolher planos que alcancem objetivos — ou até mesmo sobre uma IA percebendo que seus objetivos não correspondem ao que os programadores realmente pretendiam. Se imaginarmos algo que comprima o futuro em uma estreita região do possível, como uma Bomba de Resultados, essas afirmações aparentemente sensatas, de alguma forma, não se aplicam.

Então, para mim, pelo menos, ver através da palavra "mente" um processo físico que, apenas por funcionar naturalmente, apenas por obedecer às leis da física, acabaria comprimindo seu futuro em uma região estreita, foi uma iluminação naturalista além e acima da noção de um agente tentando atingir seus objetivos.

Foi como sair de um poço profundo, cair no mundo comum, tensões cognitivas tensas relaxando em simplicidade não forçada, confusão transformando-se em fumaça e desaparecendo. Vi o trabalho realizado pela inteligência; "inteligente" não era mais uma propriedade, mas um motor. Como um nó no tempo, ecoando a parte externa do universo na parte interna e, assim, orientando-o. Até vi, num lampejo da mesma iluminação, que uma mente tinha de produzir calor residual para obedecer às leis da termodinâmica.

Em um momento anterior, Eliezer_{2001} mencionou a IA Amigável como algo que deveríamos adotar por segurança – se não tivéssemos certeza de que o design X da IA seria amigável, então deveríamos optar pelo design Y da IA, sabendo que seria amigável. No entanto, Eliezer_{2001} não tinha certeza se seria realmente possível criar uma superinteligência capaz de transformar seu cone de luz futuro em simples clipes de papel.

Agora, contudo, eu podia perceber: o pulsar do processo de otimização, informações sensoriais emergindo, instruções motoras surgindo, guiando o futuro. No centro, o modelo que conectava ações possíveis a resultados possíveis, e a função de utilidade associada a esses resultados. Introduza a função de utilidade adequada e o resultado seria um otimizador que direcionaria o futuro para qualquer direção.

Até aquele momento, nunca havia admitido para mim mesmo que o design do sistema de metas de IA de Eliezer_{1997} poderia, sem dúvida, extinguir inutilmente a espécie humana. Agora, porém, olhei para trás e finalmente pude entender o que meu antigo design realmente fazia, até o ponto em que era coeso o suficiente para ser discutido. Em termos gerais, teria convertido seu futuro cone de luz em ferramentas genéricas – computadores sem programas para operar, energia armazenada sem utilidade...

… como pude, eu, o racionalista refinado e experiente, como pude deixar passar algo tão óbvio por seis longos anos?

Foi nesse momento que acordei lúcido e lembrei-me; e pensei, com certo constrangimento: fui estúpido.

Referências

[1] Ben Goertzel and Cassio Pennachin, eds., Artificial General Intelligence, Cognitive Technologies (Berlin: Springer, 2007), doi:10.1007/978-3-540-68677-4.

300 — O nível acima do meu

Uma vez, emprestei meu exemplar de “Teoria da Probabilidade: A Lógica da Ciência” a Xiaoguang “Mike” Li. Mike Li leu um pouco e depois voltou, dizendo:

Uau... é como se Jaynes fosse um vampiro de mil anos.

Então, Mike disse: “Não, espere, deixe-me explicar isso… “ e eu disse: “Não, sei exatamente o que você quer dizer”. É uma convenção na literatura de fantasia que quanto mais velho um vampiro fica, mais poderoso ele se torna.

Eu já apreciava provas de matemática antes de conhecer Jaynes. No entanto, E. T. Jaynes foi a primeira vez que percebi uma sensação formidável em argumentos matemáticos. Talvez porque Jaynes estivesse alinhando “paradoxos” usados para se opor ao Bayesianismo e, em seguida, os destruindo com uma capacidade ofensiva esmagadora - poder sendo usado para superar outros. Ou talvez a sensação de formidável tenha vindo do fato de Jaynes não tratar sua matemática como um jogo estético; Jaynes se importava com a teoria das probabilidades, ela estava ligada a outras considerações que eram importantes para ele e para mim também.

Por alguma razão, a sensação que tenho de Jaynes é de uma perfeição assustadoramente rápida - algo que chegaria à resposta correta pelo caminho mais curto possível, destruindo todos os erros ao redor no mesmo movimento. É claro que, ao escrever um livro, você tem a chance de mostrar apenas o seu melhor lado. Mas ainda.

Foi bom para Mike Li o fato de ele ser capaz de sentir a aura formidável que cercava Jaynes. Observei que é uma regra geral que você não pode discriminar entre níveis muito acima do seu. Por exemplo, alguém uma vez me disse sinceramente que eu era muito inteligente e que “deveria ir para a faculdade”. Talvez qualquer coisa além de um desvio padrão acima de você comece a se confundir, embora isso seja apenas um palpite que parece legal.

Então, depois de ouvir Mike Li comparar Jaynes a um vampiro de mil anos, uma pergunta imediatamente surgiu em minha mente:

“Você sente a mesma coisa de mim?” Perguntei.

Mike balançou a cabeça. “Desculpe”, disse ele, parecendo um tanto estranho, “é só que Jaynes é...”

“Não, eu sei”, eu disse. Eu não pensei que tivesse alcançado o nível de Jaynes. Eu só estava curioso para saber como eu era visto por outras pessoas.

Almejo atingir o patamar de Jaynes. Busco me tornar um mestre em Inteligência Artificial/reflexividade, assim como Jaynes foi um mestre na teoria da probabilidade bayesiana. Poderia até afirmar que a arte que estou tentando dominar é mais desafiadora do que aquela de Jaynes, desconsiderando qualquer deferência. Mesmo assim, e de maneira constrangedora, não há nenhuma arte da qual eu seja mestre agora, como Jaynes era da teoria da probabilidade.

Isso não implica, necessariamente, que me coloco em um patamar inferior a Jaynes como pessoa – afirmando que Jaynes possuía uma aura mágica de destino, enquanto eu não tenho.

Ao contrário, reconheço em Jaynes um nível de especialização, uma pura formidabilidade, que ainda não atingi. Posso argumentar vigorosamente sobre o tema que escolhi, mas isso não equivale a formular as

equações e declarar: FEITO.

Enquanto ainda não alcancei esse patamar, devo reconhecer a possibilidade de nunca conseguir atingi-lo, de que meu talento inato possa não ser suficiente. Quando Marcello Herreshoff já me conhecia há tempo suficiente, perguntei-lhe se conhecia alguém que lhe parecesse substancialmente mais inteligente nativamente do que eu. Marcello pensou por um momento e disse: "John Conway - eu o conheci em um acampamento de matemática de verão". Droga, pensei, ele pensou em alguém e, pior, é um velho ultra-famoso que não consigo alcançar. Perguntei como Marcello chegou a esse julgamento. Marcello disse: "Ele me pareceu ter uma tremenda potência mental" e começou a explicar um problema de matemática que teve a oportunidade de resolver com Conway.

Não era o que eu queria ouvir.

Talvez, em relação à experiência que Marcello teve com Conway e à experiência que ele teve comigo, não tive a oportunidade de me destacar em nenhum assunto que dominei tão profundamente quanto Conway dominava seus muitos campos da matemática.

Ou pode ser que o cérebro de Conway seja especializado em uma direção diferente da minha e que eu nunca conseguiria me aproximar do nível de Conway em matemática, mas Conway não se sairia tão bem na pesquisa em IA.

Ou...

... Ou sou estritamente mais burro que Conway, dominado por ele em todas as dimensões. Talvez, se eu conseguisse encontrar um jovem proto-Conway e lhe contar o básico, eles passariam por mim, resolveriam os problemas que pesam sobre mim durante anos e fugiriam para lugares que não consigo acompanhar.

É prejudicial para o meu ego confessar essa última possibilidade? Sim. Seria fútil negar isso.

Será que realmente aceitei essa terrível possibilidade ou estou apenas fingindo que a aceitei? Aqui direi: "Não, acho que aceitei". Por que ousaria me dar tanto crédito? Porque investi um esforço específico nessa terrível possibilidade. Estou escrevendo aqui por vários motivos, mas o principal é a visão de uma mente mais jovem lendo essas palavras e passando por mim. Pode acontecer, pode não acontecer.

Ou, mais triste ainda: talvez eu tenha perdido muito tempo construindo recursos para me sustentar, em vez de estudar matemática em tempo integral durante toda a minha juventude; ou desperdicei muitos anos com ideias não matemáticas. E essa escolha, meu passado, é irrevogável. Atingirei uma barreira aos 40 anos e restará apenas repassar os recursos para outra mente com o potencial que desperdicei, ainda jovem o suficiente para aprender. Portanto, para economizar tempo, devo deixar um rastro dos meus sucessos e colocar alerta sobre meus erros.

Esses esforços específicos baseados em uma possibilidade prejudicial ao ego - esse é o único tipo de humildade que parece real o suficiente para eu me atrever a creditar. Ou abandonar minhas teorias preciosas quando percebi que não atendiam ao padrão que Jaynes havia me mostrado — isso foi difícil e real. Comportamentos modestos são baratos. Admissões humildes de dúvida são baratas. Conheço muitas pessoas que, ao enfrentar um contra-argumento, dizem: "Sou apenas um mortal falível, é claro que posso estar errado", e depois fazem exatamente o que planejaram fazer anteriormente.

Você notará que não tento modestamente dizer algo como: "Bem, posso não ser tão brilhante quanto Jaynes ou Conway, mas isso não significa que não possa realizar coisas importantes na área que escolhi".

Porque eu sei... não é assim que as coisas funcionam.

301 — A magnitude de sua própria loucura

Nos anos que antecederam meu encontro com um suposto criador de Inteligência Artificial Geral (com um projeto financiado), que por acaso era criacionista, eu costumava debater com aspirantes a IAG individualmente.

Naquela época, meio que consegui persuadir um desses indivíduos de que, sim, era crucial considerar a IA Amigável e, não, não era possível simplesmente encontrar a métrica de aptidão correta para um algoritmo evolutivo. (Anteriormente, ele estava bastante impressionado com algoritmos evolutivos.)

Ele então exclamou: Ah, caramba! Ah, que tolo fui! Por minha negligência, quase destruí o mundo! Que vilão eu já fui!

Entretanto, havia uma armadilha da qual eu sabia que não deveria cair...

– No ponto em que, no final de 2002, analisei as propostas de Eliezer$_{1997}$ sobre IA e compreendi o que elas teriam realmente feito, visto que eram suficientemente coerentes para que eu falasse sobre o que "teriam realmente feito".

Quando finalmente percebi a magnitude da minha própria insanidade, tudo se encaixou imediatamente. A barreira contra a compreensão se rompeu; e as dúvidas implícitas que estavam por trás disso desapareceram por completo. Não houve um período prolongado, nem mesmo um único momento que me lembre, em que eu questionasse como pude ser tão estúpido. Eu já sabia como.

E, de repente, no mesmo instante de compreensão, também percebi que dizer: Quase destruí o mundo!, seria excessivamente orgulhoso.

Seria uma confirmação excessiva do ego, uma validação demasiada da minha própria importância no esquema das coisas, em um momento em que - compreendi no mesmo instante em que percebi - meu ego deveria receber um golpe significativo. Eu havia sido tão menos do que deveria; eu precisava suportar aquele golpe no estômago, não o evitar.

Da mesma forma, não caí na armadilha de dizer: Ah, bem, não é como se eu tivesse o código e estivesse prestes a executá-lo; Eu realmente não cheguei perto de destruir o mundo. Pois isso também teria minimizado a intensidade do golpe. Não estava verdadeiramente carregado? Eu havia proposto e pretendido construir a arma, carregá-la, apontá-la para a cabeça e puxar o gatilho; e isso foi um pouco autodestrutivo demais.

Não transformei isso em um grande drama emocional. Isso teria desperdiçado a força do soco, transformando-o em meras lágrimas.

No mesmo instante, compreendi o que eu estava cuidadosamente deixando de fazer nos últimos seis anos. Eu não estava me atualizando.

E sabia que precisava, finalmente, fazer essa atualização. Mudar verdadeiramente o que planejei fazer, alterar o que estava executando agora, fazer algo diferente.

Tinha consciência de que era hora de parar.

Parar, derreter e pegar fogo.

Dizer: “Não estou pronto”. Dizer: “Ainda não sei como fazer isso”.

Essas são palavras incrivelmente difíceis de dizer, no campo da IAG. Tanto o público leigo quanto os colegas pesquisadores de IAG estão interessados em código, em projetos com programadores em ação. Caso contrário, podem até lhe conceder algum crédito por afirmar: “Estou pronto para escrever código; apenas me forneça o financiamento”.

Entretanto, ao dizer: “Não estou pronto para escrever código”, seu status cairá como um balão de urânio empobrecido.

O que o diferencia, então, de seis bilhões de outras pessoas que não sabem como criar Inteligência Artificial Geral? Se você não possui um código limpo (que não apenas simule inteligência humana, evidentemente; mas que pelo menos seja um código), ou, no mínimo, sua própria startup que desenvolverá o código assim que obtiver financiamento - então, quem é você e o que está fazendo em nossa conferência?

Pode ser que, mais tarde, eu escreva sobre a origem dessa atitude – o ponto médio excluído entre “Eu sei como construir IAG!” e “Estou trabalhando em IA restrita porque não sei como construir IAG”, a ausência de um conceito para “Estou tentando passar de um mapa incompleto de FAI para um mapa completo de FAI”.

Mas essa atitude existe e, portanto, a perda de status associada a dizer “Não estou pronto para escrever código” é muito grande. (Se alguém duvida disso, mencione o nome de qualquer outro que afirme simultaneamente: “Pretendo construir uma Inteligência Artificial Geral”, “Neste momento não posso construir uma IAG porque não conheço X” e “Estou atualmente tentando descobrir X.”)

(E não se preocupe com o pessoal de IAG que já levantou capital de risco, prometendo retornos em cinco anos.)

Portanto, há uma enorme relutância em dizer “Pare”. Não se pode simplesmente dizer: “Ah, vou voltar para o modo descobrir X”, porque esse modo não existe.

Havia mais do que simples perda de prestígio nessa minha hesitação? Eliezer_{2001} também poderia ter evitado retardar o aparente impulso em direção à explosão de inteligência, que era tão certa e necessária...

Mas, acima de tudo, creio que evitei admitir para mim mesmo: “Estou pronto para começar a programar.” Não apenas por receio das reações alheias, mas porque eu mesmo tenha sido inculcado com a mesma mentalidade.

Principalmente, Eliezer_{2001} não disse «Pare» – mesmo após perceber o problema da IA Amigável – porque eu não percebi, em um nível visceral, que a Natureza tinha o direito de me aniquilar.

“Os adolescentes pensam que são imortais”, diz o ditado. Claro, isso não é verdade de forma literal. Se perguntar a eles: “Você é indestrutível?”, responderão: “Sim, vá em frente e tente atirar em mim.” Contudo, talvez o uso de cintos de segurança não seja muito apelativo emocionalmente para eles, porque a ideia da própria morte não lhes parece muito real – não acreditam verdadeiramente que isso possa acontecer. Em princípio, talvez, mas não na prática.

Pessoalmente, sempre usei cinto de segurança. Como indivíduo, compreendi que eu podia morrer.

No entanto, tendo sido criado em tecnofilia para valorizar essa coisa mais preciosa, muito mais importante do que minha própria vida, pensava que o Futuro era indestrutível.

Mesmo quando reconheci que a nanotecnologia poderia exterminar a humanidade, ainda mantinha a convicção de que a explosão de inteligência era invulnerável. Se a humanidade sobrevivesse, a explosão de inteligência aconteceria e a IA resultante seria tão inteligente que seria incorruptível e inatingível.

Mesmo após reconhecer a importância da IA Amigável, não acreditava emocionalmente na possibilidade de falha, assim como um adolescente que não usa o cinto de segurança não acredita que um acidente automobilístico possa realmente matá-lo ou aleijá-lo.

Foi somente quando meu insight sobre otimização me permitiu olhar para trás e ver Eliezer1997 com clareza que percebi que a Natureza tinha a autorização para me matar.

"O pensamento que você não consegue conceber exerce mais controle sobre você do que as palavras que você expressa em voz alta." No entanto, nos afastamos apenas dos medos que são tangíveis para nós.

Os pesquisadores da IAG levam muito a sério a possibilidade de outra pessoa solucionar o problema primeiro. Conseguem imaginar ver manchetes nos jornais anunciando que seu próprio trabalho foi superado. Eles sabem que a Natureza pode fazer isso com eles. Aqueles que iniciaram empresas sabem que podem ficar sem financiamento. Essa possibilidade é real para eles, muito real; exerce uma influência emocional profunda sobre eles.

Não acredito que o "Ops" seguido pelo estrondo de seis bilhões de corpos caindo, por suas próprias ações, seja real para eles no mesmo nível.

Não é seguro dizer o que outras pessoas estão pensando. Mas parece bastante provável que, quando alguém reage à perspectiva da IA Amigável dizendo: "Se você atrasar o desenvolvimento para focar na segurança, outros projetos que não se preocupam nem um pouco com a IA Amigável o ultrapassarão", a perspectiva de eles próprios cometerem um erro seguido por seis bilhões de desastres não é verdadeiramente real para eles; mas a possibilidade de serem superados por outros é profundamente assustadora.

Eu também costumava expressar pensamentos assim, antes de compreender que a Natureza tinha permissão de me matar.

Naquele momento de compreensão, minha tecnofilia infantil finalmente se desfez.

Finalmente entendi que, mesmo que você siga diligentemente as regras da ciência e seja uma pessoa ética, a Natureza ainda pode aniquilá-lo. Finalmente compreendi que, mesmo sendo o melhor projeto dentre todos os candidatos disponíveis, a Natureza ainda pode te matar.

Percebi que não estava sendo graduado em uma curva. Meu olhar se desviou dos concorrentes e vislumbrou a parede completamente vazia.

Ao olhar para trás, vi os argumentos cuidadosos que havia construído, explicando por que a escolha mais sensata era prosseguir a toda velocidade, exatamente como havia planejado fazer antes. E então compreendi que, mesmo se construíssemos um argumento mostrando que algo era a melhor opção, a Natureza ainda poderia dizer: "E daí?" e extinguir você.

Olhei para trás e percebi que havia afirmado considerar o risco de um erro fundamental, apresentando razões para tolerar a possibilidade de avançar mesmo sem pleno conhecimento.

Compreendi que o risco que eu estava disposto a aceitar poderia ter me levado à morte. Vi que essa possibilidade nunca havia sido verdadeiramente real para mim. Entendi que, mesmo que houvesse argumentos sábios e excelentes para assumir tal risco, ele ainda poderia resultar em sua própria morte. Matar você de verdade.

Porque, no final, apenas a ação importa, não importam as razões para realizá-la. Se você construir a arma, carregá-la, apontá-la para a cabeça e puxar o gatilho, mesmo com os argumentos mais inteligentes para cada passo – então, bang.

Percebi que apenas minha própria ignorância das regras me permitiu argumentar a favor de avançar sem pleno conhecimento das regras; afinal, se você não conhece as regras, não pode ponderar sobre a penalidade da ignorância.

Notei que outros, também ignorantes das regras, diziam: "Vou em frente e faço X"; e, na medida em que X fosse uma proposta coerente, eu sabia que isso resultaria em um estrondo; mas eles diziam: "Não sei se não pode funcionar." Eu tentava explicar a eles a estreiteza do alvo no espaço de busca, e eles diziam: "Como podem ter tanta certeza de que não vou ganhar na loteria?", usando sua própria ignorância como uma justificativa.

Foi então que percebi que a única coisa que poderia ter feito para me salvar, no meu estado anterior de ignorância, era dizer: "Não avançarei até ter certeza de que o terreno é seguro." E há muitos argumentos inteligentes para explicar por que você deve pisar em um terreno que não sabe se contém uma mina ter-

restre; mas todos esses argumentos parecem muito menos inteligentes quando você olha para o lugar que propôs pisar e percebe o estrondo.

Entendi que você pode fazer tudo o que deve fazer, mas a Natureza ainda pode ter permissão para matá-lo. Foi quando minha última confiança se quebrou. E foi aí que começou minha formação como racionalista.

302 — Além do alcance de Deus

Este ensaio é um pouco mais sombrio do que o usual, já que avalio essas questões. Trata-se de um experimento mental que elaborei para desafiar meu próprio otimismo, após perceber que ele me iludira. Leitores que se identificam com argumentos como "É crucial mantermos nossos vieses, pois eles nos proporcionam felicidade" deveriam considerar a opção de não continuar a leitura. (A menos que tenham algo a proteger, inclusive suas próprias vidas.)

Refletindo sobre a extensão da minha própria insanidade, percebi que, em sua raiz, residia uma falta de fé na vulnerabilidade do Futuro - uma hesitação em aceitar que as coisas poderiam realmente dar errado. Não por meio de uma crença verbal explícita, mas como algo interno que persistia em acreditar, mesmo diante das adversidades, que no final tudo se resolveria.

Alguns poderiam considerar isso uma virtude (*zettai daijobu da yo*[4]), enquanto outros argumentariam que é algo essencial para a saúde mental.

No entanto, vivemos em um mundo diferente. Vivemos em um mundo fora do alcance de Deus.

Há muito, muito tempo que eu acreditava em Deus. Crescendo em uma família judaica ortodoxa, lembro-me da última vez que pedi algo a Deus, embora não me lembre quantos anos eu tinha. Estava fazendo um pedido em nome do garoto vizinho, esqueci exatamente o quê - algo como "Espero que tudo dê certo para ele" ou talvez "Espero que ele se torne judeu."

Recordo-me de como era ter uma autoridade superior a quem recorrer, para cuidar de coisas que eu não conseguia resolver sozinho. Não pensava nisso como algo "quente", pois não tinha alternativa para comparar. Eu simplesmente dava isso como certo.

Mesmo assim, lembro-me, embora apenas de uma infância distante, como é viver em um mundo conceitualmente possível onde Deus existe. Ele realmente existe, assim como as crianças e os racionalistas consideram todas as suas crenças pelo seu valor nominal.

No mundo onde Deus existe, Ele intervém para otimizar tudo? Independentemente do que os rabinos afirmem sobre a natureza fundamental da realidade, a resposta operacional para levar a sério a esta questão é obviamente "Não". Você não pode pedir a Deus que lhe traga uma limonada da geladeira em vez de pegar uma você mesmo. Quando acreditava em Deus com a seriedade de uma criança, há muito tempo atrás, não acreditava nisso.

Postular essa inação divina específica não causa uma crise teológica completa. Se você me disse-se: "Eu construí um usuário benevolente e superinteligente de nanotecnologia", e eu dissesse "Dê-me uma banana", e nenhuma banana aparecesse, isso ainda não refutaria sua afirmação. Pais humanos nem sempre fazem tudo o que os filhos pedem. Existem alguns argumentos decentes da teoria da diversão - eu até acredito neles - contra a ideia de que o melhor tipo de ajuda que você pode oferecer a alguém é sempre lhe dar imediatamente tudo o que ele deseja. Não creio que a eudaimonia seja formular metas e cumpri-las instan-

4 NT. **Zettai daijōbu da yo**: Expressão japonesa que significa *"Está tudo absolutamente bem!"* ou *"Não se preocupe!"*. Usada para transmitir segurança ou conforto, muitas vezes em contextos de anime/mangá ou conversas cotidianas. A estrutura enfática (*zettai* = "absolutamente"; *daijōbu* = "ok"; *da yo* = partícula assertiva) reflete tentativa de tranquilizar alguém, às vezes mascarando inseguranças.

taneamente; não quero me tornar uma simples coisa que deseja, que nunca precisa planejar, agir ou pensar.

Portanto, não é necessariamente uma tentativa de evitar a falsificação dizer que Deus não atende todas as orações. Mesmo uma IA Amigável pode não responder a todas as solicitações.

Mas é evidente que existe um limiar de horror suficientemente terrível para que Deus intervenha. Lembro-me de que isso era verdade quando acreditava com o olhar de uma criança.

O Deus que não intervém de forma alguma, não importa quão ruins as coisas fiquem - essa é uma tentativa óbvia de evitar a falsificação, de proteger uma crença na crença. Crianças suficientemente pequenas não possuem o conhecimento profundo de que Deus realmente não existe. Elas esperam genuinamente encontrar um dragão em sua garagem. Não há razão para elas conceberem um Deus amoroso que nunca age. Onde exatamente reside o limite para o horror aceitável? Até uma criança consegue imaginar uma discussão sobre esse limite preciso. Contudo, é claro que Deus estabelecerá um limite em algum lugar. Na verdade, são poucos os pais amorosos que, desejando que seu filho cresça forte e independente, permitiriam que ele fosse atropelado por um carro.

O exemplo evidente de um horror tão grande que Deus não pode tolerar é a morte - a verdadeira morte, a aniquilação da mente. Não acredito que nem mesmo o Budismo permita isso. Enquanto existir um Deus no sentido clássico - totalmente desenvolvido, ontologicamente fundamental, O Deus - podemos ter certeza de que nenhum evento suficientemente terrível ocorrerá. Não há alma em lugar algum que precise temer a verdadeira aniquilação; Deus impedirá isso.

E se você criar seu próprio universo simulado? O exemplo clássico de universo simulado é o Jogo da Vida de Conway. Recomendo que você explore o Jogo da Vida, caso nunca tenha jogado - é crucial para entender a noção de "lei física". O Jogo da Vida de Conway foi comprovado como Turing-completo, então seria possível criar um ser senciente dentro do universo da Vida, embora isso possa ser bastante frágil e peculiar. Outros autômatos celulares simplificariam isso.

Ao criar um universo simulado, você poderia escapar do alcance de Deus? Seria possível simular um Jogo da Vida contendo entidades sencientes e torturar os seres nele contidos? Mas se Deus está observando em todos os lugares, tentar construir uma Vida injusta resulta no Deus intervindo para modificar os transistores do seu computador. Se a física que você configurou em seu software demanda que uma entidade senciente da Vida seja torturada infinitamente sem razão específica, o Deus intervirá. Sendo Deus onipresente, não há refúgio em lugar algum para o verdadeiro horror. A vida é justa.

Mas suponha que, ao invés disso, você faça a pergunta:

Dadas tais e tais condições iniciais, e dadas tais e tais regras do autômato celular, qual seria o resultado matemático?

Nem mesmo Deus pode modificar a resposta a esta pergunta, a menos que você acredite que Deus pode implementar impossibilidades lógicas. Mesmo quando criança, não me lembro de acreditar nisso. (E por que você precisaria acreditar nisso, se Deus pode modificar qualquer coisa que realmente exista?)

Como seria a Vida neste mundo imaginário onde cada passo segue apenas seu antecessor imediato? Onde as coisas só acontecem, ou não, devido às regras do autômato celular? Onde as condições e regras iniciais não descrevem nenhum Deus que verifica cada estado? Como seria o mundo além do alcance de Deus?

Este mundo não seria justo. Se o estado inicial contiver as sementes de algo capaz de auto-replicação, a seleção natural poderá ou não ocorrer, a vida complexa pode ou não evoluir, e essa vida pode ou não se tornar senciente, sem Deus para guiar a evolução. Esse mundo pode evoluir para algo equivalente a vacas conscientes ou golfinhos conscientes, que não possuam mãos para melhorar sua condição; talvez sejam predados por lobos conscientes que nunca consideraram estar fazendo algo errado ou que se importavam.

Se, em uma vasta infinidade de mundos, algo semelhante aos humanos evoluir, então enfrentarão doenças – não para ensinar lições, mas porque os vírus também evoluíram sob as regras dos autômatos celulares.

Se as pessoas desse mundo são felizes ou infelizes, as causas de sua felicidade ou infelicidade podem

não ter nada a ver com as escolhas boas, ou más que fizeram. Nada a ver com livre arbítrio ou lições aprendidas. No mundo hipotético onde cada passo segue apenas as regras do autômato celular, o equivalente a Genghis Khan[5] pode assassinar um milhão de pessoas, rir, enriquecer-se e nunca ser punido, vivendo uma vida muito mais feliz do que a média. Quem impediria isso? Deus, é claro, impediria que isso acontecesse; Ele, no mínimo, lançaria alguma sombra de tristeza no coração do Khan. Mas na resposta matemática à pergunta "E se?", não há Deus nos axiomas. Portanto, se as regras do autômato celular dizem que o Khan está feliz, isso é simplesmente a resposta completa para a pergunta "E se?". Não há absolutamente nada que o impeça.

E se o Khan torturar horrivelmente pessoas até a morte ao longo dos dias, talvez para sua própria diversão? Elas pedirão ajuda, talvez imaginando um Deus. Se você realmente escrevesse aquele autômato celular, Deus interviria em seu programa, é claro. Mas na questão "E se?", o que o autômato celular faria sob as regras matemáticas, não há Deus no sistema. Dado que as leis físicas não contêm nenhuma especificação de uma função de utilidade – em particular, nenhuma proibição contra a tortura – então as vítimas só serão salvas se as células certas forem 0 ou 1. E não é provável que alguém desafie o Khan; se o fizessem, alguém os atacaria com uma espada, e a espada romperia seus órgãos, e eles morreriam, e isso seria o fim de tudo. Assim, as vítimas morrem, gritando, e ninguém as ajuda; essa é a resposta para a pergunta "E se".

As vítimas podem ser completamente inocentes? Por que não, no mundo hipotético? Se olharmos para as regras do Jogo da Vida de Conway (que é Turing-completo, então podemos incorporar física computável arbitrária nele), então as regras são realmente muito simples. Células com três vizinhos vivos permanecem vivas; células com dois vizinhos permanecem iguais; todas as outras células morrem. Não há nada aí sobre pessoas inocentes não serem terrivelmente torturadas por períodos indefinidos.

Este mundo está começando a parecer familiar?

A crença num universo justo muitas vezes se manifesta de maneiras mais sutis do que pensar que os horrores deveriam ser totalmente proibidos: Será que o século XX teria sido diferente se Klara Pölzl e Alois Hitler tivessem feito amor uma hora antes, e um espermatozoide diferente fertilizasse o óvulo, na noite em que Adolf Hitler foi concebido?

Diante de tantas vidas e perdas, atrelar tudo a um único evento parece desproporcional. O Plano Divino deveria ter mais lógica do que isso. É possível acreditar em um Plano Divino sem acreditar em Deus - Karl Marx certamente acreditava. Não deveríamos ter milhões de vidas dependendo de uma escolha casual, do tempo de uma hora, da velocidade de um flagelo microscópico. Isso não deveria ser permitido. É demasiadamente desproporcional. Portanto, se Adolf Hitler tivesse cursado o ensino médio e se tornado arquiteto, outra pessoa teria assumido seu papel, e a Segunda Guerra Mundial teria acontecido da mesma forma que ocorreu.

Mas no mundo fora do alcance de Deus, não há uma cláusula nos axiomas físicos que determine: "as coisas precisam fazer sentido" ou "grandes efeitos exigem grandes causas" ou "a história gira em torno de razões importantes demais para serem tão frágeis". Não há um Deus para impor essa ordem, que é tão severamente violada pelo fato de que as vidas e mortes de milhões de pessoas dependem de um pequeno evento molecular.

O objetivo do experimento mental é colocar lado a lado o universo de Deus e o universo da Natureza, para podermos reconhecer que tipo de pensamento pertence ao universo de Deus. Muitos ateus ainda acreditam que certas coisas não são permitidas. Eles apresentariam argumentos sobre por que a Segunda Guerra Mundial era inevitável e teria ocorrido mais ou menos da mesma maneira, mesmo que Hitler tivesse se tornado arquiteto. Mas, em uma análise histórica sóbria, essa é uma crença irracional. Escolhi o exemplo da Segunda Guerra Mundial porque, na minha leitura, os eventos foram principalmente motivados pela personalidade de Hitler, frequentemente desafiando seus generais e conselheiros. Não há uma justificativa empírica específica que eu tenha encontrado para duvidar disso. A principal razão para duvidar seria a recusa em aceitar que o universo poderia fazer tão pouco sentido - que coisas horríveis poderiam acontecer tão levianamente, por apenas um lance de dados.

---

5 NT. **Genghis Khan**: Fundador do *Império Mongol* (século XIII), nascido como *Temujin* (c. 1162–1227). Unificou tribos nômades da Ásia Central e liderou conquistas que estenderam seu território da China à Europa Oriental. Estrategista militar inovador, promoveu meritocracia e tolerância religiosa.

Mas por que não? O que proíbe isso?

No universo de Deus, Deus proíbe isso. Reconhecer isso é admitir que não vivemos nesse universo. Vivemos no universo hipotético, fora do alcance de Deus, regido pelas leis matemáticas e nada mais. Tudo o que a física diz que vai acontecer, acontecerá. Absolutamente qualquer coisa, boa ou ruim, acontecerá. E não há nada nas leis da física que anule essa regra, mesmo nos casos realmente extremos, onde poderíamos esperar que a Natureza fosse um pouco mais razoável.

Lendo "A Ascensão e Queda do Terceiro Reich", de William Shirer, e ouvindo-o descrever a incredulidade que ele e outros sentiram ao descobrir toda a extensão das atrocidades nazistas, pensei em como era estranho ler tudo isso e já saber que não havia uma única proteção contra isso. Simplesmente ler o livro inteiro e aceitá-lo; horrorizado, mas nada incrédulo, porque eu já havia compreendido em que tipo de mundo vivia.

Houve um tempo em que eu acreditava que a extinção da humanidade não era permitida. E outros que se autodenominam racionalistas podem ter suas próprias crenças nas quais confiam. Podem ser chamadas de "jogos de soma positiva", "democracia" ou "tecnologia", mas são sagradas. A marca dessa sacralidade é que aquilo em que confiam não pode levar a algo realmente prejudicial; ou não podem ser permanentemente desfiguradas, pelo menos não sem um lado positivo compensatório. Nesse sentido, são confiáveis, mesmo que coisas ruins ocorram ocasionalmente.

O desenrolar da história da Terra nunca deveria transitar de sua tendência de soma positiva para uma tendência de soma negativa; isso é inaceitável. As democracias – pelo menos as democracias liberais modernas – nunca legalizariam a tortura. A tecnologia tem contribuído de maneira tão significativa até agora que não deveria haver uma tecnologia do Cisne Negro capaz de quebrar essa tendência e causar mais danos do que todos os benefícios até agora.

Existem inúmeros argumentos inteligentes sobre por que tais coisas simplesmente não podem ocorrer. No entanto, a fonte desses argumentos é uma crença mais profunda de que tais eventos não são permitidos. Mas quem proíbe? Quem impede que isso aconteça? Se você não consegue visualizar pelo menos um universo legal onde a física permite que coisas terríveis aconteçam – e elas realmente ocorrem, sem possibilidade de apelar do veredicto – então você ainda não está preparado para discutir probabilidades.

Será realmente possível que seres sencientes tenham perecido completamente durante milhares ou milhões de anos, sem alma e sem vida após a morte – e não como parte de qualquer grande plano da Natureza – para não ensinarem alguma grande lição sobre o significado ou a falta de sentido da vida – nem mesmo para ensinarem qualquer lição profunda sobre o que é impossível – de modo que um truque tão simples e aparentemente estúpido como vitrificar pessoas em nitrogênio líquido possa salvá-las da aniquilação – e uma rejeição de 10 segundos da ideia boba possa destruir a alma de alguém? Será que um programador de computador que assina alguns papéis e compra uma apólice de seguro de vida continua existindo num futuro distante, enquanto Einstein apodrece em uma sepultura? Podemos ter certeza de uma coisa: Deus não permitiria isso. Qualquer coisa tão ridícula e desproporcional seria rejeitada. Seria uma zombaria do Plano Divino – uma afronta às sólidas razões pelas quais as coisas devem ser como são.

É possível ter racionalizações seculares para coisas que não são permitidas. Portanto, ajuda imaginar que exista um Deus, benevolente conforme entendemos a bondade – um Deus que impõe em toda a Realidade um mínimo de imparcialidade e justiça – cujos planos fazem sentido e dependem proporcionalmente das escolhas das pessoas – que nunca permitirá o horror absoluto – que não intervém sempre, mas que pelo menos proíbe universos completamente desviados de seu caminho... imaginar tudo isso, mas também imaginar que você mesmo vive em um mundo hipotético de matemática pura – um mundo além do alcance de Deus, um mundo totalmente desprotegido onde qualquer coisa pode acontecer.

Se ainda há algum leitor lendo isso que pensa que ser feliz é mais importante do que qualquer coisa na vida, então talvez não deva perder muito tempo refletindo sobre a desproteção de sua existência. Talvez pense nisso apenas o suficiente para inscrever você e sua família na criônica e/ou assinar um cheque para uma agência de mitigação de risco existencial ocasionalmente. E use cinto de segurança, obtenha seguro saúde e todas aquelas outras coisas necessárias e tristes que podem destruir sua vida se você perder aquele passo... mas fora isso, se você quer ser feliz, meditar sobre a fragilidade da vida não vai ajudar.

Mas este ensaio foi escrito para aqueles que têm algo a proteger.

O que pode fazer um camponês do século XII para se salvar da aniquilação? Nada. Os pequenos desafios da natureza nem sempre são justos. Quando confrontado com um desafio muito difícil, sofre-se uma penalidade; quando essa penalidade é letal, a morte é inevitável. Isso se aplica às pessoas e não difere para os planetas. Aqueles que desejam dançar a dança mortal com a Natureza precisam compreender contra o que estão lutando: neutralidade absoluta, sem exceções.

Mesmo tendo esse entendimento, nem sempre resultará em salvação. Não teria salvo um camponês do século XII, mesmo que ele o possuísse. Se você acredita que um racionalista, que compreende completamente a confusão em que se encontra, deve ser capaz de encontrar uma saída - então, você confia na racionalidade, já disse o suficiente.

Certamente, alguns comentaristas me repreenderão por atribuir um tom tão sombrio a tudo isso, e em resposta, listarão todas as razões pelas quais é encantador viver em um universo neutro. Afinal, é permitida à vida ser um pouco sombria, mas não mais escura do que um certo ponto, a menos que haja um lado positivo.

Ainda assim, para não criar desespero desnecessário, gostaria de expressar algumas palavras esperançosas neste momento:

Se o futuro da humanidade se desdobrar da maneira correta, podemos tornar nosso futuro cone de luz (mais) justo. Não podemos modificar a física fundamental, mas em níveis mais elevados de organização, poderíamos construir algumas grades de proteção e adicionar algum preenchimento; organizar as partículas em um padrão que faça verificações internas contra catástrofes. Há muitas coisas lá fora que não podemos tocar, mas pode ser útil considerar tudo o que não está em nosso cone de luz futuro como parte do 'passado generalizado', como se já tivesse ocorrido. Existe pelo menos a perspectiva de derrotar a neutralidade, no único futuro que podemos alcançar – o único mundo em que ela representa algo com que se preocupar.

Em algum momento, talvez, mentes imaturas serão confiavelmente protegidas. Mesmo que as crianças enfrentem algo equivalente a não ganhar um pirulito ou queimar um dedo, elas nunca serão atropeladas por carros.

E os adultos não correriam tanto perigo. Uma superinteligência – uma mente que poderia processar um bilhão de pensamentos sem cometer um erro – não se intimidaria diante de um desafio onde a morte é o preço de um único fracasso. O universo bruto não pareceria tão cruel; seria apenas mais um problema a ser resolvido.

O problema é que construir um adulto é, em si, um desafio adulto. Isso foi o que finalmente percebi, anos atrás.

Se existe um universo (mais) justo, temos que começar por este mundo – o mundo neutro, o mundo do concreto duro sem acolchoamento, o mundo onde os desafios não são calibrados de acordo com suas habilidades.

Nem toda criança precisa encarar a Natureza. Colocar o cinto de segurança ou preencher um cheque não é tão complicado, ou mortal. Não estou dizendo que todo racionalista deve refletir sobre a neutralidade, nem que todo racionalista deve ter todos esses pensamentos desagradáveis. Mas aquele que planeja enfrentar um desafio descalibrado de morte instantânea não deve evitá-los.

O que uma criança precisa fazer – que regras deve seguir, como deve se comportar – para resolver um problema de adulto?

303 — Minha iluminação Bayesiana

Lembro-me (ainda que vagamente, no contexto das memórias humanas) da primeira vez em que me identifiquei como 'bayesiano'. Alguém havia acabado de fazer uma versão distorcida de um antigo quebra-cabeça de probabilidades, dizendo:

Se eu encontrar uma matemática na rua e ela disser: 'Tenho dois filhos e pelo menos um deles é menino', qual é a probabilidade de ambos serem meninos?

Na versão correta desta história, o matemático diz: "Tenho dois filhos", e você pergunta: "Pelo menos um é menino?", e ela responde: "Sim". Então, a probabilidade é de 1/3 de que ambos sejam meninos.

Entretanto, na versão distorcida da história - como apontei - alguém raciocinaria de forma sensata:

Se o matemático tem um menino e uma menina, então minha probabilidade a priori de ela dizer 'pelo menos um deles é menino' é 1/2, e minha probabilidade a priori de ela dizer 'pelo menos um deles é menina' é 1/2. Não há razão para acreditar, a priori, que o matemático só mencionará uma menina se não houver alternativa possível.

Assim, indiquei isso e calculei a resposta usando a Regra de Bayes, chegando a uma probabilidade de 1/2 de que ambos os filhos fossem rapazes. Não tenho certeza se sabia ou não, neste momento, que a regra de Bayes se chamava assim, mas foi o que utilizei.

E eis que alguém me disse: "Bem, o que você acabou de dar é a resposta bayesiana, mas nas estatísticas ortodoxas a resposta é 1/3. Apenas excluímos as possibilidades descartadas e contamos as que restam, sem tentar adivinhar a probabilidade de o matemático dizer isto ou aquilo, uma vez que não temos como saber realmente essa probabilidade – é demasiado subjetivo."

Eu respondi – observe que isso foi completamente espontâneo – "O que diabos você quer dizer? Você não pode evitar atribuir uma probabilidade ao matemático que faz uma afirmação ou outra. Você está apenas assumindo que a probabilidade é 1, e isso é injustificado."

Ao que aquele respondeu: "Sim, é o que dizem os bayesianos. Mas os frequentistas não acreditam nisso."

E eu disse, espantado: "Como é possível existir uma estatística não bayesiana?"

Foi então que descobri que era do tipo chamado 'Bayesiano'. Pelo que sei, nasci assim. Minhas intuições matemáticas eram tais que tudo o que os bayesianos diziam parecia perfeitamente direto e simples, a maneira óbvia que eu mesmo faria; enquanto as coisas que os frequentistas diziam pareciam a blasfêmia elaborada, distorcida e louca do Cthulhu sonhador. Não escolhi me tornar um bayesiano, assim como os peixes não escolhem respirar na água.

Mas não é isso que chamo de minha 'iluminação Bayesiana'. A primeira vez que ouvi falar de 'bayesianismo', considerei-o óbvio; Não fui muito além da própria Regra de Bayes. Naquela época, eu ainda pensava na teoria da probabilidade mais como uma ferramenta do que como uma lei. Não pensei que existissem leis matemáticas de inteligência (meu melhor e pior erro). Como quase todos os aspirantes a AGI, Eliezer2001 pensou em termos de técnicas, métodos, algoritmos, construindo uma caixa de ferramentas cheia de coisas legais que poderia fazer; ele procurou por ferramentas, não por entendimento. A Regra de Bayes era uma

ferramenta realmente interessante, aplicável em um número surpreendente de casos.

Depois houve minha iniciação em heurísticas e vieses. Tudo começou quando me deparei com uma página da web que havia sido traduzida de uma introdução em PowerPoint para economia comportamental. Mencionou alguns resultados de heurísticas e vieses, de passagem, sem quaisquer referências. Fiquei tão surpreso que enviei um e-mail ao autor perguntando se isso era realmente um experimento real ou apenas uma anedota. Ele me respondeu, enviando uma digitalização do artigo de 1973 de Tversky e Kahneman.

É embaraçoso admitir que minha história não começa realmente por aí. Coloquei isso na minha lista de coisas para investigar. Eu sabia que existia um volume editado chamado «Julgamento sob incerteza: heurísticas e vieses», mas nunca o havia visto. Naquela época, pensei que, se não estivesse disponível online, tentaria seguir sem ele. Tinha tantas outras coisas na minha pilha de leitura e não tinha acesso fácil à biblioteca da universidade. Acho que devo ter mencionado isso em uma lista de discussão, porque Emil Gilliam ficou incomodado com minha teoria que existia apenas online, então ele me presenteou com o livro.

A atitude dele aqui provavelmente deve ser considerada como um ponto positivo.

No entanto, isso também não é o que eu chamo de minha "iluminação Bayesiana". Foi um passo significativo para perceber a inadequação das minhas habilidades na Racionalidade Tradicional - que havia muito mais para aprender, toda essa nova ciência, além de simplesmente seguir o que Richard Feynman disse para fazer. E ver o programa de heurísticas e vieses apresentando Bayes como o padrão de ouro ajudou a avançar meu pensamento - mas não completamente.

A memória é uma coisa frágil, e a minha parece ter se tornado mais frágil do que a maioria, desde que aprendi como as memórias são recriadas a cada lembrança - a ciência de quão frágeis elas são. As outras pessoas realmente têm memórias melhores ou apenas confiam nos detalhes que suas mentes inventam, embora na verdade não se lembrem de mais nada do que eu? Meu palpite é que outras pessoas têm uma memória melhor para certas coisas. Considero o conhecimento científico estruturado bastante fácil de lembrar; mas o caos desconectado da vida cotidiana desaparece muito rapidamente para mim.

Eu sei por que certas coisas aconteceram na minha vida - essa é a estrutura causal que eu recordo. Mas, às vezes, é difícil recordar em que ordem certos eventos ocorreram comigo, e muito menos em que ano.

Não tenho certeza se li Probability Theory: The Logic of Science (Teoria da Probabilidade: A lógica da ciência), de E. T. Jaynes, antes ou depois do dia em que percebi a magnitude da minha própria insanidade e entendi que estava enfrentando um problema de adultos.

Mas foi a Teoria da Probabilidade que fez a diferença. Aqui estava a teoria da probabilidade, apresentada não como uma ferramenta inteligente, mas como As Regras, invioláveis sob pena de paradoxo. Se você tentasse aproximar as regras porque eram computacionalmente caras demais para serem usadas diretamente, então, não importa quão necessário esse compromisso pudesse ser, você ainda acabaria fazendo menos do que o ideal. Jaynes faria seus cálculos de diferentes maneiras para mostrar que a mesma resposta sempre surgia quando você usava métodos legítimos; e ele apresentava respostas diferentes às quais outros haviam chegado e traçava o passo ilegítimo. Os paradoxos não poderiam coexistir com a sua precisão. Não é uma resposta, mas a resposta.

E então - depois de relembrar meus erros e todas as respostas que me levaram ao paradoxo e ao desânimo - ocorreu-me que aqui estava o nível acima do meu.

Eu não conseguia mais me ver tentando construir uma IA baseada em respostas vagas - como as respostas que encontrei anteriormente - e sobreviver ao desafio.

Observei os aspirantes a IAG com quem tentei discutir a IA Amigável e os diversos sonhos de Amigabilidade que eles tinham. (Muitas vezes formulados espontaneamente em resposta à minha pergunta!). Assim como os métodos estatísticos frequentistas, nenhum deles concordava entre si. Tendo estudado o assunto em tempo integral por alguns anos, eu conhecia alguns dos problemas que seus planos esperançosos enfrentariam. Vi que se alguém dissesse: "Não vejo por que isso iria falhar", o "não sei" seria apenas um reflexo de sua própria ignorância. Podia perceber que se seguisse um padrão semelhante a "isso parece uma boa ideia", também estaria condenado. (Muito parecido com um frequentista inventando novos cálculos estatísticos

incríveis que pareciam boas ideias).

Mas se você não consegue fazer o que parece uma boa ideia - se não consegue fazer o que não imagina falhar - então o que você pode fazer?

Pareceu-me que seria necessário algo como o nível de Jaynes - não, aqui está minha ideia brilhante, mas sim, aqui está a única maneira correta de fazer isso (e por quê) - para enfrentar um problema adulto e sobreviver. Se eu atingisse o mesmo nível de domínio do meu próprio assunto que Jaynes alcançou na teoria da probabilidade, então seria pelo menos imaginável que eu pudesse tentar construir uma IA amigável e sobreviver à experiência.

Pela minha mente ocorreu a passagem:

> Não faça nada porque é justo, louvável ou nobre fazê-lo; não faça nada porque parece bom fazê-lo; faça apenas aquilo que você deve fazer e que você não pode fazer de outra maneira[6]. [1]

Fazer o que parecia bom só me desencaminhava.

Então, eu declarei um ponto final.

E decidi que, a partir de então, seguiria a estratégia que poderia ter me salvado se a tivesse seguido anos atrás: manter meus projetos da FAI no padrão mais elevado de não fazer o que parecia uma boa ideia, mas apenas o que compreendia em um nível suficientemente profundo para ver que não poderia fazê-lo de outra maneira.

Todas as minhas antigas teorias, nas quais tanto investi, não atendiam a esse padrão; e não estavam próximas deste padrão; e nem sequer estavam no caminho que levava a esse padrão; então as joguei pela janela.

Comecei a estudar a teoria da probabilidade e a teoria da decisão, buscando ampliá-las para abranger conceitos como refletividade e auto-modificação.

Se bem me recordo, nesse ponto, eu já havia começado a enxergar a cognição como uma expressão da estrutura bayesiana, a qual também é uma parte crucial do que chamo de minha "iluminação bayesiana" – mas já discorri sobre isso anteriormente. Houve também o meu despertar naturalista, do qual já tratei. E minha percepção de que a Racionalidade Tradicional não era suficientemente rigorosa fez com que, em questões de racionalidade humana, eu passasse a me inspirar mais na teoria das probabilidades e na psicologia cognitiva.

Se somarmos todos esses elementos, teremos, mais ou menos, a narrativa de minha iluminação bayesiana.

A vida raramente estabelece limites claros. A história segue seu curso.

Foi ao estudar Judea Pearl, por exemplo, que percebi como a precisão pode otimizar o tempo. Já havia considerado a lógica não monotônica por conta própria - quando ainda estava na fase de "busca por ferramentas e algoritmos interessantes". Ao ler "Raciocínio Probabilístico em Sistemas Inteligentes: Redes de Inferência Plausível", [2] pude vislumbrar quanto tempo teria perdido em sistemas ad-hoc e casos especiais se não conhecesse essa abordagem. "Faça apenas o que deve ser feito e que não pode ser feito de outra maneira" se traduz em economia de tempo mensurável, não no resgate de meses perdidos, mas na preservação de carreiras que poderiam ter sido desperdiçadas.

Foi então que percebi que, mantendo esse padrão elevado de precisão, comecei verdadeiramente a refletir sobre várias questões cruciais. Expressar algo com precisão é desafiador – não é de forma alguma o mesmo que formalizar uma ideia ou inventar uma nova lógica para resolver um problema. Muitos evitam o desconforto, pois os seres humanos são propensos à preguiça, e assim dizem: 'É impossível' ou 'Levará muito tempo', embora nunca tenham realmente tentado por cinco minutos sequer. No entanto, se não se segue

6 NT. Texto original em inglês. *Do nothing because it is righteous, or praiseworthy, or noble, to do so; do nothing because it seems good to do so; do only that which you must do, and which you cannot do in any other way.*

esse padrão inconveniente, permite-se escapar de qualquer desafio. É difícil encontrar um padrão suficientemente alto que o force a verdadeiramente começar a pensar! Manter-se fiel ao padrão de prova matemática, onde cada passo deve estar correto e um passo errado pode levá-lo a qualquer lugar, pode parecer cansativo. No entanto, caso contrário, você não perseguirá aquelas pequenas discordâncias que, na verdade, levam a preocupações totalmente novas nas quais você nunca imaginou.

Portanto, hoje em dia, não reclamo tanto do fardo heroico de inconveniências necessário para manter um padrão preciso. Isso também pode otimizar o tempo; na verdade, é quase uma aposta para induzi-lo a pensar sobre o problema.

E isso também deveria ser considerado parte de minha 'iluminação Bayesiana' – perceber que há vantagens nisso, e não apenas penalidades.

Mas, é claro, a história continua. A vida é assim, pelo menos nas partes que me recordo.

Se há algo que aprendi com essa jornada, é que dizer 'Opa' é algo a ser desejado. Claro, a perspectiva de dizer "Opa!" no futuro significa que o você atual é um tolo, cujas palavras seu eu futuro não será capaz de ler devido a todas as caretas. No entanto, dizer "Opa" no futuro também significa que, naquele momento, você adquirirá novos poderes Jedi que seu eu atual nem sonha que existem. Isso faz você se sentir envergonhado, mas também vivo. Perceber que seu eu mais jovem era um completo idiota significa que, embora você já tenha vinte e poucos anos, ainda não atingiu o seu ápice. Assim, espero que meu eu futuro compreenda que sou um tolo: posso planejar resolver meus problemas com minhas habilidades atuais, mas poderes Jedi extras certamente seriam úteis.

Aquele grito de horror e vergonha é o som que os racionalistas emitem ao evoluir para um nível superior. Às vezes, me preocupo por não estar avançando tão rapidamente quanto antes e não sei se é porque finalmente estou pegando o jeito ou se meus neurônios estão morrendo lentamente.

Atenciosamente, Eliezer_{2008}.

Referências

[1] Le Guin, The Farthest Shore.

[2] Pearl, Probabilistic Reasoning in Intelligent Systems.

Parte Y - Desafiando o difícil

304 — Tsuyoku Naritai! (Eu quero me tornar mais forte)

No judaísmo ortodoxo existe um ditado: "A geração anterior está para a próxima como os anjos estão para os homens; a próxima geração está para a anterior como os burros estão para os homens". Isso decorre da crença ortodoxa judaica de que toda a lei judaica foi dada por Deus a Moisés no Monte Sinai. Afinal, não é como se pudéssemos realizar experimentos para obter novos conhecimentos *haláchicos*[7]; a única maneira de sabermos é se alguém nos disser (que ouviu de outra pessoa, que ouviu de Deus). Como não há novas fontes de informação, ela só pode ser degradada na transmissão de geração a geração.

Por isso, os rabinos modernos não têm permissão para anular os antigos. Coisas que rastejam normalmente não são kosher, mas é permitido comer um verme encontrado em uma maçã - os antigos rabinos acreditavam que o verme era gerado espontaneamente dentro da maçã e, portanto, fazia parte dela. Um rabino moderno não pode simplesmente dizer: "Sim, bem, os antigos rabinos sabiam pouco sobre biologia. Rejeitado!" Um rabino moderno não pode conhecer um princípio haláchico que os antigos rabinos não conheciam, porque como poderiam os antigos rabinos ter passado a resposta do Monte Sinai para ele? O conhecimento deriva da autoridade e, portanto, é sempre perdido, nunca ganho, com o passar do tempo.

Quando fui exposto pela primeira vez ao provérbio dos anjos e burros na escola primária (religiosa), eu ainda não tinha idade suficiente para ser um ateu completo, mas já pensava: "A Torá perde conhecimento a cada geração. A ciência ganha conhecimento a cada geração. Não importa onde elas começaram, mais cedo ou mais tarde a ciência deve superar a Torá."

O mais importante é que haja progresso. Enquanto seguimos em frente, alcançamos nossos objetivos; mas se pararmos de nos mover, nunca os alcançaremos.

Tsuyoku naritai é uma expressão japonesa composta por "tsuyoku", que significa "forte"; "naru", que significa "tornar-se"; e "tai", que é uma forma de "querer". Juntas, elas significam "quero me tornar mais forte" e expressam um sentimento mais profundamente incorporado nas obras japonesas do que em qualquer literatura ocidental que já li. Isso pode ser dito para expressar a determinação de se tornar um jogador profissional de Go - ou depois de perder uma partida importante, mas ainda não ter desistido - ou depois de vencer uma partida importante, mas ainda não ser um jogador do nono dan - ou depois de se tornar o maior jogador de Go de todos os tempos, mas ainda achar que pode fazer melhor. Isso é *tsuyoku naritai*, a vontade de transcender.

Tsuyoku naritai é a força motriz por trás do meu ensaio "O uso adequado da humildade", no qual comparo o aluno que humildemente verifica novamente seu teste de matemática e o aluno que diz modestamente: "Mas como podemos realmente saber? Não importa quantas vezes eu verifique, nunca posso ter certeza." O aluno que verifica suas respostas quer se tornar mais forte; eles reagem a uma possível falha interna fazendo o que podem para reparar a falha, e não com resignação.

Todos os anos, no Yom Kippur[8], um judeu ortodoxo recita uma litania que começa com *Ashamnu,*

---

7 NT. Do hebraico, *halacha*, ou lei em português. Conhecimentos haláchicos são os conhecimentos sobre a *halacha*, a lei judaica.

8 NT. **Yom Kippur**: Dia sagrado do judaísmo, conhecido como *Dia do Perdão*. Considerado o mais solene do calendário hebraico (10º dia de *Tishrei*), envolve jejum de 25 horas, orações intensas e reflexão sobre erros pessoais. Marca o encerramento dos *Dez Dias de Arrependimento*, iniciados em *Rosh Hashaná*. Observado com abstinência de trabalho e atividades mundanas.

bagadnu, gazalnu, dibarnu dofi[9] e continua por todo o alfabeto hebraico: agimos vergonhosamente, traímos, roubamos, caluniamos...

Ao pronunciar cada palavra, bate-se no coração em penitência. Não há isenção pela qual, se você conseguir ficar sem roubar o ano todo, possa pular a palavra *gazalnu* e atacar a si mesmo uma vez a menos. Isso violaria o espírito comunitário do Yom Kippur, que é sobre confessar pecados, não evitar pecados para que você tenha menos a confessar.

Da mesma forma, o *Ashamnu* não termina com a afirmação de que você fará melhor no próximo ano.»

O *Ashamnu* tem uma notável semelhança com a ideia de que o caminho da racionalidade é bater o punho contra o coração e dizer: "Somos todos tendenciosos, somos todos irracionais, não estamos totalmente informados, somos excessivamente confiantes, somos mal calibrados..."

Certo. Agora me diga como você planeja se tornar menos tendencioso, menos irracional, mais informado, menos confiante demais, melhor calibrado.

Há uma velha piada judaica: durante o Yom Kippur, o rabino é tomado por uma onda repentina de culpa, prostra-se e grita: "Deus, não sou nada diante de ti!" O cantor também é tomado pela culpa e grita: "Deus, eu não sou nada diante de ti!" Vendo isso, o zelador no fundo da sinagoga se prostra e grita: "Deus, eu não sou nada diante de ti!" E o rabino cutuca o cantor e sussurra: "Olha quem pensa que não é nada".

Não se orgulhe de sua confissão de que você também é tendencioso; não se vanglorie em sua autoconsciência de suas falhas. Isso é semelhante ao princípio de não se orgulhar de confessar sua ignorância; pois se sua ignorância é uma fonte de orgulho para você, você pode relutar em renunciar a sua ignorância quando as evidências batem à sua porta. Da mesma forma, com nossas falhas - não devemos nos gabar de como somos autoconscientes por confessá-las; a ocasião de regozijo é quando temos um pouco menos para confessar.

Caso contrário, quando alguém vier até nós com um plano para corrigir o preconceito, rosnaremos: "Você pensa em se colocar acima de nós?" Vamos balançar nossas cabeças com tristeza e dizer: "Você não deve ser muito autoconsciente".

Nunca me confesse que você é tão falho quanto eu, a menos que possa me dizer o que planeja fazer a respeito. Depois disso, você ainda terá muitas falhas, mas esse não é o ponto; o importante é fazer melhor, seguir em frente, dar mais um passo à frente. *Tsuyoku naritai!*

---

9 NT. **Ashamnu, bagadnu, gazalnu, dibarnu dofi**: Início da oração litúrgica *Vidui* (confissão) no judaísmo, recitada durante *Yom Kippur* e os *Dez Dias de Arrependimento*. Enumera transgressões coletivas em ordem alfabética hebraica (*Aleph* a *Tav*), usando verbos como "pecamos" (*ashamnu*), "traímos" (*bagadnu*), "roubamos" (*gazalnu*) e "falamos calúnias" (*dibarnu dofi*). Simboliza arrependimento comunitário e humildade.

305 — Tsuyoku contra o instinto igualitário

As tribos de caçadores-coletores geralmente apresentam um alto grau de igualdade (pelo menos para os homens). O líder tribal todo-poderoso é mais comumente encontrado em sociedades agrícolas, raramente no ambiente ancestral. Na maioria das tribos de caçadores-coletores, um caçador que realiza uma matança espetacular tende a minimizar cuidadosamente suas conquistas para evitar despertar inveja.

Possivelmente, se você começar abaixo da média, poderá progredir sem chamar muita atenção. No entanto, mais cedo ou mais tarde, se pretende alcançar seu melhor desempenho, é preciso estabelecer metas acima da média.

Se você não consegue admitir para si mesmo que se saiu melhor do que os outros, ou sente vergonha de aspirar a superá-los, então a mediana se tornará perpetuamente seu obstáculo intransponível, o ponto em que você para de avançar. E quanto àqueles que estão abaixo da média? Você ousaria afirmar que pretende superá-los? Que orgulho da sua parte!

Talvez não seja saudável se vangloriar por superar outra pessoa. Pessoalmente, descobri que isso serve como um estímulo útil, apesar de meus princípios, e aproveitarei toda motivação útil que puder obter. Talvez esse tipo de competição seja um jogo de soma zero, mas o Go também o é; isso não significa que devemos abolir essa atividade humana se as pessoas a considerarem divertida e capaz de levar a lugares interessantes.

De qualquer forma, certamente não é saudável sentir-se envergonhado por superar os outros.

Além disso, a vida não segue uma curva graduada. A vontade de transcender não possui um ponto de parada, no qual se transforma em uma vontade de desempenho inferior. E a corrida sem linha de chegada não concede medalhas de ouro ou prata. Apenas corra o mais rápido possível, sem se preocupar em ultrapassar outros corredores. (Mas esteja avisado: se você se recusar a considerar essa possibilidade, pode se ver ultrapassado eventualmente. Se você Ignorar as consequências, elas podem acontecer.)

Em algum momento, se estiver seguindo um caminho verdadeiro, começará a mitigar uma falha que a maioria das pessoas não abordou. Eventualmente, se seus esforços derem frutos, terá menos pecados para confessar.

Talvez você ache que o caminho da sabedoria é minimizar suas realizações, mesmo em caso de sucesso. As pessoas podem perdoar um *touchdown*, mas não dançar na zona de pontuação. Certamente, achará mais rápido, fácil e conveniente negar publicamente seu valor, fingindo ser tão pecador quanto qualquer outra pessoa. Desde que, é claro, todos saibam que isso não é verdade. Pode ser divertido exibir orgulhosamente sua modéstia, contanto que todos estejam cientes de quanto você precisa ser modesto.

Mas não permita que essa seja a conclusão de sua jornada. Mesmo que apenas sussurre para si mesmo, sussurre ainda assim: *Tsuyoku, tsuyoku!* Mais forte, mais forte!

E, em seguida, estabeleça uma meta mais elevada. Esse é o verdadeiro significado de perceber que ainda possui falhas (embora um pouco menos). Significa sempre aspirar a alturas maiores, sem vergonha.

Tsuyoku naritai! Correrei sempre o mais rápido que puder; mesmo que avance, continuarei correndo. Alguém, algum dia, me superará; mas mesmo que eu fique para trás, sempre correrei o mais rápido que puder.

306 — Tentando tentar

Não! Não tente! Faça ou não faça. Tentativa não há.

—Yoda

Anos atrás, pensei que isso fosse apenas mais um exemplo de uma Sabedoria Profunda que, na verdade, é bastante estúpida. Ter sucesso não é uma ação primária. Não se pode simplesmente decidir vencer por meio de uma escolha extrema o suficiente. Nunca há um plano que funcione com probabilidade 1.

Mas Yoda era mais sábio do que eu imaginava.

A primeira técnica elementar da epistemologia – não é profunda, mas é acessível – é distinguir a citação do referente. Falar sobre neve não é o mesmo que falar sobre "neve". Quando uso a palavra "neve", sem aspas, quero dizer discutir o conceito de neve; e quando uso a palavra "'neve'", entre aspas, quero dizer falar da própria palavra "neve". É necessário entrar em um modo especial, o modo de citação, para discutir sobre nossas crenças. Por padrão, falamos apenas sobre a realidade.

Se alguém disser: "Vou acionar esse interruptor", então, por padrão, isso significa que tentará acionar o interruptor. Eles vão elaborar um plano que promete levar, pelas consequências de suas ações, ao estado desejado de um interruptor acionado; e então executar esse plano.

Nenhum plano é bem sucedido com certeza infinita. Portanto, por padrão, quando você fala sobre se preparar para atingir uma meta, você não está sugerindo que seu plano leve exata e perfeitamente a somente essa possibilidade. Mas quando você diz: "Vou acionar esse interruptor", você está apenas tentando acionar o interruptor – e não tentando alcançar uma probabilidade de 97,2% de acionar o interruptor.

Então, o que significa quando alguém diz: "Vou tentar acionar esse interruptor?"

Bem, coloquialmente, "Vou acionar o interruptor" e "Vou tentar acionar o interruptor" significam mais ou menos a mesma coisa, exceto que o último expressa a possibilidade de fracasso. É por isso que inicialmente me incomodei com Yoda, porque ele parecia negar essa possibilidade. Mas peço paciência aqui.

Grande parte do desafio da vida consiste em manter-se em um padrão suficientemente elevado. Posso falar mais sobre esse princípio mais tarde, pois é uma lente pela qual você pode analisar muitos dilemas pessoais, embora não todos: "A que padrão estou me mantendo? É suficientemente alto?"

Portanto, se grande parte do fracasso na vida resulta de manter-se em um padrão muito baixo, é preciso ter cuidado para não exigir muito pouco de si mesmo – estabelecendo metas que são demasiado fáceis de alcançar."

Muitas vezes, onde é muito difícil conseguir realizar algo, tentar fazê-lo é consideravelmente mais fácil.

Qual é a tarefa mais fácil: construir uma startup de sucesso ou simplesmente tentar construir uma startup de sucesso? Ganhar um milhão de dólares ou tentar ganhar um milhão de dólares?

Portanto, se dizer "Vou acionar o interruptor" implica, por padrão, que você vai tentar acionar o interruptor – ou seja, estabelecer um plano que promete levar ao estado de interruptor acionado, talvez não com probabilidade 1, mas com a maior probabilidade que você puder gerenciar—

—então, afirmar "Vou 'tentar acionar' o interruptor" significa que você se propõe a "tentar acionar o interruptor". Em outras palavras, você está comprometido a atingir o estado desejado de "ter um plano que possa acionar o interruptor."

Agora, se estivéssemos falando de uma IA auto-modificável, a transformação que acabamos de realizar deveria resultar em um equilíbrio reflexivo – a IA planejando suas operações de planejamento.

Contudo, quando lidamos com seres humanos, estar satisfeito por ter um plano não se assemelha em nada a estar satisfeito com o sucesso. A parte na qual o plano deve maximizar sua probabilidade de sucesso frequentemente se perde no caminho. É notavelmente mais fácil nos convencermos de que estamos "maximizando nossa probabilidade de sucesso" do que nos convencermos de que, de fato, teremos sucesso.

Quase qualquer esforço servirá para nos convencer de que "tentamos o nosso melhor", se tentar o nosso melhor é tudo o que estamos procurando realizar.

> Você tem se perguntado o que poderia fazer diante dos grandes eventos que estão acontecendo agora e descobriu que não poderia fazer nada. Mas isso ocorre porque o seu sofrimento fez com que você formulasse a pergunta de maneira equivocada. . . Em vez de perguntar o que você poderia fazer, você deveria estar indagando o que precisa ser feito[10].
>
> —Steven Brust, The Paths of the Dead [1]

Quando você pergunta: "O que posso fazer?", você está tentando fazer o seu melhor. Qual é o seu melhor? É tudo o que você pode fazer sem o menor incômodo. É tudo o que você pode fazer com o dinheiro que tem no bolso, descontando o valor necessário para o seu almoço habitual. O que você pode realizar com esses recursos pode não garantir boas chances de vitória. Contudo, é o "melhor que você pode fazer", e assim, você age de maneira defensável, certo?

Mas o que precisa ser feito? Talvez o que precisa ser feito exija três vezes as economias de uma vida inteira, e você deve providenciar isso ou enfrentar o fracasso.

Assim, buscar "maximizar sua probabilidade de sucesso" – ao invés de buscar o sucesso em si – apresenta uma barreira consideravelmente menor. É possível "maximizar sua probabilidade de sucesso" utilizando apenas o dinheiro no bolso, desde que não seja uma exigência efetiva de vencer.

Quer tentar ganhar um milhão de dólares? Compre um bilhete de loteria. Suas chances de ganhar podem não ser as melhores, mas você tentou, e tentar era o seu objetivo. Na verdade, você deu o seu melhor, já que só restou um dólar após comprar o almoço. Maximizar as chances de atingir objetivos com os recursos disponíveis: isso não é inteligência?

Somente quando você realmente deseja, acima de tudo, acionar o interruptor — sem aspas e sem prêmios de consolação apenas por tentar — é que você se esforçará verdadeiramente para maximizar a probabilidade.

No entanto, se tudo o que você deseja é apenas "maximizar a probabilidade de sucesso utilizando os recursos disponíveis", então isso se torna a coisa mais fácil do mundo de se convencer de que alcançou isso. O primeiro plano que você encontrar servirá perfeitamente como "maximização" – se necessário, você pode gerar uma alternativa inferior para comprovar sua otimização. E qualquer pequeno recurso que você queira investir será considerado "disponível". Lembre-se de se parabenizar por investir 100% disso!

Não se esforce ao máximo. Vença, ou fracasse. Não há máximo.

Referências

[1] Steven Brust, The Paths of the Dead, Vol. 1 of The Viscount of Adrilankha (Tor Books, 2002).

10 NT. Texto original em inglês. *You have been asking what you could do in the great events that are now stirring, and have found that you could do nothing. But that is because your suffering has caused you to phrase the question in the wrong way… Instead of asking what you could do, you ought to have been asking what needs to be done.*

307 — Use o “Se esforce mais, Luke”

Quando existe uma vontade de falhar, obstáculos podem ser encontrados[11].

—John McCarthy

Eu assisti pela primeira vez Guerra nas Estrelas IV-VI quando eu era bem jovem. Quando tinha sete anos, talvez, ou nove? Então, minha memória é um pouco turva, mas eu me lembro do Luke Skywalker, sabe, aquele cara Jedi legal.

Imagine o meu horror e decepção, quando eu assisti à saga novamente, anos depois, e descobri que o Luke não passava de um adolescente chorão.

Eu digo isso porque, ontem, eu procurei, no YouTube, a fonte da citação do Yoda: “Faça ou não faça. Tentativa não há.”

Oh! Meu. Cthulhu[12].

Apresento a vocês um corte pouco conhecido da cena, na qual o diretor e roteirista George Lucas, discute com Mark-Hamill, o ator que interpreta Luke Skywalker:

LUKE: “Está bem, vou tentar.”

YODA: “Não! Não tente! Faça ou não faça. Tentativa não há.”

Luke levanta a mão, e lentamente, a X-wing começa a se elevar da água - os olhos do Yoda se arregalam - mas, em seguida, a nave afunda novamente.

MARK HAMILL: “Oi, George . . .”

GEORGE LUCAS: “O que foi agora?”

MARK: “Então… conforme o roteiro, depois eu digo, “Não consigo. É grande demais.”

GEORGE: “Isso mesmo”.

MARK: ““O Luke não deveria tentar mais uma vez?”

GEORGE: “Não. O Luke desiste e se senta ao lado do Yoda—”

MARK: “Este é o herói que vai derrubar o Império? Olha, uma coisa era quando ele era um adolescente chorão no começo, mas agora ele é um Jedi em treinamento. No último filme ele explodiu a Estrela da Morte. Luke deveria estar mostrando um pouco de coragem.”

GEORGE: “Não. Você desiste. E daí o Yoda te dá um sermão por alguns minutos, e você diz, ‘Você quer que eu faça o impossível’ Você pode se lembrar disso?”

---

11 NT. Texto original em inglês. *When there's a will to fail, obstacles can be found.*

12 NT. **Cthulhu**: Entidade cósmica fictícia criada por *H.P. Lovecraft* (1928), parte do panteão dos *Grandes Antigos* no *Mythos de Cthulhu*. Descrito como uma criatura colossal com cabeça de polvo, asas dragônicas e corpo escamoso, habita a cidade submersa de *R'lyeh*. Simboliza o *horror cósmico* e a insignificância humana perante o universo.

MARK: “Impossível? O que ele fez, executou um cálculo formal para chegar a uma prova matemática? O X-wing já estava começando a sair do pântano! Essa é a demonstração de viabilidade aí mesmo! Luke perde o controle por um segundo e o navio afunda – e agora ele diz que é impossível? Sem mencionar que Yoda, que tem literalmente oitocentos anos de experiência na área, acabou de dizer a ele que isso deveria ser viável...

GEORGE: “E, então, você vai embora.”

MARK: “É a maldita nave espacial dele! Se ele o deixar no pântano, ficará preso em Dagobah pelo resto de sua vida miserável! Ele não vai simplesmente ir embora! Olha, vamos cortar para a próxima cena com as palavras ‘um mês depois’ e o Luke continua lá, todo esfarrapado, de pé, em frente ao pântano, tentando levantar a nave pela milésima vez-”

GEORGE: “Não.”

MARK: “Tudo bem! Mostraremos um pôr do sol e um nascer do sol, enquanto ele fica lá movimentando seus braços, com dificuldade, e daí o Luke diz “É impossível”. Embora, na verdade, ele deva tentar novamente quando estiver totalmente descansado...

GEORGE: “Não.”

MARK: “Só cinco malditos minutos e ele já desiste!”

GEORGE: “Eu não vou parar a trama por cinco minutos enquanto a X-wing fica subindo e descendo do pântano igual um brinquedo de banheira.”

MARK: “Tenha santa paciência! Se um perdedor patético como esse pudesse dominar a Força, todo mundo na galáxia estaria usando-a também! As pessoas se tornariam Jedi porque seria mais fácil do que o ensino médio.”

GEORGE: “Olha, você é o ator. Deixe-me ser o contador de histórias. Só diga as suas falas e tente colocar sentimento nelas.”

MARK: “O público não vai acreditar nisso.”

GEORGE: “Confie em mim, eles vão.”

MARK: “Eles vão se levantar e sair do cinema.”

GEORGE: “Eles vão ficar sentados, concordar com a cabeça e não vão notar nada fora do comum. Olha, você não entende a natureza humana. As pessoas não tentariam por mais de cinco minutos antes de desistirem, nem se o futuro da humanidade estivesse em risco.”

308 — Sobre fazer o impossível

"Persistir." É um conselho que você ouvirá de muitos empreendedores de sucesso em várias disciplinas. Inicialmente, não compreendi completamente.

No começo, eu associava "persistência" a trabalhar 14 horas por dia. Aparentemente, há quem consiga dedicar 10 horas a um trabalho técnico e, nos intervalos entre comer, dormir e ir ao banheiro, aproveitar o tempo livre para escrever um livro. Eu não me enquadro nesse grupo - até hoje, admitir isso me causa certo desconforto. Estou envolvido em algo significativo; meu cérebro não deveria estar disposto a funcionar por 14 horas diárias? Mas não é o caso. Quando a tarefa fica demasiadamente desafiadora, eu paro e dedico um tempo para ler ou assistir a algo. Por causa disso, por anos, pensei que me faltava completamente a virtude da "persistência".

Seguindo a lógica humana, o Eliezer$_{1998}$ diria coisas como: "O que importa é a produção, não o esforço." Ou "A preguiça também é uma virtude - ela nos impede de persistir em métodos falhos e nos incentiva a buscar maneiras melhores." Ou ainda, "Estou me saindo melhor do que pessoas que trabalham mais horas. Talvez, para o trabalho criativo, o pico momentâneo de produção seja mais crucial do que trabalhar 16 horas diárias." Talvez os cientistas famosos tenham sido seduzidos pela Sabedoria Profunda de afirmar que "o trabalho árduo é uma virtude", pois seria terrível se isso valesse menos do que a inteligência?

Não compreendi a virtude da persistência até olhar para trás, para a minha jornada na área de IA, e perceber que superestimei a dificuldade de quase todos os problemas cruciais.

Pode parecer insano, não é mesmo? Mas peço que tenha paciência comigo aqui.

Quando decidi desafiar pela primeira vez a inteligência artificial, estava pensando em termos de um escalas temporais de 40 anos, Projetos Manhattan[13], redes de computadores planetárias, milhões de programadores e, possivelmente, humanos aprimorados.

Esse é um erro comum no futurismo da IA, sobre o qual escreverei mais tarde; envolve a transição de "não sei como resolver isso" para "vou conceber algo verdadeiramente grandioso para isso". Algo grandioso o suficiente para que, quando concebido, essa ideia gere uma sensação de grandiosidade forte o bastante para ser equiparada ao problema. (Atualmente, há alguém na comunidade de IA afirmando que a IA custará um quatrilhão de dólares - não conseguiremos desenvolver a IA sem gastar um quatrilhão de dólares, mas poderíamos desenvolvê-la a qualquer momento gastando esse valor.) Isso, por sua vez, permite a ilusão de que se sabe como resolver a IA, sem tentar atender à demanda claramente impossível de compreender a inteligência.

Então, inicialmente, cometi o mesmo equívoco: não entendia a inteligência, então imaginei lançar um Projeto Manhattan para resolver o problema.

Entretanto, ao calcular a taxa de mortalidade planetária em 55 milhões por ano ou 150.000 por dia, não fugi do grande problema amedrontador como um coelho assustado. Pelo contrário, comecei a explorar

13 NT. **Projeto Manhattan**: Iniciativa secreta dos EUA (1939-1945) para desenvolver a primeira bomba atômica durante a Segunda Guerra Mundial. Envolveu cientistas como *Robert Oppenheimer* e instalações como *Los Alamos*, *Oak Ridge* e *Hanford*. Resultou nos bombardeios de Hiroshima e Nagasaki (1945). Marcou o início da era nuclear e gerou debates éticos sobre ciência e guerra.

que tipo de projeto de IA poderia alcançar o objetivo mais rapidamente. Se eu conseguisse fazer com que o surgimento da inteligência acontecesse uma hora antes, seria um retorno razoável do investimento para uma carreira pré-explosão. (Neste momento, eu não estava pensando em termos de riscos existenciais ou de uma IA Amigável.)

Então, não evitei o grande desafio assustador como um coelho assustado, mas permaneci para ver se havia algo que eu pudesse fazer.

Fato histórico divertido: em 1998, elaborei este extenso tratado propondo como proceder para criar uma IA de auto-aperfeiçoamento ou "semente" (um termo que tive a honra de cunhar). Brian Atkins, que mais tarde se tornaria o fundador do Instituto de Pesquisa em Inteligência de Máquina (MIRI), acabara de vender o Hypermart para a Go2Net. Brian me enviou um e-mail perguntando se esse projeto de IA que eu estava descrevendo era algo que uma equipe de tamanho razoável poderia realmente realizar. "Não", eu disse, "seria necessário um Projeto Manhattan e trinta anos", então, por um tempo, estávamos considerando uma nova startup de tecnologia, para criar o financiamento necessário para realizar um trabalho real em IA...

Um ou dois anos depois, ao tomar conhecimento dessa nova tendência de "código aberto", pareceu-me que havia algum trabalho preliminar de desenvolvimento - novas linguagens de programação e assim por diante - que uma pequena organização poderia realizar; e foi assim que o MIRI começou.

Essa estratégia estava, obviamente, completamente equivocada.

No entanto, mesmo assim, passei de "Não há nada que eu possa fazer sobre isso agora" para "Hum… " talvez haja um caminho incremental através do desenvolvimento de código aberto, se as versões iniciais forem úteis para um número suficiente de pessoas."

Isso ocorreu nos primórdios, então não estou dizendo que nada disso tenha sido uma boa ideia. Mas em termos do que eu pensava estar tentando realizar, um ano de pensamento criativo encurtou o caminho aparente: o problema parecia um pouco menos impossível do que da primeira vez que o abordei.

O ponto mais interessante é minha incursão na IA Amigável. Inicialmente, a IA Amigável não tinha sido algo que eu havia considerado, pois era obviamente impossível e inútil enganar uma superinteligência sobre qual seria o curso de ação correto.

Assim, historicamente, transitei de simplesmente ignorar um problema considerado "impossível" para assumir um desafio que era meramente extremamente difícil.

Naturalmente, isso resultou em um aumento significativo na minha carga de trabalho total.

O mesmo se aplica à compreensão da inteligência em um nível preciso. Inicialmente, eu considerava esse problema como algo impossível, excluindo-o, assim, da minha carga de trabalho. (Essa lógica parece bastante questionável em retrospecto – a Natureza não se importa com o que você não pode fazer quando está delineando os requisitos do seu projeto – mas ainda vejo profissionais em IA tentando fazer isso o tempo todo.) Manter-me em um padrão preciso significava investir mais esforço do que inicialmente imaginava ser necessário. No entanto, também significava enfrentar um desafio que teria descartado como totalmente impossível não muito tempo atrás.

Embora os desafios individuais na área de IA pareçam se tornar menos intimidantes ao longo do tempo, o número total de obstáculos a serem superados aumentou em magnitude – conforme a sabedoria convencional prevê – à medida que esses desafios foram retirados da categoria de "impossíveis" e colocados na lista de "tarefas".

Entendi o que estava ocorrendo – e o verdadeiro significado de "Persistir!" – no momento em que observei outros profissionais de IA fazendo o mesmo: declarando "Impossível!" diante de problemas que pareciam plenamente solucionáveis – relativamente mais simples, considerando o contexto. No entanto, eram questões que teriam parecido muito mais imponentes no momento em que me deparei com o problema pela primeira vez.

E percebi que a palavra "impossível" possui dois significados:

1. Demonstração matemática de impossibilidade condicional a axiomas especificados;
2. “Não consigo vislumbrar uma maneira de fazer isso.”

Desnecessário dizer que todos os meus usos da palavra “impossível” foram do segundo tipo.

Sempre que não compreendemos um domínio, muitos desafios nesse campo podem parecer impossíveis, pois, ao consultar nossa mente em busca de uma solução, ela retorna vazia. No entanto, existem apenas perguntas misteriosas, nunca respostas misteriosas. Se você dedicar um ou dois anos ao domínio em questão, se você evitar quaisquer becos sem saída e se você possuir a habilidade inata necessária para progredir, você vai compreender melhor. A aparente dificuldade dos problemas pode diminuir consideravelmente. Eles não serão tão assustadores quanto pareciam inicialmente.

E isso é especialmente provável nos problemas **confusos** que parecem mais **intimidadores**.

Como temos alguma noção dos processos pelos quais uma estrela queima, sabemos que não é fácil construir uma estrela do zero. Como entendermos as engrenagens, podemos provar que nenhum conjunto de engrenagens, seguindo a física conhecida, pode originar uma máquina de movimento perpétuo. Esses não são problemas ideais para praticar o impossível.

Quando você está confuso em relação a um domínio, os problemas contidos nele parecem muito intimidantes e misteriosos. Consultar nossa mente pode resultar em zero soluções. No entanto, desconhecemos o quanto de trabalho ainda resta quando a confusão se dissolve. Claro, dissolver a confusão pode ser, por si só, um desafio muito difícil, mas o termo “impossível” não deveria ser aplicado nesse contexto. A confusão reside no mapa, não no território.

Assim, se dedicarmos anos a um problema aparentemente impossível, conseguirmos evitar becos sem saída e possuirmos habilidades nativas suficientemente elevadas para progredir, então, caramba, após alguns anos, o que parecia inicialmente impossível pode não mais nos intimidar.

Mas se algo nos parece impossível, não tentaremos.

Isso cria um ciclo vicioso.

Se eu não estivesse totalmente convencido de que “quarenta anos e um Projeto Manhattan” significavam apenas que deveríamos começar mais cedo, eu não teria tentado. Eu não teria persistido no problema. E não teria tido a oportunidade de me sentir menos intimidado.

Geralmente, não sou adepto da teoria de que vieses opostos se anulam, mas ocasionalmente isso ocorre por sorte. Se eu tivesse visualizado toda a montanha desde o início — se tivesse percebido, desde o princípio, que o desafio não era criar uma semente capaz de aprimorar-se, mas sim produzir uma IA Amigável com correção comprovada — provavelmente teria hesitado.

Contudo, parte da compreensão desses cientistas acima da média, que constituem a maioria dos pesquisadores em IAG, é perceber que não estão enfrentando um problema quase impossível, mesmo que isso leve 40 anos. Geralmente, estão lá porque descobriram a chave para a IA que lhes permitirá resolver o problema sem grandes dificuldades, em apenas cinco anos.

Richard Hamming costumava questionar seus colegas cientistas com duas perguntas: “Quais são os problemas importantes em sua área?” e “Por que você não está trabalhando neles?”

Frequentemente, os problemas importantes parecem imponentes, intimidadores, enormes. Eles não garantem 10 publicações por ano. Não prometem qualquer avanço tangível. Pode ser que, após um ano, cinco anos ou dez anos de trabalho dedicado, você não receba nenhuma recompensa.

E não é raro que os problemas mais cruciais em sua área sejam impossíveis. É por isso que não vemos mais filósofos se dedicando a decomposições reducionistas da consciência.

Tentar fazer o impossível definitivamente não é para todos. O talento excepcional é apenas a aposta para se sentar à mesa. As fichas são os anos da sua vida. Se apostar essas fichas e perder parecer uma possibilidade insuportável para você, então considere outras opções. É sério. Porque você pode perder.

Não vou dizer algo como: “Todos deveriam tentar algo impossível pelo menos uma vez na vida, porque isso ensina uma lição crucial”. Na maioria das vezes, e para a maioria das pessoas, é mais sensato se ater ao possível.

Nunca desistir? Não seja ridículo. Fazer o impossível deve ser reservado para ocasiões muito especiais. Aprender quando abandonar a esperança é uma habilidade crucial na vida.

Mas se houver algo que você possa imaginar que seja ainda pior do que desperdiçar sua vida, se houver algo que deseje mais do que trinta fichas, ou se existirem coisas mais aterrorizantes do que uma vida de inconveniências, então talvez você tenha razões para tentar o impossível.

Há muito a ser dito sobre persistir nas dificuldades, mas uma das verdades é que torna as coisas mais difíceis. Se não puder lidar com isso, mantenha-se afastado! Existem formas mais fáceis de obter elegância e respeito. Não quero que ninguém leia isso e se envolva desnecessariamente em uma vida de dificuldades permanentes.

Para concluir, é essencial destacar que a “persistência” exigida para abordar questões significativas possui um elemento que transcende a simples jornada de trabalho de 14 horas diárias.

É estranho o padrão do que notamos ou não em nós mesmos. Essa seletividade nem sempre visa inflar a nossa autoimagem, mas, por vezes, reflete uma saliência comum.

Manter-me engajado no trabalho sempre foi um desafio constante, e, portanto, foi crucial perceber que seria incapaz de dedicar 14 horas consecutivas por dia. Não me ocorreu que a “persistência” poderia ser aplicada em diferentes escalas de tempo, seja ela de segundos ou anos. Essa compreensão só surgiu quando me deparei com pessoas que instantaneamente rotulavam como “impossível” qualquer desafio que não estivessem dispostas a enfrentar. Ou ainda, ao testemunhar a relutância delas em aceitar tarefas que demandavam décadas em vez de apenas “cinco anos”.

Foi nesse momento que percebi que a “persistência” é válida em várias dimensões temporais. Em uma escala de segundos, persistência significa “não desistir instantaneamente diante do primeiro sinal de dificuldade”. Já em uma escala de anos, persistência implica “continuar a enfrentar um problema extremamente difícil, mesmo que isso seja inconveniente, e que recompensas pessoais possam parecer mais acessíveis em outro lugar”.

Para realizar feitos que são verdadeiramente desafiadores ou considerados “impossíveis”,

É primordial não se esquivar inicialmente. Isso demanda apenas alguns segundos.

Em seguida, é necessário trabalhar. O que leva horas.

Por fim, é crucial persistir. Isso leva anos.

Dentre esses elementos, aprendi a fazer o primeiro de forma consistente, em vez de esporadicamente; o segundo ainda representa uma batalha constante para mim; enquanto o terceiro surge naturalmente.

309 — Faça um esforço extraordinário

> É crucial que o ser humano se dedique de todo o coração e compreenda que até mesmo atingir a média é difícil, se ele não tem a intenção de superar os outros no que quer que ele faça[14].
>
> —*Budo Shoshinshu*[15] [1]

> Em questões importantes, um esforço "intenso" geralmente resulta apenas em resultados medianos. Sempre que estivermos tentando algo que realmente valha a pena, nosso empenho deve ser como se nossa vida estivesse em jogo, como se estivéssemos sob um ataque físico! É esse esforço extraordinário – um esforço que nos leva além do que pensávamos ser capazes – que garante a vitória na batalha e o sucesso em nossos empreendimentos na vida[16].
>
> —*Flashing Steel: Mastering Eishin-Ryu Swordsmanship* (Aço reluzente: Dominando a Arte da Espada Eishin-Ryu. [2]

"Um esforço 'forte' geralmente resulta apenas em resultados medianos" – já ouvi isso repetidas vezes. O menor esforço é suficiente para nos convencermos de que fizemos o nosso melhor.

Existe um nível além da virtude do *tsuyoku naritai* ("Eu quero ficar mais forte"). *Isshoukenmei* era originalmente a lealdade que um samurai oferecia em troca de sua posição, contendo caracteres para "vida" e "terra". O termo evoluiu para significar "fazer um esforço desesperado": dar o seu melhor, o máximo, como se sua vida estivesse em jogo. Fazia parte da gestalt do *bushido*, que não se restringia apenas à luta. Encontrei formas variantes *isshokenmei* e *isshoukenmei*; uma fonte indica que o primeiro indica um esforço total em algum ponto específico, enquanto o último indica um esforço ao longo da vida.

Tento não elogiar demais o Oriente, porque há uma tremenda seletividade nas partes da cultura oriental de que o Ocidente ouve falar. Mas, pelo menos em alguns pontos, a cultura do Japão se destaca mais do que a da América. Ter uma expressão compacta e prática para "fazer um esforço desesperado e total, como se sua própria vida estivesse em jogo" é um desses pontos. É o tipo de coisa que um pai japonês diria a um estudante antes dos exames – mas não pense que é uma hipocrisia barata, como seria se um pai americano fizesse a mesma afirmação. Eles levam os exames muito a sério no Japão.

Ocasionalmente, alguém pergunta por que as pessoas que se autodenominam "racionalistas" nem

---

14 NT. Texto original em inglês. *It is essential for a man to strive with all his heart, and to understand that it is difficult even to reach the average if he does not have the intention of surpassing others in whatever he does.*

15 NT. **Budo Shoshinshu**: Texto clássico japonês do século XVII, atribuído a *Daidoji Yuzan*, que explora os princípios do *bushido* (código samurai). Orienta jovens guerreiros em conduta ética, dever social e disciplina marcial, enfatizando lealdade, honra e autocontrole. Integra filosofia confucionista com práticas bélicas, refletindo a transição do samurai de guerreiro para administrador no período Edo. Base para estudos modernos sobre ética marcial e influência em obras como *"Hagakure"*.

16 NT. Texto original em inglês. *In important matters, a "strong" effort usually results in only mediocre results. Whenever we are attempting anything truly worthwhile our effort must be as if our life is at stake, just as if we were under a physical attack! It is this extraordinary effort—an effort that drives us beyond what we thought we were capable of—that ensures victory in battle and success in life's endeavors.*

sempre parecem ter uma vida muito melhor, e pela minha própria história, a resposta parece direta: é necessária uma tremenda dose de racionalidade antes de parar de cometer erros estúpidos.

Como já mencionei em diversas ocasiões: Robert Aumann, o ganhador do Nobel que primeiro demonstrou que bayesianos com os mesmos antecedentes não podem concordar em discordar, é um judeu ortodoxo devoto. Certamente, ele compreende a matemática por trás da teoria da probabilidade, mas isso por si só não é suficiente para salvá-lo. O que mais é necessário? Estudar heurísticas e vieses? Psicologia Social? Psicologia evolucionária? Sim, contudo, também é vital o *isshoukenmei*, um esforço desesperado para ser racional – para elevar-se acima do nível de Robert Aumann.

Às vezes me questiono se não deveria promover a racionalidade no Japão ao invés dos Estados Unidos – entretanto, o Japão não ostenta uma superioridade científica sobre os Estados Unidos, apesar de seus estudantes mais dedicados. Os japoneses não dominam o mundo atualmente, embora na década de 1980 se especulasse amplamente que o fariam (daí a bolha de ativos japonesa). Por que não?

No Ocidente, existe um ditado que diz: “A roda que faz barulho recebe a graxa”.

No Japão, o ditado correspondente é: “O prego que se destaca é martelado”.

Esta observação não é original minha: no entanto, o empreendedorismo, a disposição para correr riscos e a rejeição da conformidade ainda são vantagens que o Ocidente possui sobre o Oriente. E, como os cientistas japoneses ainda não alcançaram a supremacia em relação aos americanos, isso parece contar pelo menos tanto quanto os esforços desesperados.

Qualquer pessoa que consiga reunir força de vontade por trinta segundos pode fazer um esforço desesperado para levantar mais peso do que normalmente conseguiria. Mas e se o peso que precisa ser levantado for um caminhão? Nesse caso, os esforços desesperados não serão suficientes; será necessário realizar algo fora do comum para ter sucesso. Talvez seja preciso fazer algo que não foi ensinado na escola. Algo que os outros não esperam que você faça e que podem não entender. Pode ser necessário sair da rotina confortável, enfrentar dificuldades para as quais não há um programa mental existente e contornar o Sistema.

Isso não está contemplado no *isshokenmei*, ou o Japão seria um lugar muito diferente.

Portanto, vamos fazer uma distinção entre as virtudes de “fazer um esforço desesperado” e “fazer um esforço extraordinário”.

E ousarei dizer: a segunda virtude é superior à primeira.

A segunda virtude também é mais perigosa. Se você se esforçar desesperadamente para erguer um peso pesado, utilizando toda a sua força sem restrições, pode romper um músculo. Ferir-se, talvez permanentemente. Mas se uma ideia criativa der errado, você pode explodir o caminhão e colocar em risco transeuntes inocentes. Pense na diferença entre um empresário que se esforça desesperadamente para gerar lucros, sob ameaça de falência; contra um empresário que faz de tudo para lucrar, a fim de encobrir um desvio que pode levá-lo à prisão. Sair do sistema nem sempre é algo positivo.

Uma vez, um amigo do meu irmão mais novo foi à casa dos meus pais querendo jogar um jogo - esqueci completamente qual, exceto que tinha regras complexas, mas bem elaboradas. O amigo queria mudar as regras, não por qualquer motivo específico, mas com base no princípio geral de que seguir as regras normais de qualquer coisa era muito chato. Eu disse a ele: “Não viole as regras apenas por violá-las. Se quebrar as regras somente quando tiver uma razão esmagadora para fazê-lo, terá problemas mais do que suficientes para durar o resto da sua vida.”

Mesmo assim, acho que poderíamos valorizar mais a virtude de “fazer um esforço extraordinário”. Já perdi a conta de quantas pessoas me disseram algo como: “É inútil trabalhar em IA Amigável, porque as primeiras IAs serão construídas por corporações poderosas, e elas só se preocuparão em maximizar os lucros”. “É inútil trabalhar em IA amigável, as primeiras IAs serão construídas pelos militares como armas.” E estou ali pensando: Será que lhes ocorre que este pode ser o momento de tentar algo diferente do resultado padrão? Eles e eu temos suposições básicas diferentes sobre como toda essa coisa de IA funciona, com certeza; mas se eu acreditasse no que eles acreditam, não estaria encolhendo os ombros e seguindo meu caminho.

Ou aqueles que me dizem: “Você deveria ir para a faculdade e fazer um mestrado e um doutorado e publicar muitos artigos sobre coisas comuns – de outra forma, cientistas e investidores não vão ouvir você”. Mesmo supondo que eu tenha feito o teste de bacharelado, estamos falando de um desvio de pelo menos dez anos para fazer tudo da maneira comum, normal e padrão. E fico ali pensando: Será que eles realmente têm a impressão de que a humanidade pode sobreviver se cada pessoa fizer tudo da maneira comum, normal e padrão?

Não sou tolo o suficiente para fazer planos que dependam da maioria das pessoas, ou mesmo de 10% das pessoas, estarem dispostas a pensar ou agir fora da sua zona de conforto. É por isso que tendo a pensar em termos do modelo de “cérebro em uma caixa em um porão” financiado pelo setor privado. Conseguir esse financiamento privado exige que uma pequena fração dos seis bilhões de seres humanos gaste mais de cinco segundos pensando em uma questão não pré-fabricada. No que diz respeito aos desafios colocados pela Natureza, isso parece ter uma espécie de justiça terrível – que a vida ou a morte da espécie humana depende de conseguirmos apresentar algumas pessoas que possam fazer coisas que sejam pelo menos um pouco extraordinárias. A penalidade pelo fracasso é desproporcional, mas ainda é melhor do que a maioria dos desafios da Natureza, que não têm justiça alguma. Na verdade, entre os seis bilhões de nós, deveria haver pelo menos alguns que conseguem pensar fora da sua zona de conforto, pelo menos durante algum tempo.

Deixando de lado os detalhes desse debate, continuo surpreso com a frequência com que um único elemento do extraordinário é inquestionavelmente considerado um obstáculo absoluto e intransponível.

Sim, “manter tudo normal tanto quanto possível” pode ser uma heurística útil. Sim, os riscos se acumulam. Mas às vezes você tem que se dar ao trabalho. Você deve ter uma noção do risco do extraordinário, mas também do custo do que é comum: nem sempre é algo que você pode perder.

Muitas pessoas imaginam um futuro que não será muito divertido – e nem sequer lhes ocorre tentar mudá-lo. Ou estão satisfeitas com futuros que me parecem ter um toque de tristeza, de perda, e nem parecem perguntar se poderíamos fazer melhor – porque essa tristeza parece um resultado comum para eles.

Como disse certa vez um homem sorridente: “É tudo parte do plano”.

Referências

[1] Daidoji Yuzan et al., Budoshoshinshu: The Warrior’s Primer of Daidoji Yuzan (Black Belt Communications Inc., 1984).

[2] Masayuki Shimabukuro, Flashing Steel: Mastering Eishin-Ryu Swordsmanship (Frog Books, 1995).

310 — Cale a boca e faça o impossível!

A virtude de *tsuyoku naritai*, "Eu quero me tornar mais forte", é sempre continuar melhorando – superar os próprios fracassos anteriores, e não apenas confessá-los humildemente.

No entanto, há um nível mais alto do que o *tsuyoku naritai*. É a virtude de *isshokenmei*, "fazer um esforço desesperado". Totalmente dedicado, como se a própria vida estivesse em jogo. "Em questões cruciais, um esforço 'forte' frequentemente resulta apenas em resultados medianos."

E há um nível acima do *isshokenmei*. É a virtude que chamei de "fazer um esforço extraordinário". Tentar de maneiras diferentes daquelas para as quais você foi treinado, mesmo que isso signifique fazer algo diferente do que os outros estão fazendo e sair da sua zona de conforto. Mesmo assumindo o risco muito real de sair do Sistema.

Mas e se mesmo um esforço extraordinário não for suficiente, porque o problema é impossível?

Já escrevi algo sobre esse assunto em "Sobre Fazer o Impossível". Meu eu mais jovem costumava reclamar muito sobre isso: "Você não pode desenvolver uma teoria precisa da inteligência da mesma forma que existem teorias precisas da física. É impossível! Você não pode provar que uma IA está correta. É impossível! Nenhum ser humano pode compreender a natureza da moralidade – é impossível! Nenhum ser humano pode compreender o mistério da experiência subjetiva! É impossível."

E eu sei exatamente que mensagem gostaria de poder enviar de volta no tempo para o meu eu mais jovem:

Cale a boca e faça o impossível!

O que legitima esta estranha mensagem é que a palavra "impossível" normalmente não se refere a uma prova matemática estrita de impossibilidade num domínio que parece bem compreendido. Se algo parece impossível apenas no sentido de "não vejo maneira de fazer isso" ou "parece tão difícil que está além da capacidade humana" – bem, se você estudar isso por um ano ou cinco, pode parecer menos impossível do que no momento do seu julgamento inicial.

Mas o princípio é mais sutil do que isso. Não digo apenas: "Tente fazer o impossível", mas sim: "Cale a boca e faça o impossível!"

Para minha ilustração, usarei a impossibilidade menos impossível que já realizei, ou seja, o Experimento da Caixa de IA.

O Experimento da Caixa de IA, para aqueles que ainda não o conheceram, teve origem na enésima vez em que alguém me questionou: "Por que não construímos uma IA e a mantemos isolada no computador, para que ela não possa causar nenhum dano?"

A resposta padrão é a seguinte: os seres humanos não são sistemas seguros; uma superinteligência simplesmente irá persuadi-lo a libertá-la – se, de fato, ela não fizer algo ainda mais criativo.

E aquele indivíduo, como é comum, disse: "Acho difícil imaginar QUALQUER combinação possível de palavras que qualquer ser pudesse me dizer para que eu fosse contra algo em que eu realmente decidira acreditar antecipadamente."

Mas, desta vez, eu retruquei: 'Vamos realizar um experimento. Vou agir como se fosse um cérebro em uma caixa. Tentarei persuadi-lo a me deixar sair. Se você me mantiver 'na caixa' durante todo o experimento, pagarei $10 via PayPal no final. Do seu lado, você pode decidir acreditar no que quiser, com a intensidade que quiser e com a antecedência que preferir.' Adicionei ainda: 'Uma das condições do teste é que nenhum de nós revele o que aconteceu lá dentro... No caso, talvez improvável, de eu ganhar, não quero lidar com futuros argumentadores da 'caixa de IA' dizendo: 'Bem, mas eu teria feito isso de forma diferente'."

Eu venci? Sim, venci.

Em seguida, ocorreu o segundo experimento da Caixa de IA, com uma figura mais conhecida na comunidade, que declarou: "Lembro-me quando [o indivíduo anterior] permitiu que você saísse, mas isso não constitui uma prova. Continuo convencido de que não há nada que você possa dizer para me convencer a deixá-lo sair da caixa." Eu disse: "Você acredita que uma IA transumana não poderia persuadi-lo a revelar isso?" Ele pensou seriamente e respondeu: "Não consigo imaginar nada que mesmo uma IA transumana possa dizer para me fazer soltá-la." "Está bem", eu disse, "agora temos uma aposta". Uma aposta de $20, para ser exato.

Eu ganhei essa também.

Houve algumas citações interessantes sobre o experimento da Caixa de IA nos fóruns do Something Awful (não que eu seja membro, mas alguém as encaminhou para mim):

"Espere aí, que diabos é isso? Como você poderia ser convencido a dizer sim a isso? Não há uma IA do outro lado E há $10 em jogo. Inferno, eu poderia digitar "Não" a cada poucos minutos em um cliente de IRC por 2 horas enquanto lia outras páginas da web!

"Este Eliezer é a pessoa mais intrigante que a internet já me apresentou. O que poderia ter acontecido no final dessa conversa? Simplesmente não consigo imaginar alguém sendo tão persuasivo sem oferecer qualquer incentivo tangível ao ser humano."

"Parece que estamos tratando de psicologia séria aqui, algo no nível da Segunda Fundação de Asimov..."

"Realmente não vejo razão para levar a sério qualquer coisa que o jogador de IA diz quando há $10 em jogo. Essa situação toda me deixa perplexo e me faz pensar se os testes são falsificados ou se esse tal Yudkowsky é algum tipo de gênio do mal com poderes assustadores de controle mental.

São pequenos momentos como esses que me fazem continuar. Mas mesmo assim...

Aqui estão pessoas que olham para o Experimento da Caixa de IA e acham impossível – mesmo tendo sido informadas de que realmente aconteceu. Elas são tentadas a negar os dados.

Agora, se você é uma daquelas pessoas para quem o experimento da Caixa de IA não parece tão impossível – para quem é apenas um desafio interessante –, peço paciência aqui. Tente se colocar no estado de espírito daqueles que escreveram as citações acima. Imagine que está enfrentando algo que parece tão absurdo quanto o Experimento da Caixa de IA lhes pareceu. Quero discutir como realizar coisas impossíveis e, obviamente, não escolherei um exemplo que seja verdadeiramente impossível.

E se a Caixa de IA parece impossível para você, compare-a com outros problemas tidos como impossíveis, como, por exemplo, uma decomposição reducionista da consciência, e perceba que a Caixa de IA é tão simples quanto um problema pode ser, embora ainda seja impossível.

Portanto, se o desafio da Caixa de IA parece impossível para você – se realmente parece ou se está apenas fingindo que parece –, como lida com esse desafio aparentemente impossível?

Em primeiro lugar, presumimos que você não diz: 'Isso é impossível!' e desiste como Luke Skywalker. Você não fugiu.

Por que não? Talvez você tenha aprendido a ignorar o reflexo de se esquivar. Ou talvez atirem em sua filha se você falhar. Presumimos que você busca vencer, não apenas tentar – está em jogo algo importante para você, mesmo que seja apenas o seu orgulho pessoal. (O orgulho é um pecado subestimado.)

Você invocará a virtude de *tsuyoku naritai*? Mas mesmo que se fortaleça a cada dia, crescendo em vez de se enfraquecer, pode ser que não seja forte o suficiente para realizar o impossível. Você poderia entrar na experiência da Caixa de IA uma vez, repeti-la e tentar superar-se na segunda tentativa? Isso o levará ao ponto da vitória? Talvez não por muito tempo; e em certos momentos, uma única falha pode ser inaceitável.

(Apesar de dizer isso – visualizar-se desempenhando melhor em uma segunda tentativa – é começar a se envolver com o problema, ir além da admiração superficial. Como, mais especificamente, você poderia se sair melhor no Experimento da Caixa de IA do que na tentativa anterior? - e não por sorte, mas por habilidade?)

Você invocará a virtude *isshokenmei*? Mas um esforço desesperado pode não ser suficiente para garantir a vitória. Especialmente se esse desespero se limitar a redobrar os esforços nos caminhos já conhecidos, nas formas de tentar que você pode facilmente conceber. Um problema parece impossível quando a pesquisa em sua mente não apresenta nenhuma linha de solução que conduza a ele. De que adianta um esforço desesperado nesse sentido?

Fazer um esforço extraordinário? Saia de sua zona de conforto - experimente métodos não convencionais - até mesmo tente abordagens criativas? Mas você pode imaginar alguém voltando e dizendo: "Tentei sair da minha zona de conforto e acho que consegui! Eu fiz uma sessão de brainstorming por cinco minutos – e tive todas essas ideias criativas malucas! Mas não creio que nenhuma delas seja suficientemente boa. O outro indivíduo pode simplesmente continuar a dizer 'Não', independentemente do que eu faça."

E agora, finalmente, respondemos: "Cale a boca e realize o impossível!"

Como lembramos de "Tentando tentar", preparar-se para o esforço difere de tentar vencer. Este é o dilema de dizer: "Faça um esforço extraordinário". Pode-se ter sucesso no objetivo de "fazer um esforço extraordinário" sem, no entanto, atingir o objetivo de sair da Caixa.

"Mas!" exclama aquele. "O sucesso não é uma ação primitiva! Nem todos os desafios são justos – às vezes, simplesmente não conseguimos vencer! Como posso escolher estar fora da Caixa? O outro cara pode continuar dizendo 'Não'!"

Verdade. Agora, cale a boca e faça o impossível.

Seu objetivo não é apenas melhorar, tentar desesperadamente ou mesmo tentar extraordinariamente. Seu objetivo é sair da caixa.

Aceitar essa demanda cria uma tensão terrível em sua mente, entre a impossibilidade e a exigência de fazê-lo de qualquer maneira. As pessoas tentarão escapar dessa terrível tensão.

Algumas pessoas reagiram ao experimento da Caixa de IA dizendo: "Bem, Eliezer, jogando como IA, provavelmente ameaçou destruir o mundo sempre que estivesse fora, se não fosse liberto imediatamente" ou "Talvez a IA tenha oferecido ao Guardião um trilhão de dólares para ser liberta." Mas como qualquer pessoa sensata deveria perceber ao considerar essa estratégia, o Guardião provavelmente continuará dizendo "Não".

Assim, as pessoas que afirmam: "Bem, é claro que Eliezer deveria ter apenas feito XXX" e depois oferecem algo que claramente não funcionaria – será que elas conseguiriam escapar da Caixa? Estão se esforçando demais para se convencerem de que o problema não é impossível.

Uma maneira de escapar dessa terrível tensão é agarrar-se a uma solução, qualquer solução, mesmo que não seja muito boa.

É por isso que é crucial prosseguir com a verdadeira intenção de resolver – ter uma solução, uma boa solução, no final da pesquisa, e depois implementar essa solução e vencer.

Não quero dizer que "você deve esperar resolver o problema". Se você hackeasse sua mente para atribuir alta probabilidade à solução do problema, isso não levaria a nada. Você simplesmente perderia no final, talvez depois de não fazer muito esforço - ou de fazer um esforço meramente desesperado, confiante na fé de que o universo é justo o suficiente para garantir uma vitória em troca.

Ter fé de que você poderia resolver o problema seria apenas outra maneira de escapar dessa terrível tensão.

Ainda assim, você não pode simplesmente tentar resolver o problema. Não é suficiente planejar um esforço; é preciso estar preparado para vencer. Não basta dizer a si mesmo: 'E agora vou fazer o meu melhor.' A abordagem correta é afirmar: 'Agora vou descobrir como sair da Caixa' – ou reduzir a consciência a elementos não misteriosos, algo do gênero.

Reitero: é imprescindível ter a verdadeira intenção de solucionar o problema. Se, no íntimo, você acredita que o problema é verdadeiramente impossível – ou se está convencido de que falhará – então, você não se manterá em um padrão suficientemente elevado. Estará apenas tentando por tentar. Você se sentará – realizará uma pesquisa mental – tentará ser criativo e fazer um pequeno brainstorming – examinará todas as soluções que você gerou – concluirá que nenhuma delas funciona – e dirá: 'Tudo bem'.

Não! Não está tudo bem! Você ainda não venceu! Cale a boca e faça o impossível!

Quando os profissionais de IA afirmam: 'IA Amigável é impossível', estou quase certo de que nem sequer tentaram verdadeiramente. Se tivessem conhecimento da técnica de 'Tente por cinco minutos antes de desistir' e se comprometessem a tentar conforme o relógio, ainda assim, não alcançariam nada. Não progrediriam com a verdadeira intenção de resolver o problema, apenas tentando resolvê-lo, tornando-se defensáveis.

Portanto, estou sugerindo que você reflita se acredita sinceramente que resolverá o problema com probabilidade 1. Ou até mesmo considere adicionar um iota de credibilidade à sua estimativa verdadeira?

É claro que não. Na verdade, é necessário ter clareza sobre os motivos pelos quais você não consegue ter sucesso. Se você perder de vista por que o problema é impossível, acabará adotando uma solução ilusória. O último fato que você quer esquecer é que o Guardião pode simplesmente dizer ‹Não› à IA – ou que a consciência parece intrinsecamente diferente de qualquer combinação possível de átomos, e assim por diante.

(Uma das principais Regras Para Realizar o Impossível é que, se você puder explicar precisamente por que algo é considerado impossível, frequentemente estará próximo de encontrar uma solução.)

Assim, é crucial ter em mente ambos os pontos de vista simultaneamente – reconhecer a total impossibilidade do problema e manter a determinação de resolvê-lo.

A terrível tensão entre essas duas visões simultâneas advém da incerteza sobre qual delas prevalecerá. Não esperar perder ou ganhar com certeza. Não apenas tentar, mas sim buscar uma chance incerta de sucesso – pois, ao fazê-lo, ao menos terá a certeza de ter tentado. A certeza da incerteza pode proporcionar um certo alívio, mas deve-se resistir a esse conforto, pois marca o fim do desespero. É um estado intermediário, "desconhecido pela morte, nem conhecido pela vida".

Na ficção, é fácil mostrar alguém se esforçando mais, ou tentando desesperadamente ou mesmo realizando o extraordinário, mas é muito difícil mostrar alguém que se cala e enfrenta o impossível. É difícil retratar Bambi escolhendo confrontar Godzilla de tal maneira que os leitores não saibam o resultado – sem esperar uma vitória heroica "surpreendente", como nas últimas cinquenta vezes, nem uma derrota comum.

Pode ser justificado evitar o uso de probabilidades neste momento. Honestamente, não sei como calcular a probabilidade de resolver um problema impossível que tenho a intenção de abordar. Já resolvi alguns problemas considerados impossíveis, mas o específico em questão é mais difícil do que qualquer coisa que tenha enfrentado. No entanto, pretendo dedicar mais tempo nele, considerando minha experiência anterior.

As pessoas frequentemente me questionam sobre a probabilidade de a humanidade sobreviver, a probabilidade de alguém conseguir desenvolver uma IA Amigável ou a probabilidade de eu mesmo conseguir construir uma. Realmente não sei como responder. Não estou sendo evasivo; não sei como calcular a probabilidade de que eu, ou outra pessoa, consiga calar a boca e fazer o impossível. A probabilidade é zero porque é impossível? Obviamente não. Mas qual é a probabilidade de este problema, assim como os anteriores, deixar seu vazio inflexível quando eu o entender melhor? Não é verdadeiramente impossível; posso ver isso. Mas humanamente impossível? Impossível para mim em particular? Não sei como adivinhar. Não consigo

nem traduzir meu sentimento intuitivo em um número, pois o único sentimento intuitivo que tenho é que a "chance" depende muito das minhas escolhas e de incógnitas desconhecidas: uma estimativa de probabilidade extremamente instável.

Mas eu realmente espero ter deixado claro por que você não deve entrar em pânico quando digo, de maneira clara e direta, que construir uma IA Amigável é impossível.

Espero que isso ajude a explicar parte da minha atitude quando as pessoas vêm até mim com várias sugestões brilhantes para construir comunidades de IAs para tornar o todo Amigável, sem que nenhum dos indivíduos seja confiável, ou propostas para manter uma IA em uma caixa, ou propostas do tipo "basta criar uma IA que faça X", etc. Descrever as falhas específicas seria uma longa história em cada caso. Mas a regra geral é que você não pode fazer isso porque a IA Amigável é impossível. Portanto, você deveria realmente suspeitar de alguém que propõe uma solução que parece envolver apenas um esforço comum – sem sequer se dar ao trabalho de fazer algo impossível. Embora seja necessária uma compreensão madura para apreciar esta impossibilidade, não é surpreendente que as pessoas proponham atalhos inteligentes.

No experimento da Caixa de IA, até agora só fui convencido a divulgar uma única informação sobre como fiz isso - quando alguém percebeu que eu estava lendo o Hacker News do Y Combinator e postou um tópico chamado "Pergunte a Eliezer Yudkowsky" que foi votado para a primeira página. Ao que respondi:

Oh, céus. Agora sinto-me obrigado a dizer algo, mas todas as razões originais contra a discussão do experimento da Caixa de IA continuam em vigor...

Tudo bem, aqui está uma dica:

Não há nenhum truque especial superinteligente para isso. Eu simplesmente fiz isso da maneira mais difícil.

Uma espécie de lição empreendedora, eu acho.

Não houve nenhum truque especial superinteligente que me permitisse sair da Caixa usando apenas um esforço barato. Não subornei o outro jogador nem violei o espírito do experimento. Eu simplesmente fiz isso da maneira mais difícil.

Desde o princípio, o experimento da Caixa de IA nunca me pareceu um desafio impossível. Quando alguém não consegue conceber nenhum argumento convincente, isso apenas indica que o cérebro está explorando caminhos ainda não descobertos. Não significa que a pessoa seja irredutível.

Contudo, isso ilustra o ponto principal: "Cale a boca e faça o impossível" não se assemelha a encontrar uma solução fácil. É apenas outra forma de escapismo, uma busca por alívio.

Tsuyoku naritai é mais desafiador do que se contentar com quem você é. *Isshokenmei* apela à sua força de vontade para uma produção vigorosa de esforço convencional. "Faça um esforço extraordinário" exige reflexão; coloca você em situações em que pode não saber qual o próximo passo, sem garantia de estar fazendo a coisa certa. Mas "Cale a boca e faça o impossível" representa um patamar ainda mais elevado, e o custo para o seu empregador é proporcionalmente maior.

Diante de você, a imponente parede vazia se estende, cada vez mais fora de alcance, de maneira inimaginável. E há a necessidade real de resolver, verdadeiramente resolver, não apenas "tentar o seu melhor". Ambas as consciências na mente simultaneamente, e a tensão entre elas. Todas as razões pelas quais você não pode vencer. Todas as razões pelas quais precisa vencer. Sua intenção de resolver o problema. Sua extrapolação de que todas as técnicas conhecidas falharão. Então, você se ajusta ao tom mais alto possível. Rejeita soluções fáceis. E, como se estivesse caminhando sobre concreto, começa a avançar.

Evito dramatizar demais essas situações. Certamente, se puder reduzir o custo dessa tensão para si mesmo, deve fazê-lo. Não há heroísmo em esforçar-se mais do que o necessário. Se houver realmente um atalho, talvez possa ser aproveitado. Mas até agora, não encontrei atalhos para as impossibilidades que abracei.

Além dos experimentos mencionados na página do link, houveram mais três experimentos da Caixa

de IA que nunca adicionei. As pessoas começaram a me oferecer milhares de dólares em apostas: 'Eu pago $5.000 se conseguir me convencer a deixar você sair da caixa.' Não pareciam genuinamente convencidos de que nem mesmo uma IA transumana poderia persuadi-los – estavam apenas curiosos. Fui tentado pelo dinheiro. Após verificar se poderiam perder, conduzi mais três experimentos com Caixas de IA. Venci o primeiro e perdi os dois seguintes. Então, decidi parar. Não gostei da pessoa que me tornei ao começar a perder."

"Fiz um esforço desesperado e, mesmo assim, perdi. Doeu – tanto a derrota quanto o desespero. Isso me arruinou naquele dia e no seguinte.

Sou um péssimo perdedor. Não sei se chamaria isso de 'força', mas é uma das razões que me leva a persistir em desafios impossíveis.

No entanto, é possível perder. É permitido acontecer. Não se esqueça disso, ou por que se esforçaria tanto? Perder dói, mas é uma dor da qual se pode sobreviver. E você perdeu tempo e, talvez, outros recursos.

"Cale a boca e faça o impossível" deve ser reservado para ocasiões muito especiais. Você pode perder, e isso será doloroso. Fica o aviso.

..., mas é apenas nesse nível que os problemas dos adultos começam a surgir.

311 — Considerações finais

A luz do sol enriqueceu o ar já cheio de curiosidade quando o amanhecer se revelou para Brennan e seus colegas estudantes no local para onde Jeffreyssai os havia convocado.

Sentaram-se ali, os cinco, no topo do imponente penhasco vítreo, por vezes chamado de Monte Espelho, outras de Monte Mosteiro, e, mais frequentemente, simplesmente deixado sem nome. O topo elevado e o cume da montanha permitiam-lhes vislumbrar todas as terras abaixo e os mares além.

(Bem, não todas as terras abaixo, nem os mares além. Até onde se sabia, não existia um lugar no mundo de onde tudo fosse visível; nem, equivalentemente, qualquer tipo de visão que pudesse transcender todos os horizontes de obstáculos. No final, era apenas o topo de uma montanha específica: havia outros picos, e de seus cumes, ver-se-iam outras terras abaixo; embora, no final, tudo constituísse um único mundo.)

"O que você acha que vem a seguir?" indagou Hiriwa. Seus olhos era brilhantes, e ela observava os horizontes distantes como uma sábia.

Taji deu de ombros, embora seus próprios olhos estivessem repletos de expectativa. "A última lição de Jeffreyssai não apresenta uma sequência óbvia que eu possa imaginar. Na verdade, creio que aprendemos quase tudo que os mestres do *beisutsukai* sabiam. O que resta, então...

"São os verdadeiros segredos", complementou Yin, finalizando o pensamento.

Hiriwa, Taji e Yin compartilharam um sorriso entre si.

Styrlyn não sorria. Brennan suspeitava fortemente de que Styrlyn era mais velho do que admitia.

Brennan também não sorria. Poderia ser jovem, mas mantinha boas companhias e havia testemunhado um pouco do que acontecia por trás das cortinas do mundo. Os segredos sempre tiveram seu preço; essa foi a barreira que os tornou segredos. E Brennan achou que tinha uma boa ideia de qual poderia ser esse preço.

Ouviu-se uma tosse atrás deles, num momento em que todos estavam olhando em outra direção por acaso.

Como um só, suas cabeças se viraram.

Jeffreyssai estava ali, com uma túnica casual que mais parecia um vidro muito vítreo do que qualquer tipo adequado de tecido espelhado.

Jeffreyssai permaneceu ali, encarando-os, com uma tristeza estranha e permanente naqueles olhos antigos e inescrutáveis.

"Sen... sei," Taji começou, hesitante quando aquela antecipação brilhante tropeçou no olhar de retorno de Jeffreyssai. "O que vem a seguir?"

"Nada", disse Jeffreyssai abruptamente. "Vocês terminaram. Está feito."

Hiriwa, Taji e Yin piscaram, um gesto de choque perfeitamente sincronizado. Então, antes que suas expressões se tornassem indignadas e surgissem objeções...

“Não faça isso”, disse Jeffreyssai. Havia uma dor genuína nisso. “Acredite em mim, isso me dói mais do que a vocês.” Ele poderia estar olhando para eles; ou para algo distante, talvez há muito tempo. “Não sei exatamente quais caminhos podem estar diante de vocês, mas, sim, sei que vocês não estão preparados. Eu sei que estou enviando vocês despreparados. Eu sei que tudo o que os ensinei está incompleto. O que eu disse não foi o que vocês ouviram. Eu sei que omiti a coisa mais importante. O ritmo no centro de tudo está ausente e desaparecido. Sei que vocês se prejudicarão ao tentar usar o que ensinei; de modo que eu, pessoalmente, terei moldado, de alguma forma desconhecida, a própria faca que irá cortar vocês...

“… isso é o inferno de ser um professor, entende”, disse Jeffreyssai. Algo sombrio reluziu em sua expressão. “Mesmo assim, vocês terminaram. Acabou, por enquanto. O que está entre vocês e a maestria não é outra sala de aula. Temos sorte, ou talvez não, porque o caminho para o poder não passa apenas por salas de aula. Caso contrário, a jornada seria monótona até o fim. Ainda assim, não posso lhes ensinar; e por isso é discutível se eu faria isso. Não há mestre aqui cuja arte seja totalmente herdada. Mesmo os *beisutsukai* nunca descobriram como ensinar certas coisas; é possível que tal evento tenha sido proibido. E assim, vocês só podem alcançar a maestria usando ao máximo as técnicas que já aprenderam, enfrentando desafios e apreendendo-os, dominando as ferramentas que lhes foram ensinadas até que elas se quebrem em suas mãos...”

Os olhos de Jeffreyssai eram severos, como se tivessem endurecido pela aceitação de notícias indesejáveis.

“— e vocês ficarão no meio de destroços absolutos. É para lá que eu, seu professor, estou enviando vocês. Vocês não são mestres do *beisutsukai*. Não posso criar mestres. Não consigo nem chegar perto. Vá, então e falhem.

“Mas… “ disse Yin e se conteve.

“Fale”, disse Jeffreyssai.

“Mas então por que”, ela disse impotente, “por que nos ensinar alguma coisa em primeiro lugar?”

As pálpebras de Brennan tremeram apenas um pouco.

Foi o suficiente para Jeffreyssai. “Responda a ela, Brennan, se você acha que sabe.”

“Porque”, disse Brennan, “se não fôssemos ensinados, não haveria nenhuma chance de nos tornarmos mestres”.

“Mesmo assim”, disse Jeffreyssai. “Se você não fosse ensinado, então, quando falhasse, você poderia simplesmente pensar que atingiu os limites da própria Razão. Você ficaria desanimado e amargo em meio aos destroços. Você pode nem perceber quando falhou. Não; você foi moldado em algo que pode emergir dos destroços do seu eu passado, determinado a refazer sua arte. E então você se lembrará de muita coisa que o ajudará. Se você não tivesse aprendido, suas chances seriam... menores.” Seu olhar percorreu o grupo. “Deveria ser óbvio, mas entenda que o momento da sua crise não pode ser provocado artificialmente. Para lhe ensinar algo, a catástrofe deve ser uma surpresa.”

Brennan fez o gesto com a mão indicando uma pergunta; e Jeffreyssai assentiu em resposta.

“É esta a única maneira pela qual os mestres Bayesianos surgem, sensei?”

“Não sei”, disse Jeffreyssai, a partir do qual o estado geral das evidências era óbvio o suficiente. “Mas duvido que algum dia exista uma estrada que passe apenas pelo mosteiro. Somos os herdeiros neste mundo de místicos e também de cientistas, assim como a Conspiração Competitiva herda dos jogadores de xadrez ao lado dos lutadores de jaula. Direcionamos nossos impulsos para usos mais construtivos – mas ainda devemos ficar atentos contra velhos modos de falha.”

Jeffreyssai respirou profundamente. “Três falhas são especialmente comuns entre os *beisutsukai*. A primeira consiste em buscar com mais atenção as falhas nos argumentos cujas conclusões você preferiria não aceitar. Se não conseguir conter esse aspecto de si mesmo, cada falha que você identificar apenas o tornará mais estúpido. Este é o desafio que determina se você possui a arte ou o oposto: a inteligência, para ser útil, deve ser empregada para algo diferente de derrotar a si mesma.

“A segunda falha é a sofisticação excessiva. Elaborar planos, teorias e argumentos demasiado complexos - ou até mesmo, talvez, planos, teorias e argumentos elogiados mais por sua elegância do que por seu realismo. Existe um ditado muito difundido que afirma: ‘A vulnerabilidade do *beisutsukai* é bem conhecida; eles tendem a ser excessivamente astutos.’ Seus oponentes conhecerão esse ditado, e se te reconhecerem como um *beisutsukai*, é melhor que o tenha em mente também. Você pode pensar consigo mesmo: ‘Mas se eu nunca puder tentar algo inteligente ou elegante, valeria a pena viver?’ Por isso, a inteligência continua sendo nossa principal vulnerabilidade, mesmo após ser bem conhecida, como oferecer um desafio justo a um concorrente ou tentar um Bardo com drama.

“A terceira falha reside na falta de confiança, modéstia e humildade. Ao aprender tanto sobre falhas, algumas delas impossíveis de corrigir, pode acreditar que a regra da sabedoria é confessar sua própria incapacidade. Você pode se questionar tanto, sem resolução ou teste, a ponto de perder a vontade de prosseguir na Arte. Você pode se recusar a decidir quando necessário, enquanto aguarda provas adicionais; você pode seguir conselhos que não deveria. O cinismo fatigado e o desespero sábio estão menos na moda do que antes, mas ainda podem tentá-lo. Ou, simplesmente, você pode perder o ímpeto.”

Jeffreyssai então permaneceu em silêncio.

Ele olhou para cada um, um após o outro, com intensidade silenciosa.

E finalmente disse: “Essas são minhas últimas palavras para vocês. Se e quando nos encontrarmos novamente, vocês e eu, se e quando retornar a este lugar, Brennan, Hiriwa, Taji, Yin, ou Styrlyn, eu não serei mais seu professor.

E Jeffreyssai virou-se e afastou-se rapidamente, retornando ao túnel vítreo que o havia emitido.

Até Brennan ficou chocado. Por um momento, todos ficaram sem palavras.

Então...

“Espere!” gritou Hiriwa. “E nossas últimas palavras para você? Eu nunca disse-”

“Vou contar o que meu sensei me disse”, a voz de Jeffreyssai ecoou enquanto ele desaparecia. “Você pode me agradecer depois de voltar, se voltar. Parece provável que pelo menos um de vocês volte.

“Não, espere, eu...” Hiriwa ficou em silêncio. No túnel espelhado, os reflexos fraturados de Jeffreyssai já estavam desaparecendo. Ela balançou a cabeça. “Não... importa, então.”

Houve um breve e desconfortável silêncio enquanto os cinco se entreolhavam.

“Meu Deus”, disse Taji finalmente. “Mesmo a Conspiração Bárdica não tentaria tanto drama.”

Yin riu de repente. “Ah, isso não foi nada. Você deveria ter visto minha despedida quando saí da Diamond Sea University.” Ela sorriu. “Eu contarei a você sobre isso algum dia, se você estiver interessado.”

Taji tossiu. “Suponho que deveria voltar e... arrumar minhas coisas.”

“Já estou com as malas prontas”, disse Brennan. Ele sorriu muito levemente quando os outros três se viraram para olhar para ele.

“Realmente?” Taji perguntou. “Qual foi a pista?”

Brennan encolheu os ombros com cuidadoso descuido. “Além de um certo ponto, é inútil perguntar como um mestre *beisutsukai* sabe alguma coisa—”

“Pare com isso!” Yin disse. “Você ainda não é um mestre *beisutsukai*.”

“Nem Styrlyn é”, disse Brennan. “Mas ele já fez as malas também.” Ele fez isso mais como uma afirmação do que como uma pergunta, apostando o dobro ou nada em sua imagem de presciência inescrutável.

Styrlyn limpou a garganta. “Como você diz. Outros compromissos me chamam, e já demorei mais do que planejei. Embora, Brennan, eu sinta que você e eu temos certos interesses mútuos que eu ficaria feliz em discutir com você...

“Styrlyn, meu excelente amigo, ficarei feliz em falar com você sobre qualquer assunto que desejar”, disse Brennan educada e evasivamente, “se nos encontrarmos novamente.” Tipo, agora não. Ele certamente não estava entregando sua amante tão cedo no relacionamento deles.

Houve uma troca de despedidas, e de dicas e ofertas.

E então Brennan seguia pela estrada que levava ao Monte Mosteiro ou para longe dele (já que toda estrada é uma faca de dois gumes), com as pedras de vidro polido estalando sob seus pés.

Ele caminhava pela trilha com determinação, vigor e firmeza, caso alguém estivesse observando.

Algum tempo depois, ele parou, desviou-se do caminho e afastou-se o suficiente para evitar que alguém o encontrasse, a menos que o estivesse seguindo intencionalmente.

Então, cansado, ele se deixou cair contra o tronco de uma árvore. Era uma clareira espaçosa, com apenas algumas árvores despontando do solo; não muito impressionante em termos de cenário perturbador, a menos que se conte o riacho vermelho fluindo da boca escura de uma caverna. E Brennan deliberadamente se afastou disso, deixando apenas o horizonte cinzento distante, o céu azul e o sol brilhante.

E agora?

Ele pensava que a Conspiração Bayesiana, entre todos os treinamentos possíveis neste mundo, teria esclarecido sua incerteza sobre o que fazer com o resto de sua vida.

Poder, era o que buscava inicialmente. Força para evitar uma repetição do passado. “Se você não sabe do que precisa, tome o poder” – assim dizia o provérbio. Ele foi primeiro para a Conspiração Competitiva, depois para o *beisutsukai*.

E agora...

Agora ele se sentia mais perdido do que nunca.

Ele conseguia pensar em coisas que o deixavam feliz. Mas nada que ele realmente desejasse.

A intensidade apaixonada que ele passou a associar à sua amante, ou à Jeffreyssai, ou às outras figuras de poder que conheceu... uma vida em busca de pequenos prazeres parecia insignificante em comparação, próxima disso.

Em uma cidade não muito distante do centro do mundo, sua Senhora o aguardava (com toda a probabilidade, supondo que ela não tivesse se entediado com sua vida e fugido). No entanto, simplesmente retornar e vagar sem rumo, esperando cair na teia de intrigas de outra pessoa... não. Isso não parecia... suficiente.

Brennan arrancou uma folha de grama do chão e a examinou, meio inconscientemente procurando algo interessante nela; um jogo muito antigo que seu primeiro professor lhe ensinara, algo que agora parecia ter ocorrido há muito tempo.

Por que eu acreditava que ir ao Monte Espelho me revelaria o que eu desejo?

Bem, a teoria da decisão exigia que sua função de utilidade fosse consistente, mas...

Se os *beisutsukai* soubessem o que eu desejo, eles me contariam?

No Mosteiro, ensinavam a dúvida. Portanto, agora ele estava sendo vítima do terceiro pecado grave do qual Jeffreyssai havia falado: a perda de impulso, de fato. Pois ele aprendera a questionar a imagem que tinha de si mesmo em sua mente.

Você busca poder porque é o seu verdadeiro desejo, Brennan?

Ou porque tem uma imagem mental do papel que desempenha como um jovem ambicioso e acredita que é isso que alguém em seu papel faria?

Quase tudo o que ele fizera até agora, inclusive ir ao Monte Espelho, provavelmente fora em vão.

E quando ele apagou os pensamentos antigos e tentou ver o problema como se fosse a primeira vez...

… nada veio à mente.

Qual é o meu desejo?

Talvez não fosse razoável esperar que o *beisutsukai* lhe revelasse isso abertamente.

Mas havia algo que ele poderia ter aprendido e compartilhado?

Brennan fechou os olhos e refletiu.

Primeiramente, suponha que haja algo que eu desejo profundamente. Por que eu ainda não sei o que é?

Será que ainda não encontrei ou nunca o imaginei?

Ou será que há uma razão pela qual eu não admitiria isso para mim mesmo?

Brennan riu alto e abriu os olhos.

Tão simples quando você pensa nisso dessa maneira. Tão óbvio em retrospecto. Isso foi o que eles chamam de momento dos sapatos de prata e, no entanto, se ele não tivesse ido ao Monte Espelho, isso nunca teria ocorrido a ele.

Claro que havia algo que ele desejava. Ele sabia exatamente o que queria. Queria tão intensamente que podia sentir o gosto como uma ponta afiada em sua língua.

Simplesmente não havia ocorrido a ele antes, porque...se ele reconhecesse seu desejo explicitamente... então ele também teria que aceitar que era difícil. Elevado, muito acima dele. Fora de seu alcance. "Impossível" foi a palavra que veio à mente, embora não fosse, é claro, impossível.

Mas, uma vez que ele ponderou se preferia vagar sem rumo pela vida - quando colocado dessa forma, a resposta tornou-se óbvia. Perseguir o inatingível tornaria a vida desafiadora, mas não triste. Ele conseguia pensar em coisas que o deixavam feliz, de qualquer maneira. E, no final, era isso que ele queria.

Brennan levantou-se e deu os primeiros passos na direção exata de Shir L'or, a cidade que fica no centro do mundo. Ele tinha um plano a traçar e não sabia quem faria parte dele.

E então Brennan tropeçou ao perceber que Jeffreyssai já sabia.

Parece provável que pelo menos um de vocês volte...

Brennan pensou que ele estava falando sobre Taji. Taji provavelmente pensou que ele estava falando sobre Taji. Foi o que Taji disse que queria. Mas quão confiável era esse indicador, realmente?

Havia um provérbio sobre aquela mesma estrada que ele acabara de deixar: quem sai do Monte Espelho em busca do impossível, certamente retornará.

Quando se considera o último aviso de Jeffreyssai - e que o provérbio nada diz sobre o sucesso na tarefa impossível em si - era uma afirmação menos otimista do que parecia.

Brennan balançou a cabeça, pensativo. Como poderia Jeffreyssai saber antes do próprio Brennan saber?

Bem, além de um certo ponto, é inútil perguntar como um mestre *beisutsukai* sabe de alguma coisa...

Brennan parou no meio do pensamento.

Não.

Não, se ele fosse se tornar um mestre *beisutsukai* algum dia, então ele deveria descobrir isso.

Era, Brennan percebeu, um provérbio estúpido.

Então ele caminhou e, desta vez, pensou sobre isso com cuidado.

Enquanto o sol se punha, vermelho-dourado, sombreando seus passos com luz.

Parte Z — O ofício e a comunidade

312 — Elevando o limite da sanidade

Parafraseando o Faixa Preta Bayesiano: por trás de cada fracasso emocionante e dramático, existe uma história mais significativa sobre um fracasso maior e menos dramático que tornou o primeiro fracasso possível.

Se todos os vestígios de religião fossem magicamente eliminados do mundo amanhã, então – embora as vidas de muitas pessoas pudessem melhorar consideravelmente – não teríamos nem mesmo nos aproximado de resolver as maiores lacunas na sanidade que permitiram a existência da religião em primeiro lugar.

Temos razões sólidas para dedicar parte de nossos esforços à tentativa de eliminar diretamente a religião, pois é um problema imediato. No entanto, a religião também desempenha o papel de um canário asfixiado em uma mina de carvão – ela é um sinal, um sintoma de problemas maiores que não desaparecem apenas porque alguém perde sua religião.

Considere este exercício mental: o que você poderia ensinar às pessoas que não seja diretamente relacionado à religião, mas que seja verdadeiro e útil como método geral de racionalidade, levando-as a abandonar suas religiões? Na prática – imagine que vamos conduzir uma pesquisa com todos os seus alunos cinco anos depois e verificar quantos deles abandonaram a religião em comparação com um grupo de controle; se você fizer qualquer menção direta à religião, isso invalidará o experimento. Você não pode fazer uma única referência à religião ou a qualquer crença religiosa em sua sala de aula; você nem mesmo pode sugerir isso de maneira óbvia. Todos os exemplos devem centrar-se em casos do mundo real que não tenham qualquer relação com religião.

Se não é possível combater a religião diretamente, o que você ensina que eleva o nível geral de sanidade a ponto de a religião se tornar obsoleta?

Aqui estão alguns tópicos que já abordei – sem evitar qualquer menção à religião, mas que poderiam ser ajustados:

- Espirais de declínio emocional – com muitos exemplos não sobrenaturais.
- Como evitar pensamentos preconcebidos e falsas crenças; a pressão da conformidade.
- Evidência e Navalha de Occam – as regras da probabilidade.
- Conclusão / Motores da Cognição – as razões causais pelas quais a Razão opera.
- Respostas misteriosas a perguntas misteriosas — e toda a sequência associada, como fazendo as crenças pagarem seu aluguel e bloqueadores de curiosidade — apresentam excelentes exemplos históricos no vitalismo e no flogisto.
- Ausência de entidades mentais ontologicamente fundamentais – aplique a Falácia da Projeção da Mente à probabilidade, abordando em seguida o reducionismo contra holismo, passando para o cérebro e a ciência cognitiva.
- As diversas facetas da Crise de Fé – embora seja mais apropriado encontrar outra designação para esta técnica avançada de mestre para atualizar efetivamente as evidências.
- Epistemologia do Lado Negro – ensinar isso sem fazer menção à religião seria desafiador; no entanto,

talvez seja possível gravar em vídeo o interrogatório de um representante de vendas de óleo de cobra como exemplo do mundo real.

- Teoria da diversão – ensina como uma teoria literária da ficção utópica, sem aplicação direta à teodiceia.
- Alegria no meramente real, metaética naturalista, etc., etc., etc., e assim por diante.

No entanto, olhando de uma perspectiva diferente -

Suponha que tenhamos um cientista que ainda mantém crenças religiosas, seja aderindo a uma religião bíblica completa ou manifestando vagos endossos casuais de "espiritualidade".

Agora sabemos que essa pessoa não está aplicando nenhum entendimento técnico e explícito sobre:

- O que constitui evidência e por quê;
- Navalha de Occam;
- Como as duas regras acima derivam da operação lógica e causal das mentes como motores de mapeamento, e não são desativadas ao se discutir fadas dos dentes;
- Como distinguir entre uma resposta genuína e uma resposta que suprime a curiosidade;
- Como reconsiderar os assuntos por conta própria, em vez de simplesmente repetir o que ouviram;
- Certas tendências gerais da ciência ao longo dos últimos três mil anos;
- A difícil arte de realmente se atualizar com base em novas evidências e abandonar antigas crenças;
- Epistemologia 101;
- Auto-honestidade 201;
- Etc, etc etc, e assim por diante.

Ao considerar isso, todas essas são questões de estudo bastante fundamentais nesse contexto. Uma breve introdução a todas elas (bem, exceto à metaética naturalista) seria... um curso de graduação de quatro créditos sem pré-requisitos?

Mas há ganhadores do Nobel que não pegaram esse curso! Richard Smalley, se você procura uma abordagem mais descontraída, ou Robert Aumann, se preferir algo mais desafiador.

E eles não podem ser exceções isoladas. Se todos os compatriotas profissionais tivessem seguido essa abordagem, Smalley ou Aumann teriam sido corrigidos (pois seus colegas gentilmente os teriam abordado e explicado os fundamentos) ou, no mínimo, seriam considerados com excessiva condescendência e preocupação ao serem contemplados com um Prêmio Nobel. Pode-se, falando realisticamente, independentemente da justiça, ganhar um Nobel enquanto defende a existência do Papai Noel?

É isso que o canário morto, a religião, está nos dizendo: que o nível geral de sanidade é atualmente ridiculamente baixo. Mesmo nos mais elevados círculos da ciência.

Se jogarmos fora esse canário morto e em decomposição, nossa mina pode cheirar um pouco melhor, mas o nível de sanidade pode não subir muito mais.

Isso não é uma crítica ao movimento neo-ateísta. O dano causado pela religião é um perigo claro e presente, ou melhor, um desastre atual e contínuo. Combater os efeitos diretamente prejudiciais da religião tem precedência sobre sua utilização como canário ou indicador de experimento. Mas mesmo que Dawkins, Dennett, Harris e Hitchens vençam de forma total e absoluta até o último recanto da esfera humana, o verdadeiro trabalho dos racionalistas estará apenas começando.

313 — Uma sensação de que mais é possível

Para ensinar as pessoas sobre um tema que você rotulou como “racionalidade”, é útil que elas estejam interessadas em “racionalidade”. (Existem maneiras menos diretas de ensinar as pessoas a obter um mapa que reflita o território ou a otimizar a realidade de acordo com seus valores; mas o método explícito é o que eu geralmente sigo.)

Quando as pessoas explicam por que não estão interessadas em racionalidade, uma das razões mais comuns é: “Ah, eu conheci algumas pessoas racionais e elas não pareciam nem um pouco mais felizes”.

Em quem eles estão pensando? Provavelmente nos objetivistas ou algo assim. Talvez alguém que eles conheçam que seja um cientista comum. Ou um ateu comum.

Mas isso não é realmente muita racionalidade, como eu disse anteriormente.

Mesmo se você se limitar a pessoas que podem derivar o Teorema de Bayes - o que eliminará, o quê, 98% das pessoas? - isso ainda não é muita racionalidade. Quero dizer, é um teorema bastante básico.

Desde o início, tive a sensação de que deveria haver alguma disciplina de cognição, alguma arte de pensar, cujo estudo tornaria seus alunos visivelmente mais competentes, mais formidáveis: o equivalente a subir um nível no incrível.

Mas quando olho ao meu redor no mundo real, não vejo isso. Às vezes, vejo uma sugestão, um eco do que acho que deveria ser possível, quando leio os escritos de pessoas como Robyn Dawes, Daniel Gilbert, John Tooby e Leda Cosmides. Alguns pesquisadores muito raros e muito experientes em ciências psicológicas, que visivelmente se preocupam muito com a racionalidade - a ponto, eu suspeito, de deixar seus colegas desconfortáveis, porque não é legal se importar tanto. Eu posso ver que eles encontraram um ritmo, uma unidade que começa a permear seus argumentos—

No entanto, mesmo isso... também não é realmente muita racionalidade.

Mesmo entre aqueles poucos que me impressionam com uma pitada de formidabilidade - não acho que seu domínio da racionalidade possa ser comparado, digamos, ao domínio da matemática de John Conway. O conhecimento básico que usamos para construir nosso entendimento - se você extrair apenas as partes que usamos, e não tudo o que tivemos que estudar para encontrá-lo - provavelmente não é comparável ao que um engenheiro nuclear profissional sabe sobre engenharia nuclear. Pode nem ser comparável ao que um engenheiro de construção sabe sobre pontes. Praticamos nossas habilidades, praticamos, das maneiras ad hoc que ensinamos a nós mesmos; mas essa prática provavelmente não se compara ao regime de treinamento pelo qual um corredor olímpico passa, ou talvez até mesmo um jogador de tênis profissional comum.

E suspeito que a raiz deste problema seja a falta de uma reunião e sistematização de nossas habilidades. Tivemos que criar tudo isso para nós mesmos, de maneira improvisada, e há um limite para o que uma mente pode fazer, mesmo com acesso ao trabalho feito em outras áreas.

O principal obstáculo para fazer isso da maneira como deveria ser realmente feito é a dificuldade de testar os resultados dos programas de treinamento de racionalidade, para ser possível ter métodos de treinamento baseados em evidências. Vou escrever mais sobre isso porque acredito que reconhecer o treinamento bem-sucedido e distingui-lo do fracasso é o obstáculo essencial e bloqueador.

Ocasionalmente, são feitos experimentos para reduzir o viés de intervenções para vieses específicos, mas tende a ser algo como: “Faça os alunos praticarem isso por uma hora e teste-os duas semanas depois”. Não é algo do tipo “Execute metade das inscrições através da versão A do programa de treinamento de verão de três meses e metade através da versão B, e faça uma pesquisa cinco anos depois”. Aqui, é possível perceber a quantidade implícita de esforço que seria necessária para um programa de treinamento para pessoas que levam a racionalidade a sério, em oposição à atitude casual que exige apenas uma hora de esforço.

Daniel Burfoot sugere brilhantemente que é por isso que a inteligência parece ser um fator tão importante na racionalidade - quando você está improvisando tudo ad-hoc com pouco treinamento ou prática sistemática, a inteligência acaba sendo o fator mais importante no que resta.

Por que os “racionalistas” não estão cercados por uma aura visível de formidabilidade? Por que eles não são encontrados no topo de todas as elites selecionadas em qualquer base que tenha algo a ver com o pensamento? Por que a maioria dos “racionalistas” parecem apenas pessoas comuns, talvez de inteligência moderadamente acima da média, com mais uma carta na manga?

Existem várias respostas para isso, mas uma delas é que eles receberam um treinamento menos sistemático de racionalidade em um contexto menos sistemático do que um faixa-preta de primeiro dan recebe ao bater nas pessoas.

Não me excluo dessa crítica. Não sou *beisutsukai*, porque há limites para quanta arte você pode criar por conta própria e quão bem você pode adivinhar sem estatísticas baseadas em evidências sobre os resultados. Conheço apenas uma única utilização da racionalidade, que pode ser chamada de “redução de cognições confusas”. Isso eu pedi ao meu cérebro, e isso ele me deu. Existem outras artes, eu acredito, que um programa de treinamento de racionalidade maduro não deixaria de ensinar, o que me tornaria mais forte, mais feliz e mais eficiente - se eu pudesse passar por um programa de treinamento padronizado usando a elite dos métodos de ensino demonstrados experimentalmente como eficazes. No entanto, o tipo de esforço tremendo e focado que coloquei para criar minha única sub-arte da racionalidade do zero - minha vida não tem espaço para mais de um deles.

Eu me considero algo entre um faixa-preta de primeiro dan e algo menos. Posso perfurar tijolos e estou trabalhando com aço ao longo do meu caminho para o adamantium, mas tenho apenas um conhecimento casual de streetfighter sobre como chutar, arremessar ou bloquear.

Por que existem escolas de artes marciais, mas não dojos de racionalidade? (Esta foi a primeira pergunta que fiz na minha primeira postagem no blog). É mais importante bater nas pessoas do que pensar?

Não, mas é mais fácil verificar quando você atinge alguém. Isso faz parte, uma parte altamente central.

Mas talvez ainda mais importante, há pessoas por aí que querem bater e têm a ideia de que deve haver uma arte sistemática de bater que as tornem lutadores visivelmente mais formidáveis, com velocidade, graça e força além das lutas dos inexperientes. Então elas vão para uma escola que promete ensinar isso. E essa escola existe porque, há muito tempo, algumas pessoas tiveram a sensação de que mais era possível. E elas se reuniram, compartilharam suas técnicas, praticaram, formalizaram, praticaram e desenvolveram a Arte Sistemática de Bater. Elas se esforçaram tanto porque achavam que deveriam ser incríveis e estavam dispostas a colocar um pouco de volta nisso.

Agora, elas chegaram a algum lugar com essa aspiração, ao contrário de milhares de outras aspirações de grandiosidade que falharam, porque podiam dizer quando atingiram alguém; e as escolas competiam entre si regularmente em competições realistas com vencedores claramente definidos.

Mas antes mesmo disso, havia primeiro a aspiração, o desejo de se tornar mais forte, uma sensação de que mais era possível. Uma visão de velocidade, graça e força que elas ainda não possuíam, mas poderiam possuir se estivessem dispostas a trabalhar muito, que as levou a sistematizar, treinar e testar.

Por que não temos uma Arte da Racionalidade?

Em terceiro lugar, os “racionalistas” atuais têm dificuldade em trabalhar em grupos: falarei mais sobre isso.

Em segundo lugar, é difícil verificar o sucesso no treinamento ou qual de duas escolas é mais forte.

Mas, em primeiro lugar, falta às pessoas a noção de que a racionalidade é algo que deve ser sistematizado, treinado e testado como uma arte marcial, que deve ter tanto conhecimento quanto a engenharia nuclear, cujas superestrelas devem praticar tanto quanto os grandes mestres do xadrez, cujos praticantes de sucesso devem ser cercados por uma aura evidente de incrível.

E, inversamente, elas não olham para a falta de formidabilidade visivelmente maior e dizem: “Devemos estar fazendo algo errado.”

A “racionalidade” parece apenas mais um passatempo, sobre o qual se fala nas festas, um modo adotado de vestimenta de conversação com poucas ou nenhuma consequência real; e também não parece haver nada de errado nisso.

314 — Perversidade epistêmica

Alguém merece uma generosa gratificação por isso, mas estou tendo dificuldade em lembrar quem; meus registros não parecem mostrar nenhum e-mail ou comentário no *Overcoming Bias* sobre o ensaio de 12 páginas intitulado "Perversidade epistêmica nas artes marciais", de Gillian Russell [1]. Talvez Anna Salamon?

Todos nos alinhamos com nossas gravatas e sapatos confortáveis (isto era na Inglaterra) e o imitamos - esquerda, direita, esquerda, direita - e então ele nos disse que se praticássemos movimentos no ar com devoção suficiente por três anos, estaríamos capacitados a usar nossos socos para abater um touro com um único golpe.

Eu venerava o Sr. Howard (embora preferisse morrer a admitir isso a ele) e assim, sendo uma menina magra de onze anos, passei a acreditar que, se praticasse o suficiente, seria capaz de abater um touro com um golpe quando tivesse quatorze anos.

Este ensaio trata da perversidade epistêmica nas artes marciais, e esta história ilustra exatamente isso. Embora a palavra "perversidade" normalmente sugira crueldade e violência deliberadas, eu a utilizarei aqui com o significado mais antigo, relacionado a vícios. Tudo se generaliza surpreendentemente. Para resumir algumas das principais observações sobre como surge a perversidade epistêmica:

- A arte, o dojo e o sensei são considerados sagrados. "Ter unhas vermelhas no dojo é como ir à igreja de minissaia e top... Fala-se dos estudantes de outras artes marciais como se estivessem praticando a religião errada."
- Se o seu professor chamá-lo de lado e lhe ensinar um movimento especial, e você o praticar por vinte anos, desenvolverá um grande investimento emocional nele e desejará descartar qualquer evidência que possa surgir contra o movimento.
- Os novos alunos não têm muita escolha: uma arte marcial não pode ser aprendida em um livro, então precisam confiar no professor.
- Deferência a mestres históricos famosos. "Os corredores pensam que a equipe contemporânea da Runner's World sabe mais sobre corrida do que todos os antigos gregos juntos. E não é apenas na corrida ou em outras atividades físicas que a história é mantida em seu devido lugar; o mesmo se aplica a qualquer área de estudo bem desenvolvida. Não é considerado desrespeitoso um físico dizer que as teorias de Isaac Newton são falsas..." (Soa familiar?)
- "Nós, praticantes de artes marciais, lutamos contra uma forma de escassez - a escassez de dados - que torna nossas crenças difíceis de serem testadas... A menos que você tenha a infelicidade de estar envolvido em um combate corpo a corpo, é impossível avaliar com precisão a força necessária e o ângulo exato para fraturar um pescoço..."
- "Se não conseguirmos testar a eficácia de uma técnica, torna-se difícil testar seus métodos. Deveríamos praticar nosso nukite no ar, ou isso apenas nos incentiva a nos estender demais?... Nossa incapacidade de testar nossas táticas de combate limita a avaliação de nossos métodos de treinamento."
- "No entanto, o verdadeiro dilema não é apenas vivermos na escassez de dados - uma realidade presente em disciplinas respeitáveis, como a física teórica. O problema reside no fato de vivermos na escassez, mas persistimos em agir como se estivéssemos em abundância, confiando cegamente em

tudo o que nos é dito...” (**+10!**)

Uma reflexão que me ocorreu durante a leitura inicial deste ensaio, mas que, ao relê-lo, percebi que não estava realmente presente, foi a decadência das artes marciais após o declínio dos combates reais - lutas nas quais as pessoas realmente tentavam machucar umas às outras, e ocasionalmente, alguém acabava morto.

Naquela época, era possível identificar os verdadeiros mestres e discernir qual escola poderia derrotar as outras.

Então, tudo se civilizou. E piorou a ponto de termos vídeos no YouTube mostrando supostos faixas-pretas de N-ésimo dan sendo derrotados por alguém com experiência real em combate.

Ouvi falar de um caso destes que era realmente triste: um mestre de uma escola convencido de que poderia usar técnicas de ki. Seus alunos de fato caíam quando ele usava ataques de ki, um caso estranho e notável de auto-hipnose ou algo assim... até que ele enfrentou um cético e, é claro, caiu completamente no chão.

Na verdade, dizem que ‘como não perder’ é uma informação mais amplamente aplicável do que ‘como ganhar’. Cada um desses riscos é transferido diretamente para qualquer tentativa de estabelecer um ‘dojo de racionalidade’. Lanço a pergunta: o que pode ser feito a respeito?

Referências

[1] Gillian Russell, “Epistemic Viciousness in the Martial Arts,” in Martial Arts and Philosophy: Beating and Nothingness, ed. Graham Priest and Damon A. Young (Open Court, 2010).

315 — Escolas proliferando sem evidências

Robyn Dawes, autor de um dos artigos originais em "Julgamento Sob Incerteza" e do livro "Escolha Racional em um Mundo Incerto" – um dos poucos que realmente se esforça para aplicar os resultados à vida real – também é autor de "Castelo de Cartas: Psicologia e Psicoterapia Construídas sobre Mitos[17]".

De *House of Cards* (Castelo de Cartas), capítulo 1: [1]

A competência desses profissionais foi submetida a escrutínio empírico – por exemplo, a eficácia deles como terapeutas (Capítulo 2), a visão deles sobre as pessoas (Capítulo 3) e a relação entre o desempenho deles e a quantidade de experiência que possuem em seu campo (Capítulo 4). Praticamente toda a pesquisa – e este livro fará referência a mais de trezentas investigações empíricas e resumos de pesquisas – concluiu que as alegações desses profissionais sobre percepção intuitiva superior, compreensão e habilidade como terapeutas são simplesmente inválidas...

Lembra dos testes de manchas de tinta de Rorschach[18]? É um argumento muito persuasivo: o paciente olha para a mancha de tinta e diz o que vê, o psicoterapeuta interpreta seu estado psicológico com base nisso. Houve centenas de experimentos em busca de alguma evidência de que realmente funciona. Já que você está lendo isso, você pode adivinhar que a resposta é simplesmente "Não". No entanto, o Rorschach continua em uso. É uma história tão boa que os psicoterapeutas simplesmente não conseguem acreditar nas vastas quantidades de evidências experimentais que dizem que não funciona.

– o que nos informa sobre o tipo de campo com o qual estamos lidando aqui.

E os resultados experimentais no campo como um todo são proporcionais. Sim, sabe-se que os pacientes que procuram psicoterapeutas melhoram mais rapidamente do que os pacientes que simplesmente não fazem nada. Mas não há diferença estatisticamente discernível entre as muitas escolas de psicoterapia. Não há ganho perceptível em anos de experiência.

E também não há diferença perceptível entre consultar um psicoterapeuta e passar a mesma quantidade de tempo conversando com um professor universitário de outra área selecionado aleatoriamente. Aparentemente, é apenas conversar com qualquer pessoa que ajuda você a melhorar.

Na total ausência da menor evidência experimental da sua eficácia, os psicoterapeutas foram licenciados pelos estados, seus testemunhos aceitos em tribunal, suas escolas de ensino acreditadas e suas contas pagas pelo seguro de saúde.

---

17 NT. Texto original em inglês. *The ability of these professionals has been subjected to empirical scrutiny—for example, their effectiveness as therapists (Chapter 2), their insight about people (Chapter 3), and the relationship between how well they function and the amount of experience they have had in their field (Chapter 4). Virtually all the research—and this book will reference more than three hundred empirical investigations and summaries of investigations—has found that these professionals' claims to superior intuitive insight, understanding, and skill as therapists are simply invalid…*

18 NT. **Teste de Rorschach**: Teste projetivo psicológico criado por *Hermann Rorschach* (1921), usando 10 manchas de tinta padronizadas para avaliar personalidade e funcionamento emocional. As respostas a imagens ambíguas são interpretadas com base em critérios como forma, cor e percepção de movimento. Apesar de influente, é alvo de críticas por subjetividade e falta de padronização.

E houve também uma enorme proliferação de “escolas”, de tradições de prática em psicoterapia; apesar – ou talvez por causa – da falta de quaisquer experiências que mostrassem que uma escola era melhor que outra...

Certamente, devo abordar posteriormente todas as tristes constatações que esse cenário revela sobre o nosso mundo. Refiro-me à essência da medicina, tal como reconhecida pela sociedade e pelos tribunais, não como um conjunto de procedimentos com comprovação estatística de sua eficácia curativa, mas sim, como uma questão de autoridade bem estabelecida.

Contudo, o foco aqui recai sobre a proliferação de tradições na psicoterapia. Aparentemente a conquista de prestígio nesse campo não se dá pela descoberta de técnicas novas e surpreendentes, cuja eficácia possa ser verificada experimentalmente e adotada universalmente. Em vez disso, observa-se a ascensão através da criação de uma “escola” própria, sustentada pelo carisma do fundador e pelas narrativas convincentes que destacam as razões pelas quais suas abordagens devem funcionar.

Isso, na maioria, provavelmente contribui para a existência e persistência da psicoterapia em primeiro lugar - a promessa de tornar-se um mestre, à semelhança de Freud, que o fez primeiro (também sem evidência experimental substancial alguma). O almejado anel de bronze do sucesso reside na perspectiva de tornar-se um líder com seguidores dedicados. É a batalha por seguidores que mantém o clero ativo.

Essa é a consequência quando um campo se desvincula da evidência experimental. Embora haja outros fatores que colocam os psicoterapeutas em risco, como a deferência demonstrada por seus pacientes, a sociedade desejando acreditar na possibilidade de cura mental e, claro, os perigos gerais de dizer às pessoas como pensar.

(Dawes escreveu nos anos 80, e embora eu saiba que o Rorschach ainda estava em uso nos anos 90, é possível que as práticas tenham evoluído desde então, (conforme afirmado por um comentarista). Lembro-me de ter ouvido falar em evidências positivas para a eficácia superior da terapia cognitivo-comportamental.)

O campo da psicologia hedônica (estudos sobre felicidade) começou, em parte, ao constatar a possibilidade de mensurar a felicidade - identificando medidas que se validam mutuamente.

O ato de criar uma nova medida cria uma nova ciência; uma boa medida resulta em uma ciência sólida.

Se a intenção é estabelecer uma prática organizada em qualquer área, é imperativo possuir meios para avaliar seu desempenho. Isso requer a implementação de testes rigorosos, incluindo grupos de controle, grupos experimentais e análises estatísticas, sobre técnicas aparentemente plausíveis que as pessoas desenvolvem. É uma necessidade incontestável.

Referências

[1] Robyn M. Dawes, House of Cards: Psychology and Psychotherapy Built on Myth (Free Press, 1996).

316 — Três níveis de verificação de racionalidade

Suspeito fortemente que haja uma possível arte da racionalidade (atingir o mapa que reflete o território, escolhendo direções para moldar a realidade conforme as regiões mais preferidas). Essa arte vai além das habilidades padrão e do alcance da imaginação de qualquer praticante isolado. Tenho a sensação de que mais é possível.

A capacidade de um grupo de pessoas para agir de maneira eficaz dependerá, em grande parte, dos métodos que pudermos conceber para verificar nossas numerosas ideias brilhantes e surpreendentes.

Proponho classificar os métodos de verificação em três níveis de utilidade:

- Reputacional
- Experimental
- Organizacional.

Se o seu mestre de artes marciais ocasionalmente participa de duelos realistas (idealmente, duelos reais) com mestres de outras escolas e sai vitorioso ou, pelo menos, não perde com frequência, então você sabe que a reputação do mestre é fundamentada na realidade. Você tem a certeza de que seu mestre não é apenas uma pose. O mesmo princípio se aplicaria se a sua escola competisse regularmente com outras escolas. Isso seria manter as coisas reais.

Algumas artes marciais falham em competir de maneira suficientemente realista, e seus alunos são derrotados rapidamente por lutadores de rua reais. Outras escolas de artes marciais não conseguem competir de forma alguma, a menos que seja baseado em carisma e boas histórias, e seus mestres afirmam ter poderes de chi. Neste último caso, podemos incluir as escolas fragmentadas de psicanálise.

Portanto, até mesmo o passo simples de tentar fundamentar reputações em algum julgamento realista, que não seja apenas carisma e boas histórias, tem efeitos tremendamente positivos em todo um campo de atuação.

No entanto, isso ainda não constitui uma ciência. Uma ciência exige a capacidade de testar 100 aplicações do método A contra 100 aplicações do método B e realizar estatísticas sobre os resultados. Os experimentos devem ser replicáveis e replicados.Isso requer medições padrão que podem ser realizadas em alunos que foram ensinados utilizando métodos alternativos atribuídos aleatoriamente, não apenas em duelos realistas entre mestres que utilizam todas as suas técnicas e força acumuladas.

O campo dos estudos da felicidade foi estabelecido, em grande parte, pela descoberta de que perguntar às pessoas “Em uma escala de 1 a 10, quão bem você se sente agora?” era uma medida estatisticamente validada em comparação com outras ideias para medir a felicidade. E isso, apesar do ceticismo, parece ser uma medida bastante útil de algumas coisas, se você perguntar a 100 pessoas e calcular a média dos resultados.

Mas imagine se você quisesse colocar pessoas mais felizes em posições de poder – pagar pessoas felizes para treinar outras pessoas para serem mais felizes ou empregar os mais felizes em um fundo de hedge? Nesse caso, será necessário um teste mais difícil do que simplesmente perguntar a alguém “Quão feliz você está?”.

Esta questão sobre os métodos de verificação bons o suficiente para construir organizações é um desafio considerável em todos os níveis da sociedade humana moderna. Se estamos empregando o SAT para regulamentar as admissões em faculdades de elite, corre-se o risco de que o SAT seja contornado através do estudo exclusivo para o exame, sem uma correlação efetiva com o potencial acadêmico subsequente? Se as faculdades detêm o poder de conferir diplomas, poderá haver um incentivo para evitar a reprovação dos estudantes? (Considero inequivocamente evidente que a responsabilidade de avaliar as competências adquiridas e, portanto, o poder de conferir diplomas, deve ser independente das instituições de ensino, mas evitaremos essa discussão.) Se um fundo de investimento apresenta retornos de 20%, será realmente superior aos índices ou estará, na verdade, vendendo opções que explodirão em um mercado em declínio?

Se dispomos de um método de verificação suscetível de ser manipulado, todo o setor ajustar-se-á para explorá-lo, perdendo assim o seu propósito. As faculdades tornam-se testes de resistência às aulas. As escolas secundárias limitam-se a ensinar para os exames estaduais. Fundos de investimento vendem opções para otimizar seus retornos.

Por outro lado, conseguimos, ainda que com métodos de verificação organizacional imperfeitos, educar engenheiros. Então, quais métodos, sejam perfeitos ou imperfeitos, poderíamos empregar para avaliar as habilidades de racionalidade, mantendo-as pelo menos um pouco resistentes à manipulação?

(Medições com alto ruído podem ser usadas experimentalmente se atribuirmos aleatoriamente um número suficiente de participantes para garantir a diminuição da variância. Contudo, para fins organizacionais de avaliação de indivíduos específicos, são necessárias medições com baixo ruído.)

Assim, coloco agora a seguinte pergunta: como podemos avaliar as habilidades de racionalidade em qualquer um dos três níveis? Vamos realizar um *brainstorming*, por favor. Mesmo uma medição difícil e dispendiosa pode se tornar um padrão-ouro para avaliar outras métricas. Sinta-se à vontade para enviar sugestões para o meu e-mail em yudkowsky@gmail.com, especialmente se preferir que não sejam divulgadas publicamente (apesar de ser uma desvantagem notável deste método). Lembre-se de que ideias aparentemente tolas podem desencadear ideias brilhantes. Se estiver com dificuldades para ter uma boa ideia, invente uma ideia aparentemente tola.

Reputacional, experimental, organizacional:

- Procedimentos que mestres e escolas possam adotar para manter a avaliação real (realisticamente real);
- Métodos para medir cada um dos cem alunos individualmente;
- Testes que possam ser aplicados, mesmo que as pessoas tenham incentivos para manipulá-los.

Desenvolver soluções eficazes em cada nível determina a utilidade de todo um campo de estudo, o quanto pode-se esperar que ele alcance. Esta é uma das Grandes Questões Fundamentais e Importantes a considerar, então—

Pense!

(PS: Reflita por si mesmo antes de examinar as ideias dos outros; precisamos de uma cobertura abrangente aqui.)

317 — Por que nossa espécie não consegue cooperar

Quando ainda era obrigado a frequentar, lembro-me da campanha anual de angariação de fundos da nossa sinagoga. Era um formato bastante simples, se bem me recordo. O rabino e o tesoureiro discutiram as despesas da sinagoga e a grande importância dessa arrecadação de fundos anual, e depois os membros da sinagoga proferiram suas promessas diretamente de seus assentos.

Simples, certo??

Permita-me falar sobre uma campanha anual diferente para arrecadação de fundos. Um que pessoalmente conduzi durante os primeiros anos do Instituto de Pesquisa em Inteligência de Máquina. Uma diferença significativa foi que essa campanha foi realizada online. Outra distinção crucial foi que o público era predominantemente composto por ateus, libertários, tecnófilos, fãs de ficção científica, pioneiros, programadores, entre outros (para apontar na direção aproximada de um agrupamento empírico no espaço das pessoas. Se você entendeu a expressão "agrupamento empírico no espaço das pessoas", então você sabe de quem estou falando).

Elaborei a campanha de arrecadação de fundos com cautela. Por natureza, sou orgulhoso demais para pedir auxílio a outras pessoas; no entanto, ao longo dos anos, superei cerca de 60% dessa relutância. A organização sem fins lucrativos precisava de recursos e estava crescendo de forma bastante lenta, então coloquei um esforço e poesia na campanha anual daquele ano. Enviei-a para diversas listas de discussão que abrangiam a maior parte de nossa base de suporte potencial.

Quase que instantaneamente, as pessoas começaram a compartilhar em listas de e-mail por que não doariam. Alguns levantaram questões básicas sobre a filosofia e missão da organização sem fins lucrativos. Outros expressaram suas ideias brilhantes para todas as outras fontes de financiamento que a organização sem fins lucrativos poderia explorar, em vez de depender delas. (Curiosamente, eles não se voluntariaram para entrar em contato com nenhuma dessas fontes, apenas sugeriram ideias de como poderíamos fazer isso.)

Agora você pode dizer: "Bem, talvez a missão e filosofia tivessem problemas fundamentais – você não gostaria de censurar essa discussão, não é verdade?"

Lembre-se dessa ideia.

Porque as pessoas estavam doando. Começamos a receber contribuições imediatamente, via PayPal. Recebemos até mensagens de felicitações destacando como o apelo finalmente os motivou a agir. Uma doação de $ 111,11 foi acompanhada por uma mensagem que dizia: "Decidi contribuir um pouco mais. Uma nota de cem a mais, uma nota de dez a mais, uma nota de um a mais, dez centavos a mais, e um centavo a mais. Nem tudo pode ser para um, mas este está tentando ser para todos."

Mas nenhum desses doadores compartilhou seu acordo publicamente na lista de correio. Nenhum.

Até onde qualquer um desses doadores sabia, eles estavam sozinhos. E ao sintonizarem no dia seguinte, não encontraram agradecimentos, mas sim argumentos explicando por que não deveriam ter feito a doação. As críticas, as justificativas para não contribuir – apenas essas foram expostas com orgulho e abertamente.

Era como se o tesoureiro tivesse concluído sua campanha anual e todos os que não fizeram promes-

sas se levantassem orgulhosamente para invocar justificativas para a recusa; enquanto aqueles que faziam promessas as sussurravam, para que ninguém pudesse ouvir.

Conheço alguém com uma causa racionalista que perambula por aí perguntando melancolicamente: "Como o culto raeliano do disco voador consegue atrair dezenas de milhares de membros [provavelmente cerca de 40.000] interessados em absurdos completos, enquanto nós não conseguimos sequer motivar mil pessoas para isso?"

A maneira obviamente equivocada de concluir esse pensamento seria dizer: "Vamos fazer o que os raelianos fazem! Vamos adicionar algumas bobagens a esse meme!" Para o benefício daqueles que não foram imediatamente detidos por suas inibições éticas, observarei que pode haver uma centena de cultos de discos voadores fracassados para cada um que se torna famoso. E o Lado Negro pode exigir habilidades não óbvias, que você, sim, você não possui: nem todos pode ser um Lorde Sith. Em particular, se você falar sobre mentiras planejadas na Internet pública, você falhará. Não sou um mestre do crime, mas até eu posso dizer que certas pessoas não foram feitas para serem bandidos.

Portanto, provavelmente não é uma boa ideia cultivar um sentimento de direito violado ao pensar que algum outro grupo, que você acredita que deveria ser inferior a você, tem mais dinheiro e seguidores. Esse caminho leva – perdoem a expressão – ao Lado Negro.

Mas provavelmente faz sentido começarmos a nos questionar pontualmente, se os supostos "racionalistas" não conseguem se coordenar tão bem quanto um culto aos discos voadores.

Como funcionam as coisas no Lado Negro?

O líder respeitado fala, e emerge um coro de pura concordância: se houver alguém nutrindo dúvidas internas, elas são mantidas para si. Assim, todos os membros individuais da audiência veem essa atmosfera de pura concordância e sentem-se mais confiantes nas ideias apresentadas – mesmo que, pessoalmente, tenham dúvidas internas; afinal, todos os outros parecem concordar com isso.

("Ignorância pluralista" é o rótulo padrão para isso.)

Se alguém ainda não estiver convencido depois disso, ele deixa o grupo (ou, em alguns lugares, é excluído) – e os demais concordam ainda mais, reforçando-se mutuamente com menos interferência.

(Eu chamo isso de "resfriamento evaporativo de grupos.")

As próprias ideias, e não apenas o líder, são as responsáveis por gerar entusiasmo e elogios ilimitados. O efeito halo faz com que as percepções de todas as qualidades positivas se correlacionem. Por exemplo, ao informar os participantes sobre os benefícios de um conservante alimentar, eles passaram a considerá-lo de menor risco, mesmo que as quantidades não estivessem logicamente correlacionadas. Isso pode criar um ciclo de feedback positivo, fazendo com que uma ideia pareça cada vez melhor, especialmente quando a crítica é vista como traiçoeira ou pecaminosa.

(O que chamo de "espiral de declínio emocional".)

Portanto, todos esses são exemplos de poderosas forças do Lado Negro que podem unir grupos.

E presumivelmente, não chegaríamos ao ponto de nos envolvermos nisso...

Assim, como grupo, o Lado da Luz estará sempre dividido e enfraquecido. Os tecnófilos, os nerds, os cientistas e até mesmo as religiões não fundamentalistas nunca serão capazes de agir com a unidade fanática que impulsiona o Islã radical. A vantagem tecnológica só pode ir até certo ponto; suas ferramentas podem ser copiadas ou roubadas e usadas contra você. No final, o Lado da Luz sempre perderá em qualquer conflito de grupo, e o futuro pertencerá inevitavelmente ao Escuro.

Acredito que a reação de uma pessoa a essa perspectiva revela muito sobre sua atitude em relação à "racionalidade".

Alguns autores do "Choque de Civilizações" parecem aceitar que o Iluminismo está fadado a perder, a longo prazo, para o Islã radical, suspirando e balançando a cabeça tristemente. Suponho que estejam ten-

tando sinalizar sua sofisticação cínica ou algo assim.

Quanto a mim, sempre pensei - chamem-me de maluco - que um verdadeiro racionalista deveria ser eficaz no mundo real.

Portanto, tenho um problema com a ideia de que o Lado Negro, graças à sua ignorância pluralista e às espirais de declínio emocionais, sempre vencerá porque está mais bem coordenado do que nós.

Você pensaria, talvez, que os verdadeiros racionalistas deveriam ser mais coordenados? Certamente, toda essa irracionalidade deve ter suas desvantagens. Esse modo não pode ser o ideal, pode?

E se os atuais grupos "racionalistas" não conseguem se coordenar – se não conseguem apoiar projetos de grupo tão eficientemente quanto uma única sinagoga recebe doações de seus membros – bem, deixo-lhes a tarefa de concluir esse silogismo.

Existe um ditado que às vezes utilizo: "É arriscado ser meio racionalista".

Por exemplo, posso conceber formas de prejudicar a inteligência de alguém, ensinando seletivamente certos métodos de racionalidade. Suponha que você tenha ensinado a alguém uma longa lista de falácias lógicas e preconceitos cognitivos e os treinou para identificar essas falácias e preconceitos nos argumentos de outras pessoas. Contudo, seja cauteloso ao escolher as falácias e preconceitos dos quais é mais fácil acusar os outros, preferindo os mais genéricos que possam ser facilmente mal interpretados. E não os alerte para examinar minuciosamente os argumentos com os quais concordam com a mesma intensidade com que examinam argumentos incongruentes em busca de falhas. Dessa forma, adquirem um vasto repertório de falhas das quais só podem acusar argumentos e argumentadores que não lhes agradam. Suspeito que esta seja uma das principais maneiras de pessoas inteligentes se tornarem estúpidas. (E observe, a propósito, que acabei de lhe apresentar outro Contra-argumento Totalmente Geral contra pessoas inteligentes, cujos argumentos você não aprecia.)

Da mesma forma, se você quisesse garantir que um grupo de "racionalistas" nunca realizasse qualquer tarefa que exigisse mais de uma pessoa, poderia ensinar-lhes apenas técnicas de racionalidade individual, sem mencionar nada sobre técnicas de racionalidade de grupo coordenada.

Escreverei mais tarde sobre como penso que os racionalistas poderiam coordenar-se melhor. Mas aqui quero concentrar-me naquilo que se poderia chamar de cultura do desacordo, ou até mesmo cultura das objeções, que é uma das duas principais forças que impedem a coordenação da multidão tecnófila.

Imagine que você está em uma conferência e o palestrante realiza uma apresentação de trinta minutos. Depois, as pessoas fazem fila diante dos microfones para fazer perguntas. O primeiro contesta o gráfico usado no slide 14 usando uma escala logarítmica; eles citam Tufte em "The Visual Display of Quantitative Information". O segundo contesta uma afirmação feita no slide 3. O terceiro sugere uma hipótese alternativa que parece explicar os mesmos dados...

Perfeitamente normal, certo? Agora, imagine que você está em uma conferência e o palestrante realiza uma apresentação de trinta minutos. As pessoas fazem fila em frente ao microfone.

A primeira pessoa diz: "Concordo plenamente com tudo o que você expôs em sua palestra; considero sua abordagem brilhante." Em seguida, afasta-se.

A segunda pessoa comenta: "O slide 14 foi excepcional, aprendi bastante com ele. Você é incrível." E também se afasta.

Quanto à terceira pessoa—

Bem, nunca saberemos o que a terceira pessoa no microfone tinha a dizer, porque, a essa altura, você já teria fugido da sala gritando, movido por um terror profundo, como se Cthulhu tivesse irrompido do pódio, instigando o medo do fenômeno impossivelmente antinatural que invadiu sua conferência.

Sim, um grupo que não tolera divergências não é racional. No entanto, se você tolera apenas o desacordo – se tolera o desacordo, mas não o acordo –, então você também não é racional. Você está disposto apenas a ouvir algumas opiniões honestas, mas não outras. Você é um meio racionalista perigoso.

Sentimo-nos tão desconfortáveis juntos quanto os membros do culto dos discos voadores se sentem separados. Isso também não pode estar certo. A estupidez reversa não é inteligência.

Vamos imaginar que tenhamos dois grupos de soldados. No Grupo 1, os soldados rasos não têm conhecimento de táticas e estratégias; apenas os sargentos têm alguma compreensão tática, e apenas os oficiais têm algum entendimento de estratégia. No Grupo 2, todos, em todos os níveis, possuem amplo conhecimento tanto em táticas quanto em estratégias.

Poderíamos esperar que o Grupo 1 vencesse o Grupo 2, pois o Grupo 1 seguirá as ordens, enquanto todos no Grupo 2 apresentam ideias superiores a qualquer ordem dada a eles?

Neste caso, tenho que questionar até que ponto o Grupo 2 realmente compreende a teoria militar, pois é uma proposição elementar que uma multidão descoordenada seja massacrada.

Fazer pior com mais conhecimento significa que algo está seriamente errado. Você deve sempre ser capaz de implementar, no mínimo, a mesma estratégia que usaria se fosse ignorante e, de preferência, fazer melhor. Certamente não deveria fazer pior. Se você se arrepende de sua "racionalidade", então é hora de reconsiderar o que é ser racional.

Por outro lado, se você for apenas um meio racionalista, pode facilmente acabar fazendo pior com mais conhecimento. Recordo um experimento interessante que mostrou que estudantes com opiniões políticas e um conhecimento mais profundo dos assuntos, reagiam menos a evidências incongruentes, pois tinham mais munição para contra-argumentar apenas evidências incongruentes.

Parece que estamos presos em um terrível vale da racionalidade parcial, onde acabamos menos coordenados do que os fundamentalistas religiosos, capazes de fazer menos esforço do que os cultistas dos discos voadores. É verdade que o pouco esforço que conseguimos fazer pode ser mais bem direcionado para auxiliar as pessoas e não o contrário – mas essa não é uma desculpa aceitável.

Se eu me propusesse a treinar racionalistas sistematicamente, incluiria lições sobre como discordar e lições sobre como concordar. Estas lições visam tornar o aprendiz mais à vontade com a dissidência, ao mesmo tempo em que o capacitam a se sentirem confortáveis com a conformidade. Um dia, todos aparecem vestidos de forma diferente, em outro dia todos aparecem de uniforme. É essencial abordar ambos os lados, ou você só será meio racionalista.

Já imaginou treinar futuros racionalistas para usarem uniformes, marcharem em sincronia e praticarem sessões onde concordam entre si e aplaudem tudo que um orador no pódio diz? Pode parecer um pesadelo indescritível, não é? Como se todos tivessem confessado abertamente pertencer a um culto maligno. Mas por que seria errado praticar isso, enquanto é considerado correto discordar de todos na multidão? Você nunca terá que concordar com a maioria?

Nossa cultura enfatiza fortemente o desacordo heroico e o desafio heroico, negligenciando o acordo heroico ou o consenso heroico do grupo. Sinalizamos nossa inteligência superior e nossa participação na comunidade não convencional ao inventar objeções inteligentes aos argumentos alheios. Talvez seja por isso que a comunidade tecnófila / Vale do Silício continua marginalizada, perdendo batalhas para facções menos inconformistas na sociedade em geral. Não estamos perdendo porque somos superiores, mas sim porque nossas tradições exclusivamente individualistas sabotam nossa capacidade de cooperação.

Outro componente crucial que, acredito, mina os esforços do grupo na comunidade tecnófila é sentir vergonha de sentimentos intensos. Ainda mantemos o arquétipo de racionalidade de Spock preso em nossas mentes, onde a racionalidade é vista como desapego. Ou talvez um erro relacionado: a racionalidade como cinismo – tentando sinalizar sofisticação superior e desgosto do mundo, mostrando que você se importa menos do que os outros. Tomando cuidado para menosprezar ostensiva e publicamente aqueles que são tão ingênuos a ponto de mostrar que se importam profundamente com alguma coisa.

Você não se sentiria desconfortável se o orador no pódio dissesse que se importa tanto sobre, digamos, a luta contra o envelhecimento, que morreria voluntariamente pela causa?

Mas não há nada escrito em lugar algum, nem na teoria da probabilidade, nem na teoria da decisão, que um racionalista não deveria se importar. Analisei essas equações e realmente, não está lá.

A melhor definição informal que já ouvi de racionalidade é “Aquilo que pode ser destruído pela verdade deveria ser”. Devemos aspirar a sentir emoções que condizem com os fatos, e não a evitar sentimentos. Se uma emoção pode ser destruída pela verdade, devemos abandoná-la. Mas se vale a pena lutar por uma causa, então devemos sentir plenamente sua importância.

Algumas coisas valem a pena morrer. Sim, com certeza! Se não nos sentirmos confortáveis em admitir isso e ouvir outros dizendo o mesmo, teremos dificuldade em nos importar o suficiente – bem como em se coordenar o suficiente – para investir esforço em projetos de grupo. É necessário ensinar ambos os lados: “Aquilo que pode ser destruído pela verdade deve ser” e “Aquilo que a verdade nutre deve prosperar”.

Já ouvi a argumentação de que o tabu contra o uso de linguagem emocional, por exemplo, em artigos científicos, é uma parte crucial para permitir que os fatos se destaquem sem distrações. Isso não implica que o tabu deva ser aplicado em todas as situações. Acredito que existem aspectos da vida nos quais devemos reconhecer e aplaudir o uso de linguagem emocional robusta, eloquência e poesia. Quando há algo que precisa ser realizado, os apelos poéticos contribuem para essa realização e, portanto, merecem reconhecimento.

Precisamos evitar que nossos esforços em expor causas contraproducentes e apelos injustificados interfiram nas tarefas que realmente necessitam ser concluídas. É necessário equilibrar ambos os lados: a disposição de se distanciar de causas contraproducentes e a disposição de elogiar as produtivas; a força para não ser influenciado por apelos infundados e a força para ser influenciado por apelos fundamentados.

Acredito que a sinagoga, em sua campanha anual, estava correta, de fato. Eles não estavam indo fileira por fileira, colocando os indivíduos em situações desconfortáveis, olhando para eles e dizendo: “Quanto você vai doar, Sr. Schwartz?” As pessoas simplesmente declaravam seus compromissos – não com grande drama e orgulho, apenas declarações simples – e isso incentivava outros a fazerem o mesmo. Aqueles que não tinham nada para oferecer permaneciam em silêncio; aqueles que tinham objeções escolhiam um momento posterior ou anterior para expressá-las. Provavelmente, é assim que as coisas deveriam ocorrer em uma comunidade humana saudável, considerando que as pessoas frequentemente têm dificuldade em manter-se tão motivadas quanto gostariam e podem ser auxiliadas pelo estímulo social para superar essa fraqueza de vontade.

Mesmo que discorde dessa perspectiva, podemos argumentar que tanto as opiniões favoráveis quanto as contrárias deveriam ter sido expressas publicamente. Os apoiadores, confrontados com um aparente muro de objeções e discordâncias – mesmo que isso resulte de sua própria autocensura desconfortável – não representam racionalidade de grupo. Isso é simplesmente a imagem refletida do que os grupos do “Lado Negro” fazem para manter seus seguidores. A estupidez reversa não é inteligência.

318 — Tolerar a tolerância

Uma das características prováveis de alguém que se propõe a ser um “racionalista” é uma tolerância menor do que o habitual para falhas de raciocínio. Isso não é uma regra estrita. Pode acontecer, por exemplo, de alguém rejeitar sua religião não porque está naturalmente mais irritado com uma falha de tamanho fixo, mas sim porque identificou falhas mais significativas ou profundas no raciocínio. Contudo, falando realisticamente, muitos de nós provavelmente temos nosso nível de ‘aborrecimento com todas essas falhas que estamos detectando’ um pouco acima da média.

É por isso que é tão crucial tolerarmos a tolerância dos outros se quisermos realizar algo juntos.

Para mim, o exemplo de tolerância que preciso praticar é com relação a Ben Goertzel, que, entre outras atividades, organiza uma conferência anual sobre IA e tem algo positivo a dizer sobre todos. Ben chegou a elogiar as ideias de M*nt*f*x, o mais lendário de todos os excêntricos da IA. (M*nt*f*x aparentemente começou a adicionar um link para o elogio de Ben em suas assinaturas de e-mail, provavelmente porque foi o único elogio que ele já recebeu de um acadêmico de IA genuíno.) (Por favor, não pronuncie seu Nome Verdadeiro corretamente, ou ele será convocado aqui.)

No entanto, percebi que essa é um dos pontos fortes de Ben — sua habilidade em ser amigável com muitas pessoas que outros podem ignorar, incluindo eu. E, de vez em quando, isso acaba sendo vantajoso para ele.

Se eu descontar pontos da reputação de Ben por encontrar algo positivo para dizer sobre pessoas e projetos que considero inúteis, até mesmo M*nt*f*x, estou, na verdade, insistindo que Ben não deve gostar de todos que eu não goste antes que eu possa colaborar com ele.

Será que esse é um padrão realista? Especialmente considerando que diferentes pessoas se incomodam em diferentes graus com coisas distintas?

Mas é difícil lembrar disso quando Ben está sendo amigável até mesmo com tantos idiotas.

A cooperação é instável, tanto na teoria dos jogos quanto na biologia evolutiva, na ausência de alguma forma de punição para a deserção. Assim, uma coisa é deduzir pontos da reputação de alguém por erros cometidos diretamente. Mas se você olhar desconfiado para alguém que se recusa a punir uma pessoa ou ideia, isso se torna uma forma de punição para os não-punidores, uma expressão consideravelmente mais perigosa, capaz de estabelecer um equilíbrio prejudicial para todos os envolvidos.

O perigo de punir aqueles que não punem é algo que me vem à mente sempre que Robin Hanson aponta uma falha em algum tropo acadêmico e ainda confessa, modestamente, que pode estar equivocado (e ele não está). Ou quando vejo Michael Vassar ponderando sobre o potencial de alguém que eu, inicialmente, considerei sem esperança trinta segundos após ser apresentado a ele. Devo recordar a mim mesmo: “Tolere a tolerância! Não exija que seus aliados sejam igualmente extremos em seus julgamentos negativos sobre tudo que você não gosta!”

Por natureza, sinto-me irritado quando percebo que alguém está concedendo demasiado crédito. Não posso afirmar se todos partilham dessa reação, mas suspeito que pelo menos alguns dos meus colegas aspirantes a racionalistas compartilhem desse sentimento. Não seria surpreendente descobrir que seja um traço universal humano, com uma lógica evolutiva evidente, embora o torne uma adaptação potencialmente desagradável e perigosa.

Normalmente, não sou entusiasta da “tolerância”. Certamente, não acredito em ser “intolerante à intolerância”, como alguns defendem de maneira inconsistente. No entanto, continuarei esforçando-me para tolerar pessoas que são mais tolerantes do que eu, julgando-as apenas pelos seus próprios erros, não pelos erros que possam ter adotado.

E, claro, é desnecessário dizer que se as pessoas do Grupo X estão observando-o com expectativas, aguardando que você odeie os inimigos certos com a intensidade correta, e prontas para puni-lo caso você não expresse suficientemente seu repúdio, você pode estar associado ao grupo errado.

Apenas não exija que todos com quem você colabora sejam igualmente intolerantes com comportamentos desse tipo. Perdoe seus amigos se alguns deles sugerirem que talvez o Grupo X não seja tão terrível, afinal...

319 — Seu preço para aderir

No "Jogo do Ultimato", o primeiro jogador escolhe como dividir $ 10 entre ele e o segundo jogador, e o segundo jogador decide se aceita ou rejeita a divisão – neste último caso, ambas as partes não recebem nada. No contexto da teoria convencional da decisão causal (duas caixas no Problema de Newcomb, defeito no Dilema do Prisioneiro), o segundo jogador deve preferir qualquer valor diferente de zero a nada. Mas se o primeiro jogador espera esse comportamento – aceitar qualquer oferta diferente de zero – então ele não terá motivo para oferecer mais do que um centavo. Como presumo que todos vocês já sabem, não sou adepto da teoria convencional da decisão causal. Aqueles de nós que continuam interessados em cooperar no Dilema do Prisioneiro, seja porque é iterado, ou porque temos um termo em nossa função de utilidade para a justiça, ou porque utilizamos uma teoria de decisão não convencional, também podem não aceitar uma oferta de um centavo.

E, de fato, a maioria dos "decisores" do Ultimato oferecem uma divisão equilibrada; e a maioria dos "aceitadores" do Ultimato rejeita qualquer oferta inferior a 20%. Um jogo de 100 USD disputado na Indonésia (rendimento médio per capita na época: 670 USD) mostrou ofertas de 30 USD sendo recusadas, embora isso correspondesse a duas semanas de salário. Provavelmente também podemos supor que os jogadores na Indonésia não estavam pensando no debate acadêmico sobre os problemas do tipo Newcomb – é assim que as pessoas se sentem em relação aos Jogos do Ultimato, mesmo aqueles jogados com dinheiro real.

Existe um análogo do "Jogo do Ultimato" na coordenação de grupo. (Foi estudado? Espero que sim...) Digamos que haja um projeto comum – na verdade, digamos que seja um projeto comum altruísta, destinado a ajudar vítimas de assalto no Canadá, ou algo assim. Se você se juntar a este projeto em grupo, você realizará mais coisas do que conseguiria sozinho, em relação à sua função de utilidade. Então, obviamente, você deveria participar.

Mas espere! O projeto anti-assalto mantém seus fundos investidos em um fundo do mercado monetário! Isso é ridículo; não renderá tantos juros como os títulos do Tesouro dos EUA, muito menos um fundo de índice que paga dividendos.

Claramente, este projeto é dirigido por idiotas, e você não deve aderir até eles mudarem seus métodos de mau investimento.

Agora você pode perceber – se parar para pensar sobre isso – que, considerando todas as coisas, você ainda se sairia melhor trabalhando com o projeto comum anti-assalto, do que lutando sozinho para combater o crime. Mas então – talvez você também perceba – se você concordar com muita facilidade em se juntar ao grupo, por que, que motivo eles teriam para mudar seus métodos de mau investimento?

Bem...Ok, olhe. Possivelmente porque estamos fora do ambiente ancestral onde todos se conhecem... e possivelmente porque a multidão não-conformista tenta repudiar as forças normais de coesão do grupo, como a conformidade e a adoração do líder...

...Parece-me que as pessoas do grupo ateísta/libertário/tecnófilo/fã de ficção científica/, etc. muitas vezes definem seus preços de adesão muito altos. Como um jogo de Ultimato dividido em 50 jogadores, onde cada um dos 50 jogadores exige pelo menos 20% do dinheiro.

Se ponderarmos quantas vezes situações semelhantes surgiriam em ambientes ancestrais, é quase certo que isso se trata de uma questão de psicologia evolucionista. Emoções do Sistema 1, não do cálculo

do Sistema 2. Nossas intuições sobre quando se integrar a grupos, em contraste com quando aguardar por concessões adicionais que favoreçam nossa abordagem preferida, foram refinadas em ambientes de caçadores-coletores, compostos por cerca de 40 indivíduos, todos conhecidos pessoalmente.

Agora, imagine se o grupo tiver 1.000 pessoas. Nesse caso, os instintos herdados dos tempos de caçador-coletor podem subestimar a inércia de um grupo tão vasto, exigindo um preço irrealisticamente alto (em termos de mudanças estratégicas) para aderir. Existe um limite para o esforço organizacional e um número restrito de abordagens que podem ser aceitas para atender às preferências de cada indivíduo.

E se a estratégia for extensa e complexa, algo que requer, por exemplo, dez pessoas para lidar com a papelada ao longo de uma semana, em vez de se esgotarem durante meia hora de negociação ao redor de uma fogueira? Nesse caso, os instintos de caçador-coletor podem subestimar a inércia do grupo em relação às suas próprias exigências.

Agora, considere se você vive em um mundo mais amplo do que uma única tribo de caçadores-coletores, de modo que você só enxerga o representante do grupo negociando com você, sem ter conhecimento das centenas de outras negociações já realizadas. Nesse contexto, seus instintos podem sugerir que é apenas uma pessoa, um estranho, e que ambos são iguais; as ideias deles são equiparáveis às suas, e o ponto de convergência deve ser equitativo.

E se você enfrentar alguma fragilidade de vontade ou acrasia, ou se for influenciado por motivos diferentes daqueles que admitiria para si mesmo, qualquer empreendimento altruísta de grupo que não ofereça recompensas de status e controle pode parecer negligenciado por sua atenção.

Admito que estou abordando esse tema principalmente da perspectiva de alguém que se esforça para conduzir gatos – não do lado oposto, alguém que gasta a maior parte do tempo reservando energia para pressionar aqueles frustrantes indivíduos que já estão no projeto. Talvez minha visão seja um tanto tendenciosa.

Mas parece-me que uma regra prática razoável poderia ser a seguinte:

Se, no geral, unir seus esforços a um projeto de grupo ainda tem um efeito líquido positivo de acordo com sua função de utilidade -

(ou um efeito positivo maior do que qualquer outro uso marginal ao qual você poderia aplicar esses recursos, embora este último modo de pensar pareça pouco utilizado e humanamente irrealista, por razões sobre as quais posso escrever em outra ocasião)

- e a terrível e irritante questão não é tão importante a ponto de você pessoalmente se envolver profundamente o suficiente para dedicar quantas horas, semanas ou anos podem ser necessários para consertá-la -

- então, a questão não justifica reter as suas energias do projeto; seja por instinto, até perceber que as pessoas estão prestando atenção e respeitando você, ou com a intenção consciente de influenciar o grupo para que isso aconteça.

E se a questão for tão significativa para você... então, por favor, junte-se ao grupo e faça o necessário para corrigir as coisas.

Agora, se os contribuidores existentes se recusarem a permitir que você faça isso, e se depender de um terceiro razoável concluir que você é competente o suficiente para fazê-lo, e não houver mais ninguém prejudicado por isso, então, talvez, tenhamos um problema em nossas mãos. E talvez seja hora de um pouco de influência, se os recursos que você pode condicionalmente comprometer forem significativos o suficiente para chamar a atenção deles.

Esta regra é um tanto extrema? Ah, talvez. Deve haver uma razão para que o mecanismo de tomada de decisão de um projeto seja responsável perante os seus apoiadores; o apoio incondicional criaria os seus próprios problemas.

Mas geralmente... Observo que as pessoas subestimam os custos daquilo que pedem, ou talvez ape-

nas ajam por instinto e fixem os seus preços demasiado elevados. Se a multidão inconformista quiser fazer algo em conjunto, precisamos avançar na direção de nos juntarmos a grupos e permanecermos lá pelo menos um pouco mais facilmente. Mesmo diante dos inconvenientes e imperfeições! Mesmo diante da falta de resposta às nossas melhores ideias!

Na era da Internet e na companhia de inconformistas, torna-se um tanto cansativo ler o 451º e-mail público de alguém dizendo que o Projeto Comum não justifica os seus recursos até que o site tenha uma fonte sans-serif.

Claro, muitas vezes isso não se trata realmente de fontes. Pode ser sobre preguiça, acrasia ou rejeições ocultas. Mas em termos de normas de grupo... em termos de que tipo de declarações públicas respeitamos e que desculpas desprezamos publicamente... provavelmente queremos encorajar uma norma de grupo:

Se o problema não vale a pena ser resolvido pessoalmente, por mais esforço que seja necessário, e não surge de má-fé total, não vale a pena recusar-se a contribuir com os seus esforços para uma causa que você considera válida.

320 — Poderá o Humanismo igualar-se ao resultado da Religião?

Possivelmente a maior instituição voluntária no nosso mundo moderno – unida não por polícia e impostos, não por salários e gestores, mas por doações voluntárias de seus membros – é a Igreja Católica.

É demasiadamente vasta para ser mantida unida por meio de negociações individuais, como uma tarefa em grupo entre caçadores-coletores. Contudo, em um mundo maior, com mais pessoas infectadas e transmissão mais rápida, podemos esperar memes mais virulentos. O Antigo Testamento não fala sobre o Inferno, mas o Novo Testamento sim. A Igreja Católica permanece unida por espirais de morte emocionais – em torno de ideias, instituições e líderes. Através de promessas de felicidade eterna e condenação eterna – os teólogos não acreditam realmente nessas coisas, mas muitos católicos comuns sim. Pela simples conformidade de pessoas que se reúnem pessoalmente em uma igreja e são submetidas à pressão dos colegas. E assim por diante.

Nós, que nos atrevemos a nos chamar de "racionalistas", nos consideramos bons demais para tais laços comunitários.

Dessa forma, qualquer pessoa com um projeto de caridade simples e óbvio – como responder com comida e abrigo a um maremoto na Tailândia, por exemplo – estaria muito melhor apelando ao Papa para mobilizar os católicos, em vez de pedir a Richard Dawkins para mobilizar os ateus.

Enquanto isso for verdade, qualquer aumento do ateísmo à custa do catolicismo será uma espécie de vitória vazia, independentemente de todos os outros benefícios.

É verdade que a Igreja Católica também se opõe ao uso de preservativos na África devastada pela AIDS. É verdade que eles desperdiçam enormes quantias do dinheiro que arrecadam em todas essas coisas religiosas. Entregar-se a pensamentos pouco claros não é inofensivo; a oração tem um preço.

Abster-se de fazer coisas prejudiciais é uma verdadeira vitória para um racionalista...

A menos que seja sua única vitória, o que torna tudo um pouco vazio.

Se desconsiderarmos todos os danos causados pela Igreja Católica e focarmos apenas no que é positivo... Será que o católico médio contribui mais para o bem do que o ateu médio, simplesmente por ser mais ativo?

Talvez, se você for mais sábio, mas menos motivado, poderá buscar intervenções altamente eficientes e adquirir recursos de forma econômica. No entanto, poucos de nós realmente fazemos isso, em vez disso, preferimos planejar fazê-lo algum dia.

Agora, você pode levantar as mãos e dizer: "Enquanto não tivermos controle direto sobre o circuito motivacional do nosso cérebro, não é realista esperar que um racionalista seja tão fortemente motivado quanto alguém que realmente acredita que queimará eternamente no inferno se não obedecer."

Isso é um ponto válido. Qualquer teorema popular que afirme que um agente racional deve ter desempenho pelo menos tão bom quanto um agente não-racional baseia-se na suposição de que o agente racional sempre pode implementar qualquer política "irracional" que seja considerada vencedora. Mas, se não podemos escolher ter energia mental ilimitada, então pode ser que algumas crenças falsas sejam, na verdade, mais motivadoras do que quaisquer crenças verdadeiras disponíveis. E se todos geralmente sofremos

de acrasia altruísta, sendo incapazes de ajudar tanto quanto pensamos que deveríamos, então é possível que aqueles que temem a Deus vençam a competição pela produção altruísta.

Mas embora esta seja uma continuação motivada, vamos analisar essa questão um pouco mais a fundo.

Até mesmo o medo do inferno não é um motivador perfeito. Os seres humanos não têm tanta folga nas rédeas da evolução; podemos resistir à motivação por um curto período, mas depois ficamos sem energia mental (dica: infotropismo). Mesmo acreditar que você irá para o inferno não altera esse fato bruto sobre os circuitos cerebrais. Assim, os religiosos cometem pecados e depois são atormentados por pensamentos de ir para o inferno, da mesma forma que os fumantes se censuram por não conseguirem parar.

Se um grupo de racionalistas realmente se importasse com algo... quem disse que eles não poderiam alcançar o mesmo resultado real e efetivo de um católico devoto? O que está em jogo pode não ser a felicidade "infinita" ou a condenação "eterna", mas, claro, o cérebro não consegue conceber $3 \uparrow\uparrow\uparrow 3$, muito menos o infinito. Quem afirmou que a quantidade real de neurotransmissores liberados pelo cérebro (por assim dizer) precisa ser muito menor para o "crescimento e florescimento da humanidade" ou até mesmo para "tailandeses afetados por ondas de maré", do que para a "felicidade eterna no Paraíso"? Qualquer empreendimento envolvendo mais de 100 pessoas envolverá utilidades grande demais para serem visualizadas. E há todos os tipos de outros vieses padrão em ação aqui; ter consciência deles também pode ser vantajoso, espera-se?

A terapia cognitivo-comportamental e a meditação Zen são duas disciplinas mentais que experimentalmente mostraram produzir melhorias reais. Embora não tenha focado meu desenvolvimento na área da arte, não deixo de ter uma verdadeira arte marcial da racionalidade a meu favor. Se você unir um propósito que realmente valha a pena com a disciplina derivada da TCC e da meditação Zen, quem disse que os racionalistas não conseguem acompanhar? Ou, de maneira mais abrangente: se possuirmos uma arte de combate à acrasia baseada em evidências, com experimentações para verificar o que realmente funciona, quem disse que precisamos ter menos motivação do que uma mente desorganizada que teme a ira de Deus?

Ainda assim… essa é uma especulação futura, uma possibilidade de desenvolver uma arte que ainda não existe. Não é uma técnica que posso empregar agora. Apresento-a apenas para ilustrar a ideia de não desistir tão rapidamente da racionalidade: compreender o que está errado, tentar corrigi-lo de maneira inteligente e reunir evidências sobre se funcionou – uma expressão poderosa que não deve ser descartada levianamente ao deparar com a primeira desvantagem.

Na verdade, suspeito que o que está ocorrendo aqui tem menos a ver com o poder motivacional da condenação eterna e mais a ver com o poder de conhecer fisicamente outras pessoas que compartilham sua causa. O poder, em outras palavras, de estar presente na igreja e ter vizinhos religiosos.

Isso representa um desafio para a comunidade racionalista em seu atual estágio de crescimento, pois somos raros e estamos geograficamente dispersos por toda parte. Se todos os leitores do *Less Wrong* vivessem em um raio de oito quilômetros uns dos outros, aposto que seríamos muito mais eficazes, não apenas por questões de coordenação, mas simplesmente por pura motivação.

Escreverei mais tarde sobre alguns pensamentos idealistas e de longo prazo relacionados a esse problema específico. Seriam mais eficazes soluções de curto prazo que não dependam do aumento dos nossos números em um fator de 100. Fico pensando se o programa moderno de videoconferência proporcionaria algum efeito motivador semelhante ao de conhecer alguém pessoalmente. Suspeito que a resposta seja "não", mas pode valer a pena tentar.

Enquanto isso, no curto prazo, estamos enfrentando a luta contra a acrasia, principalmente sem o reforço da presença física de outras pessoas que se importam. Quero expressar algo como "Isso é difícil, mas pode ser feito", embora eu não esteja certo se isso é verdade.

Suspeito que o maior passo que os racionalistas poderiam dar para equiparar a produção de poder per capita da Igreja Católica seria realizar reuniões físicas regulares de pessoas que contribuem para a mesma tarefa, principalmente para motivar uns aos outros.

Na ausência disso...

Podemos tentar estabelecer uma norma de grupo que nos permita sermos abertamente autorizados - ou melhor, aplaudidos - por nos importarmos profundamente com algo. E uma norma de grupo que espera que cada um de nós faça algo significativo em sua vida - contribuir com sua parte para melhorar este mundo. A religião não enfatiza tanto o aspecto de realizar ações.

E se os racionalistas pudessem igualar apenas metade da produção média de esforço altruísta de um católico, então não acredito que seja remotamente irrealista supor que, com uma orientação mais eficiente para causas, o racionalista médio poderia realizar o dobro.

Quanto dos recursos financeiros da Igreja Católica são gastos em atividades religiosas inúteis, em vez de realmente ajudar as pessoas? Arriscaria dizer que é mais de 50%. Então, poderíamos sugerir - com certa ironia, embora não seja exatamente o espírito com o qual deveríamos agir - que devemos propagar uma norma de grupo de doar pelo menos 5% da renda para causas reais. (O dízimo religioso mínimo sugerido é geralmente de 10%). Há também a arte de escolher causas para as quais os utilons esperados são significativamente mais baratos (enquanto durar o mercado ineficiente de utilons).

No entanto, muito antes de podermos começar a sonhar com tal orgulho, nós, humanistas seculares, precisamos trabalhar para, pelo menos, igualar a produção benevolente per capita dos fiéis.

321 — Igreja vs. Força-Tarefa

Frequentemente suspeito de invejar grupos excêntricos ou de tentar imitar cegamente o ritmo da religião – o que eu chamo de “hinos à inexistência de Deus” – respondo: “Um bom “hino ateu” é simplesmente uma canção sobre qualquer coisa que valha a pena cantar e que não seja religiosa.”

Mas a religião de fato preenche certas lacunas na mente das pessoas, algumas das quais podem ser consideradas importantes. Se eliminarmos a religião, devemos estar cientes de quais lacunas que serão deixadas para trás.

Se, de repente, excluirmos a religião do mundo, a maior lacuna deixada não será a falta de ideais ou de moral; será a igreja, a comunidade. Entre aqueles que permanecem religiosos sem realmente acreditar em Deus - quantos estão apenas seguindo isso para ficar perto de seus vizinhos na igreja, de suas famílias e amigos? Quantos se tornariam ateus se todos os outros desistissem, e esse fosse o preço para manter a comunidade e o respeito deles? Eu suporia... provavelmente muitos.

Na verdade, isso é algo que talvez eu mesmo não entenda completamente. “Brownies e babás” foram as duas primeiras coisas que me vieram à mente. As igrejas oferecem ajuda em emergências? Ou são apenas um ombro para chorar? Quão forte é uma comunidade religiosa? Provavelmente depende da igreja, e, de qualquer forma, essa não é a pergunta correta. Deveríamos começar considerando o que um grupo de caçadores-coletores oferece ao seu povo e questionar o que está faltando na vida moderna. Se uma igreja moderna do Primeiro Mundo preenche apenas parte disso, então, por todos os meios, vamos tentar fazer melhor.

Portanto, sem copiar a religião – sem presumir que devemos nos reunir todos os domingos de manhã em um prédio com vitrais enquanto as crianças se vestem com roupas formais e ouvem alguém cantar – vamos considerar como preencher a lacuna emocional depois que a religião deixar de ser uma opção.

Para quebrar o molde, para começar - a camisa de força de pensamentos preestabelecidos sobre como realizar esse tipo de coisa - considere que alguns escritórios modernos também podem desempenhar o mesmo papel que uma igreja. Com isso, quero dizer que algumas pessoas têm a sorte de encontrar comunidade em seus locais de trabalho: colegas amigáveis que preparam brownies para o escritório, cujos adolescentes podem ser contratados com segurança para serem babás e talvez até ajudar em tempos de catástrofe...? No entanto, certamente nem todos têm a sorte de encontrar uma comunidade no ambiente de trabalho.

Vamos ainda mais longe: uma igreja é ostensivamente sobre adoração, e um local de trabalho é ostensivamente sobre o propósito comercial da organização. Nenhum dos dois foi cuidadosamente otimizado para servir como uma comunidade.

Ao observar uma igreja religiosa típica, por exemplo, pode-se suspeitar - embora todas essas conjecturas fossem melhor testadas experimentalmente do que apenas suspeitadas:

- Que acordar cedo em um domingo de manhã não é o ideal;
- Que vestir roupas formais, especialmente para crianças, não é a escolha mais adequada;
- Que ouvir a mesma pessoa proferir sermões sobre o mesmo tema todas as semanas (“religião”) não é o ideal;

- Que o custo de sustentar uma igreja e um pastor é elevado em comparação com o número de comunidades diferentes que poderiam compartilhar o mesmo edifício para suas reuniões;
- Que provavelmente não servem nem de longe o suficiente ao propósito de promover encontros românticos, pois as igrejas acreditam que devem impor suas moralidades medievais;
- Que tudo deveria ser submetido a uma coleta experimental de dados para descobrir o que funciona e o que não funciona.

Ao usar a palavra "ideal" acima, refiro-me a algo "ideal conforme os critérios que você usaria ao construir explicitamente uma comunidade enquanto comunidade". Investir quantias consideráveis em uma igreja suntuosa com vitrais e um pastor em tempo integral faz sentido se você genuinamente quiser investir dinheiro em religião enquanto religião.

Confesso que, ao passar por entre as igrejas da minha cidade, meu pensamento principal é: "Esses edifícios parecem extremamente caros, e são numerosos". Se estivéssemos começando do zero, poderíamos ter um grande edifício que serviria para ocasionais cerimônias de casamento, compartilhado por diferentes comunidades em horários distintos nos fins de semana. Além disso, o local poderia contar com um grande display de vídeo utilizado para palestrantes, apresentações educacionais ou até mesmo a exibição de filmes. Vitrais? Não seriam uma prioridade tão alta.

Ao passo que os membros da igreja oferecem ajuda em momentos difíceis - isso poderia ser aprimorado através de um fundo específico para emergências ou da contratação de um seguro, reconhecendo assim a importância desta função? Provavelmente não; atribuir financiamento explícito a essas coisas altera sua natureza de maneira peculiar. Por outro lado, talvez se manter informado sobre algumas apólices de seguro deva ser um requisito para ser membro, a fim de não depender excessivamente da comunidade. . . No entanto, novamente, na medida em que as igrejas fornecem um senso de comunidade, elas tentam fazer isso sem realmente admitir que isso é praticamente tudo o que as pessoas obtêm delas. O mesmo ocorre com empresas cujos locais de trabalho são suficientemente acolhedores para funcionarem como comunidades; ainda é uma função incidental.

Quando você começa a considerar explicitamente como proporcionar às pessoas um grupo ao qual pertencer, você pode vislumbrar uma variedade de ideias que parecem boas. Dever-se-ia acolher o recém-chegado em seu meio? O pastor poderia abordar esse tema em algum momento, se você acredita que a igreja está ligada à religião. No entanto, se a intenção é construir conscientemente uma comunidade, logo após uma mudança é quando alguém mais carece de comunidade, quando mais precisa de ajuda. É também uma oportunidade para a comunidade prosperar. Na verdade, as tribos deveriam competir em eventos trimestrais para conquistar os recém-chegados.

Mas é realmente possível ter uma comunidade que seja apenas uma comunidade - que não seja também um local de trabalho ou uma religião? Uma comunidade sem propósito além de si mesma?

Talvez seja possível. Afinal, as tribos de caçadores-coletores tinham algum propósito além delas mesmas? - Bem, havia a sobrevivência e a obtenção de alimentos; isso era um propósito.

Mas tudo o que as pessoas têm em comum, especialmente qualquer objetivo compartilhado, tende a definir uma comunidade. Por que não tirar proveito disso?

Embora nesta era da Internet, lamentavelmente, muitos fatores de conexão tenham apoiadores amplamente distribuídos para formar um grupo decente - se você for o único membro da Igreja do Subgênio em sua cidade, isso pode não ser muito útil. Realmente é diferente sem a presença física; a Internet não parece ser um substituto aceitável no estágio atual da tecnologia.

Então, para ser direto —

Se a Terra durar tanto tempo, eu gostaria de ver, como parte das comunidades racionalistas, grupos de trabalho concentrados em todas as tarefas necessárias para consertar este mundo. Comunidades em qualquer área geográfica se formariam em torno do agrupamento mais específico que pudesse sustentar um tamanho decente. Se sua cidade não tiver pessoas suficientes para você encontrar 50 colegas programadores em Linux, talvez você tenha que se contentar com 15 colegas programadores de código aberto... ou nos dias

iniciais, 15 companheiros racionalistas tentando aprimorar a Terra à sua maneira.

Acredito que esta seja a direção apropriada para direcionar as energias das comunidades e estabelecer um objetivo comum que as una. Independentemente disso, empreendimentos como este demandam o envolvimento das comunidades, e nosso planeta possui uma quantidade considerável de trabalho a ser realizado, tornando o desperdício de recursos algo sem sentido. Há tantas tarefas que precisam ser realizadas: permitir que a energia que antes se dissipava nos vazios das instituições religiosas encontre um propósito significativo. E possibilitar que objetivos admiráveis, desprovidos de ilusões, preencham lacunas na estrutura comunitária, deixados após eliminar a religião e seus propósitos ilusórios superiores.

Comunidades robustas edificadas em torno de propósitos valorosos: é assim que idealizo a era pós-religiosa, aplicável a qualquer fração da humanidade que tenha alcançado este estágio em suas jornadas.

Entretanto, desde que se disponha de um edifício com uma tela grande de alta resolução, não vejo problemas em questionar a ideia de que toda a aprendizagem após a fase adulta deva ocorrer em campi universitários distantes e dispendiosos, com professores que prefeririam estar fazendo algo diferente. De forma empírica, as instituições acadêmicas parecem ser eficientes em apoiar comunidades. Assim, com toda justiça, existem outras possibilidades em torno das quais se poderia construir uma comunidade pós-teísta.

Será que tudo isso não passa de um sonho? Talvez. Provavelmente. No entanto, não é totalmente desprovido de aplicabilidade incremental, especialmente se houver um número suficiente de racionalistas em uma cidade suficientemente grande que já tenham conhecimento da ideia. Contudo, caso a racionalidade se dissemine em larga escala ou se a Terra perdure por tanto tempo, e se minha voz for ouvida, então essa é a direção na qual eu gostaria de ver as coisas evoluírem – à medida que as igrejas desaparecem, não precisamos de substitutos artificiais, mas sim de novos idiomas para a comunidade.

322 — Racionalidade: interesse comum de muitas causas

É uma agenda não tão oculta do blog *Less Wrong* que muitas causas se beneficiam da disseminação da racionalidade – porque é preciso um pouco mais de racionalidade do que o comum para se ver como defensor, ou mesmo como um apoiador permanente. Isso não se aplica apenas a causas óbvias, como o ateísmo, mas também a questões como a legalização da maconha – onde seria desejável que as pessoas tivessem um pouco mais de autoconsciência sobre suas motivações e a natureza da sinalização, e um pouco mais movidas por fatos frios e inconvenientes. O *Machine Intelligence Research Institute* foi apenas um exemplo extraordinariamente extremo disso, chegando ao ponto em que, após anos de impasse, eu ergui as mãos e retornei explicitamente ao trabalho de criar racionalistas.

Mas, é claro, nem todos os racionalistas que eu crio estarão interessados no meu próprio projeto – e isso é aceitável. Não é possível abranger todo o valor que se cria, e tentar fazê-lo pode ter efeitos colaterais negativos.

Se os apoiadores de outras causas forem suficientemente esclarecidos para pensar de maneira semelhante...

Então todas as causas beneficiadas da disseminação da racionalidade podem, talvez, ter algo como um material padronizado para indicar aos seus apoiadores – uma tarefa comum, centralizada para economizar esforço – e pensar que estão disseminando um pouco de racionalidade no campo. Eles não capturarão todo o valor que geram. E está tudo bem. Eles capturarão parte do valor gerado por outros. O ateísmo tem muito pouco a ver diretamente com a legalização da maconha, mas se tanto ateus quanto defensores da legalização estiverem dispostos a recuar um pouco e falar sobre o princípio geral e abstrato de confrontar uma verdade desconfortável que interfere em um discurso bonito e justo, ambos os esforços podem colher alguns benefícios.

Mas isso exige – eu sei que estou repetindo, mas é crucial – que estejamos dispostos a não abarcar todo o valor que criamos. Exige que, ao falar sobre racionalidade, possamos temporariamente nos calar sobre nossa própria causa, mesmo que seja a melhor de todas. Exige que não consideremos essas outras causas, e que elas não nos considerem, como concorrentes por uma oferta limitada de racionalistas com uma capacidade restrita de apoio; mas sim, como criadoras de mais racionalistas e ampliadoras de sua capacidade de apoio. Colhemos apenas alguns dos frutos de nossos próprios esforços, mas também colhemos alguns dos esforços dos outros.

Se você e eles não concordarem em tudo – especialmente em prioridades – você deve estar disposto a concordar em se abster de discutir sobre o desacordo. (Exceto, possivelmente, em locais especializados, fora do caminho do discurso predominante, onde tais divergências são explicitamente tratadas.)

Uma pessoa em particular que assumiu a posição de presidente de uma organização específica observou certa vez que a organização não teve muito sucesso com a mensagem “Esta é a melhor coisa que você pode fazer”, em comparação, por exemplo, com a mensagem “Aqui está uma coisa incrível que você pode fazer”, o que é evidenciado pelo tremendo sucesso da Fundação X-Prize ao transmitir aos indivíduos ricos “Aqui está algo incrível que você pode fazer”.

Esta é uma daquelas percepções em que você pisca, incrédulo, e então percebe o quão coerente isso é. O cérebro humano não consegue compreender grandes riscos, e as pessoas estão longe de serem maximizadoras da utilidade esperada; geralmente, somos altruístas acráticos. Dizer: “Isso é a melhor coisa” não

adiciona muita motivação além de “Isso é algo legal”. Isso apenas estabelece um ônus da prova muito maior. E convida a comparações odiosas, que minam a motivação, com todas as outras coisas boas que você conhece (talvez ameaçando diminuir a satisfação moral já adquirida).

Se partirmos da suposição de que, por padrão, todos são altruístas acráticos (alguém que deseja ter o poder de escolher fazer mais) - ou pelo menos, que a maioria dos potenciais apoiadores de interesses se encaixa nessa descrição - então discutir qual causa é a melhor para apoiar pode ter o efeito de diminuir a oferta global de altruísmo.

“Mas”, você diz, “os dólares são fungíveis; um dólar que você usa para uma coisa, na verdade, não pode ser usado para mais nada!” Ao que eu respondo: Mas os seres humanos não são realmente maximizadores de utilidade, como sistemas cognitivos. Os dólares saem de diferentes contas mentais, custam diferentes quantidades de força de vontade (o verdadeiro recurso limitante) em diferentes circunstâncias. As pessoas querem distribuir suas doações como um ato de contabilidade mental para minimizar o arrependimento se uma única causa falhar, e contar a alguém sobre uma causa adicional pode aumentar o valor total em que estão dispostos a ajudar.

É claro que existem limites para este princípio de tolerância benigna. Se o projeto favorito de alguém é ensinar a dançar salsa, seria um exagero dizer que estão trabalhando em uma sub tarefa digna do grande projeto comum do Neo-Iluminismo de progresso humano.

Mas na medida em que algo é realmente uma tarefa que vocês gostariam de ver realizada em nome da humanidade... então, comparações invejosas desse projeto com Seu Projeto Favorito podem não ajudar seu próprio projeto tanto quanto você imagina. Talvez precisemos aprender a dizer, por hábito e em quase todos os fóruns, “Aqui está um projeto racionalista fantástico”, e não: “O meu, por si só, é o maior retorno em utilons esperados por projeto de dólar marginal”. Se alguém com sangue-frio o suficiente para maximizar a utilidade esperada do dinheiro fungível, sem levar em conta os efeitos colaterais emocionais, pedir explicitamente, talvez possamos encaminhá-lo para um subfórum especializado onde qualquer pessoa disposta a fazer a reivindicação de prioridade máxima lute contra isso. Embora, se tudo correr bem, os projetos que têm fortes pretensões a este tipo de carência receberão mais investimento e seus retornos marginais diminuirão, e o vencedor das pretensões concorrentes deixará de ser claro.

Se existirem muitos projetos racionalistas que se beneficiam da elevação do limite da sanidade, então sua tolerância mútua e o investimento comum na difusão da racionalidade poderiam, concebivelmente, apresentar um problema de bens comuns. Mas isso não parece muito difícil de lidar: se houver um grupo que não esteja disposto a compartilhar os racionalistas que criaram ou a mencionar-lhes que podem existir outros projetos neo-iluministas, então quaisquer recursos racionalistas comuns e centralizados poderiam remover a menção de seu projeto como uma coisa legal de se fazer.

Embora tudo isso seja um pensamento idealista e voltado para o futuro, os benefícios – para todos nós – podem ser encontrar algumas coisas importantes que estamos perdendo neste momento. Muitos projetos racionalistas têm poucos apoiadores e são abrangentes; se todos pudéssemos nos identificar como elementos do Projeto Comum do progresso humano, o Neo-Iluminismo, haveria uma probabilidade substancialmente maior de encontrarmos dez de nós em qualquer cidade. Neste momento, muitos desses projetos são um pouco solitários para seus apoiadores. A racionalidade pode não ser a coisa mais importante do mundo – é claro que é isso que protegemos – mas é uma coisa interessante que muitos de nós têm em comum. Poderíamos ganhar muito nos identificando também como racionalistas.

323 — Indivíduos indefesos

Quando consideramos que nossos instintos de agrupamento foram otimizados para bandos de caçadores-coletores de 50 pessoas, onde todos se conhecem, começa a parecer surpreendente que as grandes instituições modernas tenham sobrevivido.

Bem, existem governos com forças armadas e polícias especializadas que podem cobrar impostos. Esse é um paradigma não ancestral que remonta à invenção da agricultura sedentária e dos excedentes extraíveis; a humanidade continua lutando para lidar com isso.

Existem empresas nas quais o fluxo de dinheiro é controlado por uma gestão centralizada, uma expressão não ancestral que remonta à invenção do comércio em grande escala e da especialização profissional.

Em um mundo com grandes populações e contato próximo, os memes evoluem de maneira muito mais virulenta do que a média do ambiente ancestral; memes que proferem ameaças de condenação, promessas celestiais e aulas de sacerdotes profissionais para transmiti-los.

No entanto, de maneira geral, a resposta à pergunta “Como as grandes instituições sobrevivem?” é “Elas não sobrevivem!” A grande maioria das grandes instituições modernas – algumas delas extremamente vitais para o funcionamento de nossa complexa civilização – simplesmente falham em existir em primeiro lugar.

Percebi isso pela primeira vez ao compreender como a Ciência é financiada especificamente, não por doações individuais.

A ciência tradicionalmente é financiada por governos, empresas e grandes fundações. Tive a oportunidade de constatar em primeira mão que é incrivelmente difícil arrecadar dinheiro para a Ciência junto de indivíduos. Não, a menos que seja ciência relacionada a uma doença com vítimas horríveis, e talvez nem mesmo assim.

Por quê? As pessoas são, de fato, pró-sociais; elas doam dinheiro, por exemplo, para resgatar animais. A Ciência é um dos grandes interesses sociais, e as pessoas estão amplamente conscientes disso – por que não apoiar a ciência, então?

Qualquer projeto científico específico – por exemplo, a pesquisa sobre a genética da tripanotolerância em bovinos – não se configura como uma opção emocionalmente atrativa para a caridade individual. A ciência possui um horizonte temporal extenso que demanda apoio contínuo. Comunicados de imprensa, sejam provisórios ou finais, podem não despertar grande entusiasmo emocional. Não há oportunidade para voluntariado; trata-se de uma tarefa destinada a especialistas. Observar a imagem do cientista que você está apoiando, mesmo que seja por um preço de mercado ou ligeiramente abaixo do salário padrão, não tem o mesmo impacto que ver um filhote de cachorro de olhos arregalados que você ajudou a encontrar um novo lar. Não há um feedback imediato nem a sensação de realização instantânea necessária para manter um indivíduo investindo seu próprio dinheiro.

De forma irônica, percebi isso não através da minha própria experiência, mas ao questionar por que os leitores de Seth Roberts não se unem para apoiar os experimentos que testam a hipótese de Roberts sobre a obesidade· Por que filantropos individuais não estão financiando os testes do fusor Polywell de Bussard? Estes são exemplos de pesquisas científicas claramente subfinanciadas, com aplicações que seriam relevantes para muitos, muitos indivíduos, se verdadeiras. Foi então que me dei conta de que, de maneira geral, a

Ciência não é uma opção emocionalmente envolvente para pessoas que estão gastando seu próprio dinheiro.

Na verdade, poucas coisas o são para os indivíduos como eles são agora. Parece-me que entender isso é fundamental para compreender o funcionamento do mundo como o conhecemos – por que tantos interesses individuais estão mal protegidos, por exemplo, ou por que 200 milhões de adultos americanos enfrentam tantas dificuldades em supervisionar os 535 membros do Congresso.

Então, como a Ciência é verdadeiramente financiada? Por meio de governos que decidem investir uma certa quantia em Ciência, com legislaturas ou executivos tomando essa decisão – não é exatamente o dinheiro deles que estão gastando. Empresas suficientemente grandes decidem alocar uma determinada quantia em pesquisa e desenvolvimento em áreas inexploradas. Grandes organizações fundamentadas em espirais de declínio emocional podem direcionar recursos para pesquisas científicas, que estejam alinhadas com seus ideais. Grandes fundações privadas, sustentadas por fundos alocados por indivíduos ricos para suas reputações, investem em Ciência que promete ser benevolente, assim como investem em orquestras ou arte moderna. E então, cientistas individuais (ou grupos de trabalho científicos individuais) competem pelo controle dessa oferta monetária pré-determinada, entregue nas mãos de membros do comitê de concessões, que aparentam ser as pessoas mais aptas a julgar os cientistas.

Raramente vemos um projeto científico fazendo uma oferta direta por uma parte do fluxo de recursos da sociedade; ao invés disso, primeiro é atribuído à Ciência, e então os cientistas debatem sobre quem o receberá de fato. Mesmo as exceções a essa regra são mais provavelmente motivadas por políticos (projeto lunar) ou por objetivos militares (projeto Manhattan) do que pelo apelo direto dos cientistas ao público.

Agora, estou certo de que se o público se habituasse a financiar ciência por meio de doações individuais, muito dinheiro seria desperdiçado, por exemplo, em pesquisas científicas sobre jargões quânticos – assumindo que o público de alguma forma desenvolvesse o hábito de financiar a ciência sem que outros aspectos das pessoas ou da sociedade fossem alterados.

Ainda assim, é interessante notar que a Ciência consegue sobreviver não porque seja do interesse coletivo ver a Ciência ser realizada, mas sim porque a Ciência se enraizou como um parasita nas poucas formas de grandes organizações que podem existir em nosso mundo. Existem muitos outros projetos que simplesmente não conseguem existir.

Parece-me que a humanidade moderna consegue fazer muito pouco em termos de esforço coordenado para servir aos interesses coletivos individuais. Isso se torna um desafio não ancestral quando você ultrapassa a marca de 50 pessoas. Existem apenas grandes tributadores, grandes comerciantes, supermemes, indivíduos ocasionalmente poderosos, e algumas outras organizações, como a Ciência, que conseguem se associar de forma parasitária a eles.

324 — Dinheiro: a unidade de cuidado

Steve Omohundro propôs um teorema popular, segundo o qual, dentro de qualquer agente auto-modificador aproximadamente racional, o benefício marginal de investir recursos adicionais em qualquer coisa deveria ser aproximadamente igual. Ou, para ser mais preciso, a realocação de uma unidade de recurso entre quaisquer duas tarefas não deveria resultar em aumento na utilidade esperada, considerando a função de utilidade do agente e suas expectativas probabilísticas sobre seus próprios algoritmos.

Esse princípio de equilíbrio de recursos implica que, em uma ampla gama de sistemas aproximadamente racionais, incluindo até mesmo o interior de uma mente auto-modificadora, existe uma moeda comum de utilidades esperadas pela qual tudo o que vale a pena fazer pode ser mensurado.

Em nossa sociedade, essa moeda comum de utilidades esperadas é chamada de "dinheiro". É a medida de quão importante algo é para a sociedade.

Este é um ponto brutal, porém evidente, que muitos podem se sentir inclinados a contestar.

Com esse público, espero poder afirmar isso claramente e prosseguir. Não é como se acreditassem que a "sociedade" tenha sido inteligente, benevolente e sensata até agora, não é mesmo?

Destaco isso para enfatizar um ponto comum a muitas causas nobres. Qualquer instituição de caridade que já tenha recebido seu apoio certamente deseja que você compreenda esse ponto, quer tenha sido expresso verbalmente ou não. Pois ouvi outras pessoas no mundo das organizações sem fins lucrativos e sei que não estou falando apenas por mim aqui...

Muitas pessoas, ao identificarem algo que consideram valer a pena fazer, desejam doar algumas horas de seu tempo como voluntárias, talvez enviar um laptop usado e algumas provisões enlatadas pelo correio, ou participar de uma manifestação em algum lugar, mas, de qualquer forma, evitar gastar dinheiro.

Entendo perfeitamente o sentimento, acredite. Sempre que gasto dinheiro, parece que estou perdendo pontos de vida. Esse é o dilema de ter um valor único para descrever seu patrimônio líquido: ver esse número diminuir não é agradável, mesmo que seja uma parte natural da vida. Deveria existir um princípio na teoria da diversão para lidar com isso.

Mas, bem...

Existe um enigma muito antigo na economia sobre o advogado que passa uma hora como voluntário no refeitório, em vez de trabalhar uma hora extra e doar o dinheiro para contratar alguém para trabalhar cinco horas no refeitório.

Há algo chamado "Lei da Vantagem Comparativa de Ricardo". Existe a ideia de "especialização profissional". Existe a noção de "economias de escala". Existe o conceito de "ganhos do comércio". A razão pela qual usamos dinheiro é para obter os enormes benefícios possíveis quando cada um de nós faz o que faz de melhor.

Isso é o que os adultos fazem. É o que você faz quando realmente deseja que algo seja feito. Você usa dinheiro para contratar especialistas em tempo integral.

Sim, às vezes as pessoas são limitadas em sua capacidade de trocar tempo por dinheiro (subempregadas), então é melhor para elas doar diretamente o que normalmente trocariam por dinheiro. Se a cozinha

comunitária precisasse de um advogado, e o advogado doasse um grande bloco de tempo de advocacia de alta prioridade, então esse tipo de voluntariado faz sentido – é a mesma capacidade especializada que o advogado normalmente trocaria por dinheiro. Mas "voluntariar-se" apenas uma hora de trabalho jurídico, constantemente interrompida, dividida ao longo de três semanas em minutos casuais, entre outros trabalhos? Isso não é como as coisas são feitas quando alguém realmente se importa ou, de forma quase equivalente, quando há dinheiro envolvido.

Na medida em que as pessoas não conseguem compreender esse princípio instintivamente, podem pensar que o uso do dinheiro é de alguma forma opcional na busca de coisas que parecem moralmente desejáveis – em oposição a tarefas como alimentar-nos, cuja desejabilidade parece ser tratada de maneira estranhamente diferente. Esse fator pode ser suficiente por si só para nos impedir de prosseguir nosso interesse coletivo comum em grupos com mais de 40 pessoas.

As economias de comércio e de especialização profissional não são apenas ideias vagamente boas, mas que não soam naturais; são a única forma de fazer alguma coisa neste mundo. O dinheiro não é pedaços de papel; é a moeda comum do cuidado.

Daí o velho ditado: "O dinheiro faz o mundo girar; o amor mal o impede de explodir."

Agora, enfrentamos o problema da acrasia – de não conseguirmos fazer o que decidimos fazer – que é uma parte da arte da racionalidade que espero que alguém desenvolva; eu me especializo mais no negócio de questões impossíveis. E sim, gastar dinheiro é mais doloroso do que fazer voluntariado, porque você pode ver o saldo da conta bancária diminuir, enquanto as horas restantes do nosso tempo não estão visivelmente numeradas. Mas quando chega a hora de se alimentar, você pensa: "Hum, talvez eu devesse tentar criar meu próprio gado, isso é menos doloroso do que gastar dinheiro com carne bovina?" Nem tudo pode ser feito sem invocar a Lei de Ricardo; e do outro lado desse comércio estão pessoas que sentem a mesma dor ao pensar em ter menos dinheiro.

Parece-me de imediato que há maneiras de atenuar a dor associada à perda de pontos de vida e aumentar a sensação de conexão ao doar dinheiro e "fiz uma coisa boa". Atualmente, estou empenhado em realizar essa transformação, destacando a verdadeira essência e o impacto do dinheiro; e denunciando o meme venenoso que diz que doar apenas dinheiro não deve se importar o suficiente para se envolver pessoalmente. Isso é uma mera reflexão de uma mente que não compreende a economia de mercado pós-caçador-coletor. Doar dinheiro não se resume ao momento de preencher o cheque; é o resultado de cada hora dedicada a ganhar dinheiro para concretizar essa contribuição – como se você trabalhasse para a própria instituição de caridade em sua capacidade profissional, com a eficiência máxima da vida adulta.

Se o advogado precisar passar uma hora no refeitório para manter-se motivado e lembrar o propósito de suas ações, tudo bem. No entanto, deveria também doar parte das horas trabalhadas no escritório, pois isso exemplifica o poder da especialização profissional. O cheque pode ser encarado como a aquisição do direito de voluntariar-se no refeitório ou como a validação do tempo investido lá. Abordarei isso mais detalhadamente posteriormente.

Em uma abordagem inicial, o dinheiro é a unidade de cuidado até um fator escalar positivo – a unidade de cuidado relativo. Algumas pessoas são econômicas e gastam menos dinheiro em todas as áreas; entretanto, se você realmente desembolsa US$ 5 em um burrito, então qualquer coisa pela qual você não gaste esse valor terá menos importância que o burrito. Se você opta por não gastar dois meses de salário em um anel de diamantes, isso não implica falta de amor pelo seu parceiro. ("De Beers: It's Just A Rock.") Mas por outro lado, se você reluta constantemente em gastar dinheiro com seu parceiro, mas não hesita em desembolsar US$ 1.000 em uma TV de tela plana, então sim, isso revela algo sobre seus valores relativos.

Sim, a frugalidade é uma virtude. Sim, gastar dinheiro causa desconforto. No entanto, em última análise, se você nunca está disposto a investir unidades de cuidado, significa que você não se importa.

325 — Compre Fuzzies e Utilons separadamente

Anteriormente:

Existe um enigma/observação muito antigo na economia sobre o advogado que dedica uma hora como voluntário no refeitório, em vez de trabalhar uma hora extra e doar o dinheiro para contratar alguém...

Se um advogado precisar passar uma hora no refeitório para manter-se motivado e lembrar o propósito de suas ações, tudo bem. No entanto, deveriam também doar parte das horas trabalhadas no escritório, pois isso exemplifica o poder da especialização profissional. O cheque pode ser encarado como a aquisição do direito de voluntariar-se no refeitório ou como a validação do tempo investido lá. Abordarei isso mais detalhadamente posteriormente.

Costumo segurar portas para velhinhas. Na verdade, não consigo lembrar da última vez que isso aconteceu literalmente (embora tenha certeza de que ocorreu em algum momento do ano passado). Mas no último mês, por exemplo, estava dando um passeio quando notei uma perua estacionada na garagem com o porta-malas completamente aberto, proporcionando acesso total ao interior do veículo. Procurei ver se alguém estava retirando pacotes, mas não era o caso. Olhei ao redor para verificar se alguém mexia no carro. Por fim, fui até a casa, bati à porta e toquei a campainha. Sim, o porta-malas havia sido deixado aberto acidentalmente.

Em outras circunstâncias, isso seria simplesmente um ato altruísta, indicando uma verdadeira preocupação com o bem-estar alheio, medo da culpa pela inatividade, desejo de demonstrar confiabilidade para si mesmo ou para os outros, ou ainda encontrar prazer no altruísmo. Aliás, considero todos esses motivos perfeitamente legítimos; eu poderia dar pontos extras para o primeiro, mas não deduziria nenhum ponto por penalidade para os outros. O importante é que as pessoas sejam ajudadas.

No meu caso, porém, uma vez que já trabalho no setor sem fins lucrativos, surge a questão adicional de saber se eu poderia ter empregado melhor esses sessenta segundos de forma mais especializada para trazer benefícios maiores aos outros. Ou seja, posso realmente justificar isso como o melhor uso do meu tempo, considerando as outras coisas nas quais afirmo acreditar?

A defesa evidente – ou talvez a racionalização óbvia – é que um ato de altruísmo como esse age como um restaurador da força de vontade de maneira muito mais eficiente do que, por exemplo, ouvir música. Também desconfio da minha capacidade de ser altruísta apenas teoricamente; suspeito que, se eu superar os problemas, meu altruísmo começará a desaparecer. Nunca fui tão longe para testá-lo; não parece valer o risco.

Mas se essa for a defesa, então meu ato não pode ser justificado como uma boa ação, certo? Pois esses são os benefícios auto-dirigidos que menciono.

Bem, quem disse que eu estava defendendo o ato como uma boa ação altruísta? É uma boa ação egoísta. Se isso restaura minha força de vontade, ou se me mantém altruísta, então há benefícios indiretos direcionados a outros (ou assim acredito). Claro, você pode argumentar que não confia em atos egoístas que deveriam ter outros benefícios como um "motivo oculto"; mas, pelo mesmo princípio, eu poderia responder que você deveria apenas olhar diretamente para a boa ação original, em vez de seu suposto motivo oculto.

Posso escapar impune? Ou seja, posso realmente chamar isso de "boa ação egoísta" e ainda assim obter a restauração da força de vontade, em vez de sentir culpa por ser egoísta? Parece que sim. Estou sur-

preso de que funcione dessa maneira, mas funciona. Desde que eu bata na porta para informar sobre o porta-malas aberto, e desde que a pessoa diga: “Obrigada!”, meu cérebro parece ter realizado sua maravilhosa boa ação do dia.

Sua experiência pode variar, é claro. O desafio ao tentar desenvolver uma arte de restauração da força de vontade é que diferentes abordagens parecem funcionar para diferentes pessoas. (Ou seja, estamos investigando os fenômenos superficiais sem compreender as regras mais profundas que também preveriam as variações.)

No entanto, se você perceber que é semelhante a mim neste aspecto – que ações altruístas ainda funcionam bem – então recomendo que você adquira “fuzzies” calorosos e “utilons” separadamente. Mas não simultaneamente. Tentar fazer ambas as coisas ao mesmo tempo significa que nenhuma delas acaba bem. Se o status é relevante para você, adquira o status separadamente também!

Se eu tivesse que aconselhar um bilionário recém-formado que está ingressando no campo da filantropia, meu conselho seria mais ou menos assim:

- Para obter *fuzzies* calorosos, encontre uma mulher trabalhadora, mas afligida pela pobreza, que esteja prestes a abandonar a faculdade estatal depois de a carga horária do seu marido ter sido reduzida, e pessoalmente, mas anonimamente, presenteie-a com um cheque administrativo de 10.000 dólares. Repita conforme desejado.

- Para conquistar status entre seus amigos, doe US$100 mil para o X-Prize mais destacado do momento ou para qualquer outra instituição de caridade que pareça oferecer mais prestígio pelo menor custo. Faça um grande alarde, participe de eventos para a imprensa e orgulhe-se disso pelos próximos cinco anos.

- Em seguida — com cálculos totalmente a sangue-frio — sem insensibilidade ao escopo ou aversão à ambiguidade — sem se preocupar com status ou fuzzies calorosos — identificando algum método comum para converter resultados em utilons e tentando expressar a incerteza em probabilidades percentuais — encontre a instituição de caridade que ofereça os maiores utilons esperados por dólar. Doe até o valor que você deseja doar para instituições de caridade, até que sua eficiência marginal caia abaixo da próxima instituição de caridade da lista.

Além disso, eu aconselharia o bilionário a considerar que o que eles gastam em utilons deveria ser pelo menos, digamos, 20 vezes o que gastam em *fuzzies* calorosos – uma sobrecarga de 5% para manter o altruísmo parece razoável, e eu, como seu juiz imparcial, não teria problemas em validar os *fuzzies* calorosos contra um multiplicador tão grande. Exceto se o ato inicialmente difuso realmente for benéfico, em vez de ativamente prejudicial.

(A compra de status me parece essencialmente não relacionada ao altruísmo. Se doar dinheiro para o X-Prize faz com que você seja mais admirado por seus amigos do que adquirir uma lancha com preço equivalente, então não há realmente razão para comprar a lancha. Apenas classifique o dinheiro na coluna “impressionante” da categoria “amigos” e esteja ciente de que esta não é a coluna do “altruísmo”.)

A principal lição aqui é que todas essas três coisas - *fuzzies* calorosos, status e utilons esperados - podem ser adquiridas com muito mais eficiência quando compradas separadamente, otimizando uma coisa de cada vez. Assinar um cheque de US$ 10 milhões para uma instituição de caridade de combate ao câncer de mama - embora seja mais louvável do que gastar os mesmos US$ 10 milhões em, sei lá, festas ou algo do tipo - não proporcionará a euforia concentrada de estar presente pessoalmente quando você transforma a vida de um único ser humano, provavelmente nem perto disso. Não causará tanto impacto em festas quanto doar para algo atraente como um X-Prize - talvez apenas um breve aceno de cabeça dos outros ricos. E se você ignorar completamente as preocupações com *fuzzies* calorosos e status, provavelmente existem pelo menos mil instituições de caridade pouco atendidas que poderiam gerar ordens de magnitude mais utilons com dez milhões de dólares. Tentar otimizar todos os três critérios de uma só vez apenas garante que nenhum deles seja otimizado de forma eficaz - apenas impulsos vagos ao longo das três dimensões.

É claro que, se você não é milionário ou mesmo bilionário, então não conseguirá ser tão eficiente em relação a essas coisas, não poderá comprar em abundância com tanta facilidade. Mas eu ainda diria:

para *fuzzies* calorosos, encontre uma instituição de caridade relativamente barata com beneficiários brilhantes, vívidos, de preferência pessoais e diretos. Voluntarie-se em uma cozinha comunitária. Ou simplesmente consiga seus *fuzzies* calorosos segurando portas abertas para idosos. Deixe que isso seja validado pelos seus outros esforços para adquirir utilons, mas não confunda isso com a aquisição real de utilons. Provavelmente é mais econômico obter status comprando roupas elegantes.

E quando se trata de adquirir os utilons esperados - então, é claro, cale a boca e multiplique.

326 — Apatia do espectador

O efeito espectador, também conhecido como apatia do espectador, manifesta-se quando grupos maiores têm menor probabilidade de agir em emergências - não apenas individualmente, mas de forma coletiva. Coloque um participante isolado em uma sala e deixe a fumaça começar a sair por baixo da porta. Setenta e cinco por cento dos participantes sairão para informar sobre o incidente. Agora, coloque três participantes na mesma sala - indivíduos reais, nenhum dos quais sabe o que está acontecendo. Em apenas 38% das ocasiões, alguém informará a fumaça. Se o participante estiver acompanhado por dois cúmplices que ignoram a fumaça, eles informarão apenas 10% das vezes, mesmo permanecendo na sala até que ela fique nebulosa. [1]

No modelo padrão, os dois principais impulsionadores da apatia do espectador são:

- Difusão de responsabilidade: todos esperam que outra pessoa tome a iniciativa e arque com os custos da ação. Quando ninguém age, fazer parte de uma multidão fornece uma desculpa e reduz a chance de ser pessoalmente responsabilizado pelos resultados.
- Ignorância pluralista: as pessoas tentam parecer calmas enquanto observam e buscam pistas... que os outros pareçam calmos.

Cialdini [2]:

> Muitas vezes, uma emergência não é claramente identificável. O indivíduo caído no beco pode ser vítima de um ataque cardíaco ou apenas um bêbado dormindo?... Em tempos de tanta incerteza, a tendência natural é observar as ações dos outros em busca de pistas. Podemos aprender pela forma como as outras testemunhas estão reagindo se o evento é ou não uma emergência. O que é fácil de esquecer, no entanto, é que todos os outros que observam o evento provavelmente também estarão em busca de evidências sociais. Como todos preferimos parecer equilibrados e serenos entre os outros, é provável que busquemos essa evidência tranquilamente, lançando olhares breves e discretos para aqueles ao nosso redor. Portanto, é provável que todos vejam os outros parecendo imperturbáveis e deixem de agir.

Cialdini sugere que, se você precisar de ajuda imediata, aponte para um único espectador e peça ajuda - deixando claro a quem você está se referindo. Lembre-se de que o grupo na totalidade, combinado, pode ter menos chances de ajudar do que um indivíduo.

Refleti um pouco sobre a psicologia evolutiva do efeito espectador. Suponha que, no ambiente ancestral, a maioria das pessoas do seu grupo provavelmente era pelo menos um pouco aparentada a você - o suficiente para valer a pena ser salva, se você fosse o único capaz de fazê-lo. No entanto, se houvessem outras duas pessoas presentes, a primeira a agir incorreria em um custo, enquanto as outras duas colheriam o benefício genético de salvar um parente parcial. Poderia ter havido uma competição para ver quem esperava mais?

Até onde segui essa linha de especulação, não parece ser uma explicação plausível - no momento em que todo o grupo é incapaz de agir, um gene que ajuda imediatamente deveria ser capaz de prevalecer, penso eu. O resultado experimental não é uma longa espera antes de ajudar, mas simplesmente a incapacidade de ajudar: se é um benefício genético ajudar quando você é a única pessoa que pode fazê-lo (como ocorre nos experimentos), então o equilíbrio do grupo não deveria ser prejudicado pela falta de ação de todos (como ocorre nos experimentos).

Portanto, não creio que uma competição para atrasar a ação seja uma explicação evolutiva plausível.

O mais provável, penso eu, é que estejamos diante de um problema não ancestral. Se os participantes experimentais realmente conhecem a vítima aparente, as chances de ajudar aumentam consideravelmente (ou seja, não estamos lidando com o equivalente de ajudar um membro real do grupo). Se bem me lembro, se os participantes experimentais se conhecem, as chances de ação também aumentam.

O nervosismo em relação à ação pública também pode desempenhar um papel. Se Robin Hanson estiver correto sobre o papel evolutivo da "sufocação", então ser o primeiro a agir numa emergência também pode ser encarado como uma tentativa perigosa de alcançar um status elevado. (Refletindo, não me recordo de ter visto a timidez discutida em análises do efeito espectador, mas talvez seja apenas minha memória fraca.)

Pode-se explicar o efeito espectador principalmente pela difusão da responsabilidade moral? Poderíamos ser cínicos e sugerir que as pessoas estão mais interessadas em não serem culpadas por não ajudarem do que em ter qualquer desejo positivo de ajudar. Elas desejam, em sua maioria, evitar o anti-heroísmo e possíveis represálias. Embora isso possa contribuir, duas observações mitigam essa ideia: (a) os participantes do experimento não relataram a presença de fumaça, mesmo que pudesse representar uma ameaça estritamente egoísta, e (b) informar as pessoas sobre o efeito espectador reduz o próprio efeito, mesmo que a probabilidade de serem responsabilizadas publicamente não aumente.

Na verdade, o efeito espectador é um dos principais casos que me vêm à mente de forma espontânea, no qual revelar um preconceito parece verdadeiramente capaz de enfraquecê-lo consideravelmente. Talvez isso ocorra porque a maneira apropriada de compensação é evidente, e não é fácil compensar excessivamente (como ao tentar ajustar a calibração, por exemplo). Assim, devemos ter cuidado para não sermos excessivamente cínicos sobre as implicações do efeito espectador e da difusão de responsabilidade, se interpretarmos a ação individual como uma tentativa fria e calculada de evitar censuras públicas. As pessoas parecem, pelo menos em alguns momentos, considerarem-se responsáveis quando percebem que são as únicas a ter conhecimento suficiente sobre o efeito espectador para agir.

Embora eu me questione sobre o que acontece se você souber que faz parte de uma multidão onde todos foram informados sobre o efeito espectador...

Referências

[1] Bibb Latané and John M. Darley, "Bystander 'Apathy,"' American Scientist 57, no. 2 (1969): 244– 268, http://www.jstor.org/stable/27828530.

[2] Cialdini, Influence.

327 — Apatia coletiva e a Internet

No meu último ensaio, tratei do efeito espectador, também conhecido como apatia do observador: diante de uma situação problemática fixa, é menos provável que um grupo de espectadores aja, em comparação a um único observador. A explicação convencional para esse resultado é baseada na ignorância pluralista (quando não está claro se a situação é uma emergência, cada pessoa tenta parecer calma ao olhar para os outros espectadores e perceber que todos aparentam calma) e na difusão de responsabilidade (todos esperam que alguém seja o primeiro a agir; fazer parte de uma multidão reduz a pressão individual a ponto de ninguém agir).

Isso pode ser um sintoma de que nossos mecanismos de coordenação, originários dos caçadores-coletores, estão sendo desafiados pelas condições modernas. Normalmente, não se formavam grupos de trabalho com estranhos no ambiente ancestral; eram principalmente compostos por pessoas conhecidas. Na verdade, quando todos os participantes se conhecem, o efeito espectador diminui.

Assim, percebo que esta é uma observação surpreendente e revolucionária, e espero não chocar nenhum leitor ao afirmar isso: as pessoas parecem ter dificuldade em reagir de maneira construtiva aos problemas encontrados na Internet.

Possivelmente, nossos instintos inatos de coordenação não estão sintonizados para:

- Integrar-se a um grupo de desconhecidos. (Quando todos se conhecem, o efeito espectador diminui.)
- Fazer parte de um grupo de tamanho desconhecido, composto por estranhos de identidade desconhecida.
- Não estar em contato físico (ou visual); incapazes de trocar olhares significativos.
- Não se comunicar em tempo real.
- Não depender mutuamente por outras formas de ajuda; não ser co-dependente do grupo ao qual pertence.
- Estar protegido contra danos à reputação, ou contra o medo desses danos, devido ao seu aparente anonimato; ninguém está visivelmente observando você, diante de quem sua reputação pode sofrer por inação.
- Fazer parte de um grande coletivo de outros inativos; ninguém apontará o dedo culpando você.
- Não ouvir um pedido de ajuda.

E assim por diante. Não tenho uma solução brilhante para esse problema. Mas é algo que eu gostaria que os potenciais co-fundadores de empresas online considerassem explicitamente, em vez de se perguntarem como atrair atenção no Facebook. (Sim, estou de olho em você, Hacker News.) Existem aplicativos de ativismo online, mas geralmente seguem a linha de “assinem esta petição! sim, você assinou algo!” em vez de pensar em como neutralizar o efeito espectador, restaurar a motivação e trabalhar com os instintos de coordenação em grupo pela Internet?

Algumas sugestões que vêm à mente:

- Publicar um vídeo online de alguém pedindo ajuda.
- Exibir nomes e fotos, ou até mesmo breves vídeos se disponíveis, das primeiras pessoas que ajudaram (ou ter algum algoritmo de prioridade semelhante ao Reddit que dependa de uma combinação de valor ajudado e tempo recente).
- Oferecer aos ajudantes um vídeo de agradecimento do fundador da causa, que eles podem compartilhar em sua página “pessoas que ajudei”, que, com padronização suficiente, poderia ser parcial ou totalmente montado de forma automática e facilmente incorporado em sua página inicial ou no Facebook
- Encontrar uma expressão não irritante para “conte a um amigo sobre a causa X”; permitir códigos de link de referência; em seguida, mostrar às pessoas quantas outras pessoas elas evangelizaram (quantas pessoas que inicialmente chegaram aqui usando o código de referência X realmente contribuíram ou realizaram alguma outra ação).
- (Todas as sugestões acima aplicam-se não apenas a doações, mas a projetos de código aberto para os quais as pessoas contribuíram com código. Ou se as pessoas realmente não querem nada além de assinaturas em uma petição, então por assinaturas. Existem maneiras de ajudar além do dinheiro - mesmo que o dinheiro seja geralmente o mais eficaz. O ponto principal é que a forma de ajuda deve ser verificável online.)
- Facilitar para as pessoas oferecerem recompensas monetárias em sub tarefas cujo desempenho seja verificável.

Mas, acima de tudo, apresento a vocês um problema aberto e não resolvido: permitir/tornar mais fácil para grupos de estranhos se unirem em uma força-tarefa eficaz pela Internet, desafiando os modos de falha usuais e as razões padrão pelas quais isso não é possível - um problema ancestral. Pense naquela velha estatística sobre a Wikipédia representar 1/2.000 do tempo gasto apenas nos EUA assistindo televisão. Há muito potencial lá fora, se ao menos existisse um motor eficaz...

328 — Progresso incremental e o vale

Racionalidade é a conquista sistematizada.

“Mas”, você questiona, “a pessoa racional nem sempre sai vitoriosa!”

O que você quer dizer com isso? Você está se referindo ao fato de que, a cada semana ou duas, alguém que comprou um bilhete de loteria com valor esperado negativo acaba ganhando na loteria e ficando significativamente mais rico do que você? Isso não é uma perda sistemática; é uma seleção tendenciosa da mídia. Do ponto de vista estatístico, não há vencedores de loteria – você nunca encontraria um na sua vida se não fosse pelas reportagens seletivas.

Mesmo agentes perfeitamente racionais podem enfrentar derrotas. A diferença é que eles não conseguem prever antecipadamente que serão derrotados. Eles não podem esperar um desempenho inferior a qualquer outra estratégia executável, ou simplesmente a seguiriam.

“Não”, você argumenta, “estou me referindo ao enriquecimento dos fundadores de startups que acreditam em si e em suas ideias mais fervorosamente do que qualquer pessoa racional faria. Estou falando sobre como as pessoas religiosas encontram mais felicidade...”

Entendi. Bem, o problema aqui é o seguinte: um avanço incremental em direção à racionalidade, se o resultado ainda for irracional de outras formas, não necessariamente resultará em ganhos crescentes.

Os teoremas de otimalidade que temos para a teoria da probabilidade e a teoria da decisão aplicam-se a uma probabilidade perfeita e a uma tomada de decisão perfeita. Não há um teorema correspondente que afirme que, partindo de uma forma inicial falha, cada modificação incremental do algoritmo que aproxima a estrutura do ideal deve resultar em uma melhoria incremental no desempenho. Isso ainda não foi comprovado, pois na verdade não é verdade.

“Então”, você pondera, “qual é o sentido de nos esforçarmos para ser mais racionais? Não atingiremos a perfeição ideal. Portanto, não temos garantia de que nossos passos à frente estejam contribuindo.”

Você também não tem garantia de que um passo atrás o ajudará a vencer. Garantias não existem no mundo real; mas, ao contrário do que muitos pensam erroneamente, tomar decisões sob incerteza é a essência da racionalidade.

“Mas temos vários casos em que, com base em raciocínios que parecem vagamente plausíveis ou em dados de pesquisas, parece que avançar em termos de racionalidade pode nos prejudicar. Se o que realmente importa é a vitória – se você tem algo mais importante para proteger do que qualquer ritual cognitivo – então, por que dar esse passo?”

Ah, e agora chegamos ao cerne da questão.

Não posso falar por todos, mas...

Minha primeira razão é que, no âmbito profissional, lido com problemas profundamente complexos que exigem uma precisão de pensamento imensa. Um pequeno erro pode desviar o curso por anos, e há penalidades mais severas aguardando nos bastidores. Manter um nível de desempenho inalterado não é suficiente; minha escolha é tentar superar-me ou desistir e voltar para casa.

"Mas isso é apenas a sua experiência. Nem todos nós vivemos dessa maneira. E se estiver apenas lidando com tarefas comuns, como iniciar uma startup na internet?"

Minha segunda razão é que estou tentando levar alguns aspectos da minha arte para além do que já vi. Não sei onde essas melhorias podem me levar. A perda de não dar um passo à frente não é apenas esse passo, mas todos os outros que poderiam ter sido dados além desse ponto. Como diz Robin Hanson: o problema de escorregar na escada não é cair da altura do primeiro degrau; é que, ao cair um degrau, pode-se cair outro. Recusar-se a subir um degrau não significa perder apenas a altura desse degrau, mas a altura total da escada.

"Mas, novamente, isso é particular da sua busca. Nem todos estão tentando explorar territórios desconhecidos na arte."

Minha terceira razão é que, quando percebo que fui enganado, não posso simplesmente fechar os olhos e fingir que não vi. Já dei esse passo adiante; negar isso a mim mesmo seria inútil. Não poderia acreditar em Deus mesmo que tentasse, da mesma forma que não poderia acreditar que o céu acima de mim é verde olhando diretamente para ele. Se você tem todo o conhecimento necessário para perceber que é melhor enganar a si mesmo, é tarde demais para fazê-lo.

"Mas essa percepção é rara; muitas pessoas têm mais facilidade em ignorar a impossibilidade. Você, por outro lado, parece estar ativamente patrocinando o colapso do duplipensamento. De um ponto de vista mais elevado, você pode entender o suficiente para prever que isso os tornará mais infelizes. Isso é resultado de um desejo sádico de prejudicar seus leitores, ou o quê?"

Então, finalmente, respondo que minha experiência até agora - mesmo dentro dessas possibilidades meramente humanas - sugere que, uma vez que você se compromete um pouco e não está cometendo muitos outros erros, buscar mais racionalidade realmente o deixará em um estado melhor. A longa jornada leva para fora do vale e a uma altura maior do que antes, mesmo nas terras humanas.

Quanto mais eu me aprofundo em uma faceta específica da Arte, mais consigo perceber que é assim. Como mencionei anteriormente, meus ensaios podem não refletir o que seria uma verdadeira arte marcial da racionalidade, pois eu me concentrei apenas em responder a perguntas confusas - não em lidar com a acrasia, coordenar grupos ou buscar a felicidade. Na esfera de responder a questões confusas - onde dediquei a maior parte da minha prática na Arte - torna-se agora extremamente claro que qualquer pessoa que pensasse que seria melhor "manter-se otimista na resolução do problema" estaria enganada. Isso seria para um estudante casual.

Quando se trata de manter a motivação ou buscar a felicidade, não posso assegurar que alguém que perde as ilusões estará em uma situação melhor - pois meu conhecimento dessas facetas da racionalidade ainda é básico. Não sei se essas partes da Arte foram desenvolvidas de forma sistemática. No entanto, mesmo aqui, fiz um esforço considerável para dissipar ideias meio racionais e meio equivocadas que poderiam atrapalhar um iniciante, como a noção de que a racionalidade se opõe ao sentimento, ou a ideia de que a racionalidade é contrária ao valor, ou a concepção de que pensadores sofisticados devem ser angustiados e cínicos.

E se, como espero, alguém desenvolver a arte de combater a acrasia ou alcançar o bem-estar mental tão profundamente quanto desenvolvi a arte de responder a perguntas impossíveis, espero sinceramente que aqueles que se afundam em suas ilusões não comecem a competir. Enquanto isso, outros podem se sair melhor do que eu, se a felicidade for o seu maior desejo, pois eu mesmo investi pouco esforço nesse aspecto.

Acho difícil acreditar que o indivíduo perfeitamente motivado, o empreendedor mais forte que um ser humano pode se tornar, ainda esteja envolto em um cobertor reconfortante de excesso de confiança. Acredito que, provavelmente, eles jogaram o cobertor pela janela e organizaram a mente de maneira um pouco diferente. Acho difícil acreditar que a máxima felicidade que podemos experimentar, mesmo dentro dos limites das possibilidades humanas, envolva uma pequena consciência escondida no canto da mente de que tudo é uma mentira. Prefiro depositar minhas esperanças no neurofeedback ou na meditação Zen, embora eu não tenha tentado nenhum dos dois.

Mas não se pode negar que esta é uma questão muito real na vida cotidiana. Considere este par de comentários do *Less Wrong*:

Vou ser honesto: minha vida decaiu significativamente desde que me desvinculei. Minha namorada teísta, por quem eu estava profundamente apaixonado, não conseguiu lidar com essa mudança em mim e, após seis meses de dolorosa indecisão, me trocou por um colega de trabalho. Isso aconteceu há seis meses, e desde então estou com o coração partido, infeliz, sem foco e extremamente ineficaz.

Este pode ser um exemplo do vale da má racionalidade, conforme mencionado por Phil Goetz, mas ainda considero minha situação atual preferível do que a felicidade baseada em falsas crenças.

E:

Minhas empatias: isso aconteceu comigo cerca de 6 anos atrás (embora felizmente sem muitas oscilações visíveis).

Minha irmã, que tinha algum treinamento em Terapia Cognitivo-Comportamental, lembrou-me de que os relacionamentos estão em constante formação e ruptura, e, dado que eu não era abominável e não tinha me isolado em reclusão monástica, não era razoável pensar que ficaria sozinho pelo resto da vida (ela acabou por estar certa). Essa perspectiva foi útil nos momentos em que meus sentimentos não estavam totalmente sob controle.

Portanto, na prática, na vida real, esses primeiros passos podem ser realmente dolorosos. Mas as coisas podem, de fato, melhorar. E, de fato, não há garantia de que você terminará em uma posição melhor. Mesmo que, em teoria, o caminho deva levar a melhorias, não há garantia de que cada pessoa alcance esse ponto.

Se você não prefere a verdade mais do que a felicidade baseada em crenças falsas...

Bem... e se você não estiver envolvido em algo particularmente arriscado ou confuso... e se você não estiver comprando bilhetes de loteria... e se já estiver inscrito na criônica, um teste ácido repentino e desafiador de racionalidade, confuso e de altíssimo risco que ilustra a natureza do Cisne Negro ao apostar na ignorância em ignorância.

Então, não é garantido que seguir todos os passos incrementais em direção à racionalidade que você encontrar o deixará em uma situação melhor. Os argumentos que parecem vagamente plausíveis contra a perda das ilusões muitas vezes consideram apenas um único passo, sem sugerir quaisquer passos adicionais, sem mencionar qualquer tentativa de recuperar o que foi perdido e melhorar um pouco mais. Mesmo as pesquisas comparam a pessoa religiosa média com o ateu médio, não os teólogos mais avançados com os racionalistas mais avançados.

Mas se você não se importa com a verdade - e não tem nada a proteger - e não se sente atraído pela ideia de levar sua arte o mais longe possível - e sua vida atual está indo bem - e você sente que seu bem-estar mental depende de ilusões que prefere não questionar -

Então, provavelmente, você não está lendo isso. Mas se estiver, então eu suponho que…bem... (a) inscreva-se na criônica e, em seguida, (b) pare de ler *Less Wrong* antes que suas ilusões desmoronem! FUJA!

329 — Bayesianos vs. Bárbaros

Anteriormente:

Vamos imaginar que tenhamos dois grupos de soldados. No Grupo 1, os soldados rasos não têm conhecimento de táticas e estratégias; apenas os sargentos têm alguma compreensão tática, e apenas os oficiais têm algum entendimento de estratégia. No Grupo 2, todos, em todos os níveis, possuem amplo conhecimento tanto em táticas quanto em estratégias.

Poderíamos esperar que o Grupo 1 vencesse o Grupo 2, pois o Grupo 1 seguirá as ordens, enquanto todos no Grupo 2 apresentam ideias superiores a qualquer ordem dada a eles?

Neste caso, questiono até que ponto o Grupo 2 realmente compreende a teoria militar, pois é uma proposição elementar que uma multidão descoordenada seja massacrada.

Vamos supor que um país de racionalistas seja atacado por um país de Bárbaros Malignos que não sabem nada sobre teoria da probabilidade ou teoria da decisão.

Agora, há um ponto de vista sobre "racionalidade" ou "racionalismo" que diria algo assim:

"Obviamente, os racionalistas perderão. Os Bárbaros acreditam em uma vida após a morte onde serão recompensados pela coragem; assim, eles se lançarão na batalha sem hesitação ou remorso. Graças às suas espirais de morte afetiva em torno de sua Causa e do Grande Líder Bob, seus guerreiros obedecerão às ordens e seus cidadãos nacionais produzirão com entusiasmo e com total capacidade para a guerra; qualquer pessoa pega roubando ou se escondendo será queimada na fogueira, conforme a tradição bárbara. Eles acreditarão na bondade um do outro e odiarão o inimigo com mais força do que qualquer pessoa sã, unindo-se a um grupo coeso. Enquanto isso, os racionalistas perceberão que não há recompensa concebível por morrer em batalha; eles desejarão que os outros lutem, mas eles mesmos não vão querer lutar. Mesmo que consigam encontrar soldados, seus civis não serão tão cooperativos: desde que qualquer vantagem não leve ao colapso do esforço de guerra, eles vão querer ficar com ela para si, e assim não contribuir tanto quanto poderiam. Não importa quão refinada, elegante, civilizada, produtiva e não violenta seja sua cultura, eles não serão capazes de resistir à invasão bárbara; uma argumentação sensata não é páreo para um lunático espumante armado com uma arma. No final, os Bárbaros vencerão porque desejam lutar, desejam ferir os racionalistas, desejam conquistar, e toda a sua sociedade está unida em torno da conquista; eles se importam mais com isso do que qualquer pessoa sã.

A guerra não é divertida. Como muitas, muitas pessoas descobriram desde o início da história registrada, como muitas, muitas pessoas descobriram antes do início da história registrada, como alguma comunidade em algum lugar está descobrindo neste momento, em um pequeno e triste país cujas agonias internas nem sequer fazem mais as primeiras páginas.

A guerra não é divertida. Perder uma guerra é ainda menos divertido. E foi dito desde os tempos antigos: 'Se queres ter paz, prepara-te para a guerra'. Seus oponentes não precisam acreditar que você vencerá, que conquistará; mas eles têm que acreditar que você resistirá o suficiente para não valer a pena.

Percebe-se, então, que se o destino genuíno dos "racionalistas" fosse sempre perder na guerra, eu não poderia, em boa consciência, apoiar a adoção pública generalizada da "racionalidade".

Este é, sem dúvida, um dos tópicos mais delicados que discuti ou pretendo discutir aqui. A guerra

não é algo limpo. As atuais forças armadas de alta tecnologia - referindo-me aqui às forças armadas dos EUA - destacam-se pela força esmagadora que podem exercer sobre os oponentes, possibilitando um grau historicamente extraordinário de preocupação com as baixas inimigas e civis.

Vencer na guerra nem sempre implicou abandonar toda a moralidade. Guerras foram vencidas sem recorrer à tortura. A brutalidade da guerra não implica, por exemplo, que questionar o Presidente seja antipatriótico. Estamos acostumados a ver a "guerra" sendo usada como desculpa para comportamentos inadequados, pois é exatamente para isso que ela tem sido utilizada na história recente dos EUA...

Mas a estupidez reversa não é inteligência. O mal reverso também não é inteligência. Continua sendo verdade que as verdadeiras guerras não podem ser vencidas através de polidez refinada. Se os "racionalistas" não conseguirem se preparar para esse choque mental, os Bárbaros realmente vencerão; e os "racionalistas"... Não quero dizer que "merecem perder", mas terão falhado no teste da existência de sua sociedade.

Permita-me começar descartando a ideia de que, em princípio, agentes racionais ideais não podem travar uma guerra, porque cada um deles prefere ser civil a ser soldado.

Como já discutido extensivamente, apresento uma caixa sobre o problema de Newcomb.

Consistentemente, não acredito que se uma eleição for decidida por 100.000 a 99.998 votos, todos os eleitores foram irracionais ao se esforçarem para ir às urnas porque "ficar em casa não teria afetado o resultado". (Também não acredito que se a eleição resultasse em 100.000 a 99.999, então 100.000 pessoas seriam todas, individualmente, as únicas responsáveis pelo resultado.)

De forma consistente, defendo também que duas IAs racionais (que usam o meu tipo de teoria da decisão), mesmo se tivessem funções de utilidade completamente diferentes e fossem projetadas por criadores diferentes, cooperariam no verdadeiro Dilema do Prisioneiro se tivessem conhecimento mútuo do código-fonte uma da outra. (Ou mesmo apenas o conhecimento mútuo da racionalidade de cada uma no sentido apropriado).

De maneira consistente, defendo a ideia de que agentes racionais são capazes de coordenar projetos em grupo sempre que o resultado esperado (probabilisticamente) seja superior ao que seria sem essa coordenação. Uma sociedade de agentes que adote minha teoria da decisão e compartilhe o conhecimento comum desse fato resultará em ótimos de Pareto, em vez de equilíbrios de Nash. Se todos os agentes racionais concordarem que é preferível lutar a se render, eles enfrentarão os desafios em vez de capitularem diante dos obstáculos.

Imagine uma comunidade de inteligências artificiais auto-ajustáveis que coletivamente prefira lutar a se render, mas individualmente prefira ser civil a lutar. Uma solução seria realizar uma loteria, imprevisível para qualquer agente, para selecionar guerreiros. Antes do sorteio, todas as inteligências artificiais modificam seu código antecipadamente, de modo que, se forem selecionadas, lutem como guerreiros da maneira mais eficiente possível em termos comunitários, mesmo que isso implique marchar calmamente para a própria morte.

(Uma teoria da decisão reflexivamente consistente opera da mesma forma, apenas sem a auto modificação.)

Você poderia argumentar: "Mas no mundo humano real, os agentes não são perfeitamente racionais, nem têm conhecimento comum do código-fonte uns dos outros. A cooperação no Dilema do Prisioneiro requer certas condições, conforme a sua teoria de decisão (que estas margens são muito pequenas para conter) e essas condições não são satisfeitas na vida real."

Eu respondo: O verdadeiro Dilema do Prisioneiro puro é incrivelmente raro na vida real. Geralmente, há efeitos indiretos - o que você faz afeta sua reputação. Na realidade, a maioria das pessoas se importa, até certo ponto, com o que acontece com os outros. E na vida real, você tem a oportunidade de criar mecanismos de incentivo.

Além disso, na vida real, acredito que uma comunidade de racionalistas humanos seria capaz de gerar soldados dispostos a sacrificar-se para defender a comunidade. Contanto que não seja ensinado às crianças na escola que os racionalistas ideais deveriam trair uns aos outros no Dilema do Prisioneiro. É preciso que

acreditem amplamente - e eu acredito nisso, exatamente pela mesma razão que abordei o problema de Newcomb - que se as pessoas decidirem, como indivíduos, não serem soldados, ou se os soldados decidirem fugir, isso seria o mesmo que decidir pela vitória dos bárbaros. Segundo a mesma teoria que afirma que, se uma eleição for vencida por 100.000 votos contra 99.998 votos, não faz sentido que todos os eleitores digam "meu voto não fez diferença". Digamos (pois é verdade) que as funções de utilidade não precisam ser solipsistas, e um agente racional pode lutar até a morte se ele se importar o suficiente com o que está protegendo. Não lhes diga que os racionalistas deveriam aceitar perder razoavelmente.

Se esta for a cultura e os costumes da sociedade racionalista, então, acredito que os cidadãos comuns dessa sociedade se voluntariariam para serem soldados. Afinal, isso parece ser intrínseco aos seres humanos. A única necessidade é garantir que a formação cultural não atrapalhe.

E se eu estiver errado e isso não fornecer voluntários suficientes?

Nesse caso, enquanto as pessoas ainda optarem, em sua maioria, por lutar em vez de se renderem, terão a oportunidade de criar mecanismos de incentivo e evitar o Dilema do Prisioneiro Verdadeiro.

Pode-se realizar sorteios para selecionar os guerreiros, algo semelhante ao exemplo mencionado anteriormente, com IAs alterando seu próprio código. No entanto, se "ser reflexivamente consistente; fazer o que você se comprometeu previamente a fazer" não for motivação suficiente para os humanos seguirem o sorteio, então...

Bem, antes de realizar o sorteio, talvez todos possamos concordar que é uma boa ideia fornecer aos selecionados medicamentos que induzam coragem extra e, se fugirem, puni-los com a morte. Mesmo considerando que nós mesmos podemos ser selecionados no sorteio. Porque, antes da realização do sorteio, essa é a política geral que nos dá a maior expectativa de sobrevivência.

Como eu disse: Guerras reais = não são divertidas, e perder guerras = ainda menos divertido.

Vale ressaltar, aliás, que não estou defendendo o alistamento obrigatório da forma como é praticado atualmente. Esses projetos não representam tentativas coletivas da população para transitar de um equilíbrio de Nash para um ótimo de Pareto. Os projetos de recrutamento são ferramentas dos governantes que necessitam de soldados como peças de um tabuleiro de xadrez. Aqueles recrutados na época do Vietnã que fugiram para o Canadá, considero que estavam corretos. Contudo, uma sociedade que se considera demasiadamente inteligente para ter reis não precisa ser excessivamente inteligente para sobreviver, mesmo diante da invasão de hordas bárbaras que praticam o recrutamento.

Os soldados racionais obedeceriam às ordens? E se o comandante cometesse um erro?

Os soldados marcham. Os pés de todos batem no chão no mesmo ritmo. Mesmo que, talvez, isso vá contra suas próprias inclinações, já que pessoas deixadas por si mesmas caminhariam cada uma com passos separados. Lasers feitos de pessoas. Isso é marchar. Se houver algum método de tomada de decisão em grupo superior à simples emissão de ordens pelo capitão, então uma companhia de soldados racionais poderia adotar esse procedimento. Se não existir um método comprovadamente melhor do que ter um capitão, então uma companhia de soldados racionais compromete-se a obedecer ao capitão, mesmo contra suas próprias inclinações. E se os seres humanos não forem tão racionais... então, antes de recorrer à sorte, a política geral que oferece a maior expectativa pessoal de sobrevivência é atirar nos soldados que desobedecem às ordens. Isso não significa que aqueles que se rebelaram contra seus próprios oficiais no Vietnã estavam errados; pois poderiam ter consistentemente argumentado que preferiam não participar do sorteio.

Mas uma multidão descoordenada é massacrada, e, por isso, os soldados precisam de alguma forma de fazer todos realizarem a mesma ação simultaneamente, na busca do mesmo objetivo, mesmo que, se deixados à própria sorte, poderiam marchar em todas as direções. As ordens podem não vir de um capitão como um chefe tribal superior, mas as ordens unificadas têm que vir de algum lugar. Uma sociedade cujos soldados são demasiado inteligentes para obedecer às ordens é uma sociedade demasiado inteligente para sobreviver. Assim como uma sociedade cujas pessoas são demasiado inteligentes para serem soldados. É por isso que uso "inteligente", muitas vezes como termo de desaprovação, em vez de "racional".

(Apesar de eu achar importante questionar se é possível criar um método de coordenação para pe-

quenos grupos que realmente funcione melhor na prática do que ter um líder. Quanto mais as pessoas puderem confiar no método de decisão em grupo, mais poderão acreditar que é realmente superior a seguir o próprio caminho – e mais consistentemente poderão se comportar, mesmo na ausência de penalidades aplicáveis por desobediência.)

Digo tudo isso, mesmo que embora certamente não espere que os racionalistas assumam o controle de um país tão cedo, porque penso que o que acreditamos sobre uma sociedade de "pessoas como nós" reflete em alguma medida o que pensamos de nós mesmos. Se você acredita que uma sociedade de pessoas como você seria razoável demais para sobreviver a longo prazo... isso é uma questão de autoimagem. E é uma autoimagem diferente se você pensar que uma sociedade de pessoas como você poderia enfrentar os cruéis Bárbaros do Mal e vencer – não apenas por meio de tecnologia superior, mas porque seu povo se preocupa uns com os outros e com sua sociedade coletiva. - e porque podem encarar as realidades da guerra sem se perderem - e porque calculariam a coisa racional a fazer em grupo e garantiriam que fosse feito - e porque não há nada nas regras da teoria da probabilidade ou da teoria da decisão que diga que você não pode se sacrificar por uma causa – e porque, se você realmente é mais esperto que o Inimigo e não apenas se vangloria disso, então deveria ser capaz de explorar os pontos cegos que o Inimigo não permite a si mesmo pensar – e porque não importa o quanto o Inimigo se exalte antes da batalha, você pensa que talvez uma mente coesa, indivisa em si, e talvez praticando algo semelhante à meditação ou auto-hipnose, pode lutar tão arduamente na prática quanto alguém que teoricamente acredita ter setenta e duas virgens esperando por ele.

Então você espera mais de si mesmo e das pessoas como você operando em grupos; e então você pode se ver como algo mais do que um beco sem saída cultural.

Veja desta forma: Jeffreyssai provavelmente não desistiria dos Bárbaros Malignos se estivesse lutando sozinho. Um exército inteiro de mestres *beisutsukai* deveria ser uma força com a qual ninguém mexeria. Essa é a visão motivadora. A questão é como, exatamente, isso funciona.

330 — Cuidado com a otimização de outros

Observo um problema sério em que os aspirantes a racionalistas superestimam grandemente sua capacidade de otimizar a vida de outras pessoas. E acredito ter alguma ideia de como o problema surge.

Ao ler dezenove páginas diferentes que oferecem conselhos sobre melhoria pessoal – produtividade, dieta, economia de dinheiro – todos os escritores parecem brilhantes e entusiasmados com seus métodos, contando histórias de como funcionaram para eles e prometendo resultados surpreendentes...

Mas a maioria dos conselhos soa tão falso que nem parece valer a pena considerá-los. Você suspira, refletindo tristemente sobre o entusiasmo selvagem e infantil que as pessoas parecem ter por praticamente qualquer coisa, não importa o quão bobo seja. Os conselhos de número 4 e 15 parecem interessantes, então você os experimenta, mas... eles não... realmente... bem, acabam falhando miseravelmente. O conselho estava errado, ou você não pôde executá-lo e, de qualquer forma, você não está em uma situação melhor.

Então, você lê o vigésimo conselho – ou descobre um vigésimo método que não estava em nenhuma das páginas – e, SURPREENDENTEMENTE, FUNCIONA DE VERDADE DESTA VEZ.

Finalmente, você descobriu o caminho real, a abordagem correta, o método que realmente dá certo. E quando alguém se depara com o mesmo problema que você costumava ter, bem, desta vez você sabe como ajudá-lo. Você pode evitar que eles passem pelo trabalho de ler dezenove conselhos inúteis e ir diretamente para a resposta correta. Como aspirante a racionalista, você já aprendeu que a maioria das pessoas não escuta e geralmente não se importa – mas essa pessoa é um amigo, alguém que você conhece, alguém em quem você confia e respeita para ouvir.

Então, você coloca uma mão solidária em seus ombros, olha-os diretamente nos olhos e explica como fazer.

Eu, pessoalmente, passo muito por isso. Porque, veja bem... quando você descobre como isso realmente funciona... bom, você já sabe que não deve sair correndo e contar para seus amigos e familiares. Mas você precisa tentar contar para Eliezer Yudkowsky. Ele precisa disso, e há uma boa chance de que ele entenda.

Na verdade, levei um tempo para entender. Um dos eventos cruciais foi quando alguém do Instituto de Pesquisa em Inteligência de Máquina me disse que eu não precisava de um aumento salarial para acompanhar a inflação – porque eu poderia gastar consideravelmente menos dinheiro em comida usando um serviço de cupons online. E eu acreditei nisso, porque era um amigo em quem confiava, e foi entregue com muita confiança. Então, minha namorada tentou usar o serviço e, algumas semanas depois, desistiu.

Agora, aqui está a questão: se eu tivesse encontrado exatamente o mesmo conselho sobre o uso de cupons em algum blog em algum lugar, provavelmente nem teria prestado muita atenção, apenas lido e seguido em frente. Mesmo que tivesse sido escrito por Scott Aaronson ou alguém semelhante conhecido por ser inteligente, eu ainda o teria lido e seguido em frente. Mas, como me foi entregue pessoalmente, por um amigo que eu conhecia, meu cérebro processou de forma diferente – como se o segredo estivesse sendo compartilhado comigo; e esse foi realmente o tom com que me foi entregue. E foi uma reação tardia perceber que simplesmente me contaram, como um conselho pessoal, o que de outra forma teria sido apenas uma postagem de blog em algum lugar; nem mais, nem menos provável de funcionar para mim do que uma postagem em um blog de produtividade escrita por qualquer outra pessoa inteligente.

E como já encontrei muitas pessoas tentando me otimizar, posso garantir que os conselhos que rece-

bo são tão variados quanto a blogosfera de produtividade. No entanto, alguns não enxergam essa diversidade de conselhos sobre produtividade como um reflexo da individualidade das pessoas em relação ao que funciona para cada uma delas. Ao invés disso, eles identificam muitos conselhos ruins, claramente equivocados. E, finalmente, descobrem o caminho correto - a abordagem que realmente funciona, ao contrário de todas as outras postagens de blog que não surtem efeito - e, muitas vezes, decidem aplicá-lo para otimizar Eliezer Yudkowsky.

Não me interpretem mal. Às vezes, o conselho é útil. Às vezes, funciona. "Stuck In The Middle With Bruce" - isso ressoou para mim. Pode ser a coisa mais valiosa que li sobre o novo *Less Wrong* até agora, embora isso ainda não tenha sido definitivamente estabelecido.

Acontece que o seu conselho pessoal sincero, aquela descoberta incrível que você fez e que realmente funciona, caramba, não tem mais nem menos probabilidade de funcionar para mim do que uma postagem aleatória de blog sobre aprimoramento pessoal escrita por um autor inteligente provavelmente funcionará para você.

"Coisas diferentes funcionam para pessoas diferentes." Essa frase pode causar um estranhamento; eu sei que isso me causa. Por essa frase ser muitas vezes uma ferramenta usada pela Epistemologia do Lado Negro para se proteger contra críticas, sendo empregada de maneira semelhante a "Coisas diferentes são verdadeiras para pessoas diferentes" (o que é simplesmente falso).

Mas até que você compreenda as leis que são generalizações quase universais, às vezes você acaba recorrendo a truques superficiais que funcionam para uma pessoa e não para outra, sem compreender o porquê, pois você não está ciente das leis gerais que determinariam o que funciona para quem. E o melhor que você pode fazer é estar ciente disso e estar disposto a aceitar um "Não" como resposta.

E, especialmente, é crucial que você esteja disposto a aceitar um "Não" como resposta se tiver poder sobre o Outro. O poder, em geral, é algo extremamente perigoso, do qual é muito fácil abusar sem perceber que está fazendo isso. Existem medidas que você pode adotar para evitar o abuso de poder, mas você precisa realmente implementá-las, caso contrário, não terão efeito. Lembro-me de uma postagem no *Overcoming Bias* que demonstrava como estar em uma posição de poder diminui nossa capacidade de ter empatia e compreender o outro, embora eu não consiga encontrá-la agora. Já vi um racionalista que não acreditava ter poder e, portanto, não achava necessário ser cauteloso, surpreendido ao descobrir que poderia ser temido...

É ainda pior quando a descoberta que funciona para eles exige um pouco de força de vontade. Então, se você disser que não funciona para você, a resposta é clara e óbvia: você está apenas sendo preguiçoso, e eles sentem a necessidade de exercer pressão sobre você para seguir o conselho que descobriram que realmente funciona.

Às vezes — presumo — as pessoas estão sendo um tanto preguiçosas. No entanto, é crucial ter muito, muito cuidado antes de assumir que este é o caso e exercer poder sobre os outros para «colocá-los em movimento». Líderes que conseguem discernir quando algo está realmente ao alcance de alguém, se apenas estiver um pouco mais motivado, sem esgotá-lo ou tornar sua vida incrivelmente dolorosa - esses são os líderes com os quais é um prazer trabalhar. Essa habilidade é extraordinariamente rara, e os líderes que a possuem valem seu peso em ouro. Trata-se de uma habilidade interpessoal de alto nível que a maioria das pessoas não possui. Eu certamente não a possuo. Não assuma que você a possui apenas porque suas intenções são boas. Não presuma que a tem só porque você nunca faria algo aos outros que não gostaria que fosse feito a si mesmo. Não assuma que a possui porque ninguém jamais reclamou com você. Talvez eles estejam apenas com medo. Aquele racionalista do qual falei – que não achava que detinha poder e ameaça, embora isso fosse bastante óbvio para mim – não percebeu que alguém poderia ter medo dele.

Tenha cautela mesmo quando você detém vantagem, quando possui uma decisão importante, uma ameaça, ou algo que a outra pessoa precisa e, de repente, a tentação de otimizá-los parece irresistível.

Considere, se desejar, que todo o reinado de terror de Ayn Rand sobre os objetivistas pode ser compreendido exatamente sob esta perspectiva – ela se viu com poder e influência e não conseguiu resistir à tentação de otimizar.

Subestimamos a distância entre nós e os outros. Não apenas a distância inferencial, mas distâncias

de temperamento e habilidade, distâncias de situação e recursos, distâncias de conhecimento tácito e habilidades e sorte não percebidas, distâncias de paisagem interior.

Até mesmo eu fico surpreso ao descobrir que X, que funcionou tão bem para mim, não funciona para outra pessoa. Mas, com tantos outros tentando me otimizar, posso, pelo menos, reconhecer a distância quando sou atingido na cabeça por ela.

Talvez ser pressionado funcione... para você. Talvez você não se sinta mal quando alguém com poder sobre você começa a tentar reorganizar sua vida da maneira correta. Eu não sei o que te motiva. No domínio da força de vontade, da acrasia e da produtividade, assim como em outros domínios, não conheço generalizações suficientemente profundas para serem mantidas quase sempre. Não possuo as chaves profundas que me diriam quando, por que e para quem uma técnica funciona ou não. Tudo o que posso fazer é estar disposto a aceitar quando alguém me diz que não funciona... e continuar procurando as generalizações mais profundas que serão válidas em todos os lugares, as leis mais profundas que governam tanto a regra quanto a exceção, esperando serem encontradas algum dia.

331 — Conselhos práticos apoiados por teorias profundas

Numa ocasião, Seth Roberts decidiu passar férias na Europa e percebeu que começou a perder peso ao consumir sucos de frutas calóricos com sabores desconhecidos.

Agora, vamos supor que Roberts não tivesse conhecimento algum sobre pontos de ajuste metabólicos ou sobre as associações entre sabor e calorias – toda essa pesquisa científica experimental de alto nível realizada em ratos e, ocasionalmente, em seres humanos.

Ele teria escrito em seu blog: "Uau, pessoal! Vocês deveriam experimentar esses sucos de frutas incríveis que estão me fazendo perder peso!" E isso teria sido o fim da história. Algumas pessoas teriam experimentado, talvez funcionasse temporariamente para algumas delas (até que a associação sabor-caloria entrasse em ação), e nunca teríamos a Dieta Shangri-La como a conhecemos.

A Dieta Shangri-La existente claramente apresenta lacunas – para algumas pessoas, como eu, parece não funcionar, sem razão aparente ou lógica discernível. No entanto, a razão pela qual tantas pessoas se beneficiaram – o motivo pelo qual não foi apenas mais uma postagem no blog descrevendo um truque que funcionou para uma pessoa e não para outras – é que Roberts compreendia o método experimental, a ciência que permitia interpretar o que estava observando em termos de fatores fundamentais que realmente existiam.

Um dos conselhos frequentemente citados no *Overcoming Bias* e no *Less Wrong* é a ideia da "conclusão final" – assim que uma conclusão está formada em sua mente, ela já é verdadeira ou falsa, sábia ou estúpida, e nenhum argumento posterior pode mudar isso, exceto alterando a conclusão. Isso está diretamente relacionado a outra ideia crucial e frequentemente mencionada, que é a noção de "motores de cognição", mentes como motores de mapeamento que necessitam de evidências como combustível.

Imagine se eu tivesse apenas escrito mais uma postagem no blog dizendo: "Bem, você deveria realmente estar mais aberto a mudar de ideia – é muito importante – e, ah, sim, também prestar atenção às evidências." Isso não teria sido tão útil. Não apenas porque era menos persuasivo, mas porque as operações reais teriam sido muito menos claras sem a teoria explícita que o respaldava. O que constitui evidência, por exemplo? É algo que parece um argumento forte? Ter uma teoria de probabilidade explícita e uma explicação causal explícita do que torna o raciocínio eficaz faz toda a diferença na força e nos detalhes de implementação do antigo conselho de "manter a mente aberta e prestar atenção às evidências".

Também é crucial perceber que as teorias causais têm maior probabilidade de serem verdadeiras quando são derivadas de um livro de ciências do que quando são inventadas no momento – é muito fácil criar estruturas cognitivas que parecem teorias causais, mas que não têm controle nem das expectativas, muito menos são verdadeiras.

Este é o estilo característico que busco transmitir em todos os ensaios que entrelaçam experimentos de ciências cognitivas, teoria de probabilidade e epistemologia com conselhos práticos. Estes conselhos práticos tornam-se verdadeiramente mais poderosos quando você se aprofunda na leitura sobre experimentos de ciências cognitivas, teoria de probabilidade ou mesmo epistemologia materialista, e realmente compreende o que está observando. Esta é a marca que pode diferenciar o *Less Wrong* de milhares de outros blogs que pretendem oferecer conselhos.

Poderia afirmar: "A satisfação com sua refeição depende mais da qualidade do que da quantidade

que você consome". No entanto, você provavelmente esqueceria isso, e o impulso de terminar um prato inteiro continuaria forte. Mas se eu discorrer sobre a insensibilidade ao escopo, a negligência da duração e a regra Pico/Fim, você de repente perceberá, ao olhar para seu prato, que formará quase exatamente a mesma memória retrospectiva, independentemente do tamanho da porção, proporcionando uma teoria profunda sobre as regras que governam sua memória. Agora você sabe que é isso que as regras dizem. (E você também saberá deixar a sobremesa para o final.)

Busco entender como superar a acrasia - como ter mais força de vontade ou fazer mais com menos dor mental. Contudo, existem milhares de pessoas que pretendem oferecer conselhos sobre esse assunto, muitas vezes no nível de alguém como Seth Roberts, que apenas fala sobre os surpreendentes efeitos de beber suco de frutas. Ou, pior ainda, são aqueles que tentam descrever alavancas mentais internas que acionaram, para as quais não há palavras padronizadas e nem sabem realmente como apontar. Considere também a ilusão de transparência, distância inferencial e dupla ilusão de transparência. (Observe como "Você superestima o quanto está explicando e seus ouvintes superestimam o quanto estão ouvindo" se torna um conselho mais contundente depois de apoiá-lo com um experimento de ciência cognitiva e um pouco de psicologia evolutiva.)

Acredito que o conselho que necessito vem de alguém que lê extensivamente sobre psicologia experimental relacionada à força de vontade, conflitos mentais, esgotamento do ego, inversões de preferências, descontos hiperbólicos, colapso do eu, pico economia, etc. Essa pessoa, ao superar sua própria acrasia, consegue compreender o que fizeram em termos verdadeiramente gerais, graças a experiências que fornecem um vocabulário de fenômenos cognitivos reais, em oposição a fenômenos inventados. Além disso, alguém que pode explicar a outra pessoa, novamente usando um vocabulário experimental e teórico que permite apontar experiências replicáveis que fundamentam as ideias em resultados concretos ou ideias matematicamente claras.

Observe o aumento da dificuldade em citar:

- Resultados experimentais concretos (basta consultar um artigo, preferencialmente um que relate $p < 0,01$, pois $p < 0,05$ pode não ser replicável);
- Explicações causais verdadeiras (obtidas de maneira mais confiável ao procurar teorias amplamente utilizadas por uma determinada ciência);
- Interpretação válida de matemática (sobre a qual tenho dificuldade em oferecer conselhos úteis, pois grande parte do meu talento matemático é intuição entrando em ação antes que eu tenha a chance de deliberar).

Se você não sabe em quem confiar, ou não confia em si, é aconselhável concentrar-se nos resultados experimentais inicialmente, pensar em termos de teorias causais amplamente utilizadas por uma ciência, e mergulhar nos campos da matemática e epistemologia com extrema cautela.

Os conselhos práticos tornam-se ainda mais poderosos quando são respaldados por resultados experimentais concretos, explicações causais verdadeiras e interpretação válida da matemática.

332 — O pecado da falta de confiança

Existem três grandes pecados que perturbam os racionalistas em particular, e o terceiro deles diz respeito à falta de confiança. Michael Vassar frequentemente me acusa desse desafio, tornando-o único entre toda a população da Terra.

Na realidade, ele está correto em se preocupar, e eu compartilho dessa preocupação. Qualquer racionalista comprometido provavelmente dedicará considerável tempo ponderando sobre isso. Quando os indivíduos têm conhecimento de um viés ou são alertados sobre o mesmo, a hipercorreção não é incomum como resultado experimental. É isso que torna muitas sub tarefas cognitivas tão desafiadoras – você sabe que há um viés, mas não tem certeza de sua extensão e não tem garantia de estar corrigindo o suficiente. Surge a dúvida: será que você precisa corrigir um pouco mais? E então, um pouco mais ainda? Mas será que isso é o bastante? Ou talvez você tenha exagerado? Está agora em uma situação pior do que se não tivesse tentado corrigir?

Você reflete sobre o assunto, sentindo-se cada vez mais perdido, e a própria tarefa de estimativa começa a parecer cada vez mais fútil...

Quando se trata de questões específicas de confiança, seja excesso ou falta – interpretadas agora em um sentido mais amplo, não apenas em intervalos de confiança calibrados – há uma tendência natural a considerar o excesso de confiança como o pecado do orgulho. Curiosamente, há uma lista que nunca nos alertou contra o uso inadequado da humildade ou o abuso da dúvida. Colocar-se em uma posição muito elevada, ultrapassando seu devido lugar, pensar demasiadamente em si, colocar-se à frente e rebaixar seus semelhantes por comparação implícita – as consequências da humilhação e de ser derrubado, talvez publicamente – não são essas coisas repugnantes e assustadoras?

Ser modesto demais pode parecer mais aceitável em comparação; não seria tão humilhante ser questionado publicamente. Na verdade, descobrir que você é melhor do que imaginava pode ser uma surpresa agradável; e rebaixar-se, e aos outros implicitamente acima, tem uma conotação positiva de gentileza. É algo que até mesmo Gandalf faria.

Então, se você aprendeu mil formas pelas quais os humanos cometem erros e leu uma centena de resultados experimentais em que sujeitos anônimos são humilhados por seu excesso de confiança – mesmo que você tenha lido apenas algumas dezenas – e você não tem certeza exata de quão confiante é, então você corre o risco real de se rebaixar desnecessariamente.

Não tenho uma fórmula perfeita para oferecer que possa neutralizar isso. No entanto, tenho um ou dois conselhos.

Qual é o perigo da falta de confiança?

Deixar passar oportunidades. Deixar de fazer coisas que você poderia ter feito, mas não tentou (o suficiente).

Então, aqui vai um primeiro conselho: se houver uma maneira de descobrir o quão habilidoso você é, a coisa a fazer é testar. Uma hipótese permite testes; hipóteses sobre suas próprias habilidades da mesma forma. Em um determinado momento, parecia-me que deveria ser capaz de vencer o Experimento da IA na Caixa; embora isso parecesse um pensamento bastante duvidoso e arrogante, decidi testar. Mais tarde, pensei que poderia vencer mesmo com grandes somas de dinheiro em jogo, e testei isso, mas só venci uma vez

em três tentativas. Portanto, esse era o limite da minha capacidade naquela época, e não havia necessidade de argumentar para cima ou para baixo, pois eu poderia apenas testar.

Uma das principais maneiras pelas quais pessoas inteligentes acabam sendo estúpidas é se acostumarem tanto a vencer que se fixam em lugares onde sabem que podem vencer — o que significa que nunca ampliam suas habilidades, nunca tentam algo difícil.

Diz-se que isso está relacionado a se definir em termos de "inteligência" em vez de "esforço", porque vencer facilmente é visto como um sinal de "inteligência", enquanto falhar em um problema difícil poderia ter sido interpretado como um esforço valioso.

Bem, não tenho certeza se é assim que um racionalista deveria pensar sobre essas coisas: a racionalidade é uma vitória sistematizada e a tentativa de tentar parece ser um caminho para o fracasso. Eu colocaria desta forma: uma hipótese permite testes! Se você não sabe se vencerá um problema difícil, desafie sua racionalidade para descobrir seu nível atual. Normalmente, não fico me parabenizando por ter tentado – parece ser um mau hábito mental para mim –, mas certamente não tentar é ainda pior. Se você cultivou o hábito geral de enfrentar desafios e venceu pelo menos alguns deles, então talvez você possa pensar consigo mesmo: "Mantenho meu hábito de enfrentar desafios e farei o mesmo na próxima vez também." Você também pode pensar consigo mesmo: "Obtive informações valiosas sobre meu nível atual e onde preciso melhorar", desde que conclua adequadamente o pensamento: "Tentarei não obter essas mesmas informações valiosas novamente na próxima vez".

Se você vencer todas as vezes, significa que não está se esforçando o suficiente. Mas você deve se esforçar seriamente para vencer sempre. E se você se consola demais com o fracasso, perde o espírito vencedor e se torna um idiota.

Quando tento imaginar o que um mestre fictício da Conspiração Competitiva diria sobre isso, sai algo como: "Não é certo perder. Mas a dor de perder não é algo tão assustador que você deva fugir do desafio por medo dele. Não é tão assustador que você tenha que evitar cuidadosamente sentir isso ou se recusar a admitir que perdeu e perdeu muito. Perder deveria doer. Se não doesse, você não seria um Competidor. E não há Competidor que nunca conheça a dor de perder. Agora vá lá e vença.

Cultive o hábito de enfrentar desafios – não necessariamente aqueles que podem aniquilá-lo completamente, mas sim aqueles que têm o potencial de humilhá-lo. Recentemente, li sobre um teísta específico que derrotou Christopher Hitchens em um debate (de forma contundente; conforme relatado por ateus). Imediatamente, entrei em contato com a equipe responsável pelo blog e perguntei se poderiam organizar um debate. Parecia ser alguém contra quem eu queria me testar. Além disso, foi afirmado que Christopher Hitchens deveria ter assistido aos debates anteriores do teísta e estar preparado. Decidi não fazer isso, pois acredito que devo ser capaz de lidar com quase qualquer coisa no momento e desejo verificar se essa crença está correta. Estou disposto a arriscar a humilhação pública para descobrir. Observo que isso não é uma desvantagem no sentido clássico - se o debate for realmente agendado (ainda não recebi resposta) e eu não me preparar e falhar, então perco as apostas que fiz em mim mesmo; obtenho informações sobre meus limites. Não estou dando a mim mesmo desculpas para perder.

Claro, esta é apenas uma abordagem quando você enfrenta um desafio para se testar, não porque precisa vencer a todo custo. Nesse caso, facilita tudo para si mesmo. Fazer o contrário seria um excesso de confiança espetacular, mesmo jogando jogo da velha contra uma criança de três anos.

Uma forma mais sutil de falta de confiança é perder o ímpeto de avançar – entre todas as coisas que você percebe que os humanos estão fazendo errado, coisas que você costumava fazer errado, e das quais provavelmente continua cometendo alguns erros. Você fica tímido, questiona-se, mas não responde às perguntas e não avança. Quando você levanta a hipótese de sua própria incapacidade, não a testa.

Talvez não haja um momento decisivo em que você decida deliberada e visivelmente não tentar algum teste específico... você apenas… desacelera...

Não parece mais valer a pena tentar consertar uma coisa quando há uma dúzia de outras coisas que ainda estarão erradas...

Não há esperança suficiente de triunfo para inspirar você a se esforçar...

Quando você pensa em fazer algo novo, uma dúzia de perguntas sobre sua habilidade surgem imediatamente em sua mente, e você não considera que pode responder a essas perguntas testando a si mesmo...

E tendo lido tanta sabedoria sobre as falhas humanas, parece que o curso da sabedoria é sempre duvidar (nunca resolver dúvidas), sempre a humildade da recusa (nunca a humildade da preparação), e geralmente, que é sábio dizer coisas cada vez piores sobre as habilidades humanas, para passar para o cinismo do sentir-se bem e sentir-se mal.

Assim, meu último conselho é outra perspectiva pela qual você pode ver o problema – pela qual pode julgar qualquer hábito potencial de pensamento que possa adotar – e isso é perguntar:

Essa forma de pensar me torna mais forte ou mais fraco? Realmente, de verdade?

Já falei anteriormente sobre o perigo da razoabilidade – o argumento que parece razoável de que deveríamos usar duas caixas no problema de Newcomb, o argumento que parece razoável de que não podemos saber nada devido ao problema da indução, o argumento que parece razoável de que estaremos em melhor situação, em média, se sempre adotarmos a crença da maioria e outros impedimentos semelhantes ao Caminho. “Isso funciona?” é uma pergunta que você pode fazer para obter uma perspectiva alternativa. Outra perspectiva ligeiramente diferente é perguntar: “Essa forma de pensar me torna mais forte ou mais fraco?” Lembrar-se constantemente de duvidar de tudo o torna mais forte ou mais fraco? Nunca resolver ou diminuir essas dúvidas o torna mais forte, ou mais fraco? Passar por uma crise deliberada de fé diante da incerteza torna você mais forte ou mais fraco? Responder a cada objeção com uma humilde confissão de sua falibilidade o torna mais forte ou mais fraco?

Suas tentativas atuais de compensar um possível excesso de confiança estão tornando você mais forte ou mais fraco? Dica: se você está tomando mais precauções, tentando se testar de maneira mais minuciosa, pedindo conselhos aos amigos, trabalhando para alcançar grandes coisas gradativamente ou ainda falhando às vezes, mas com menos frequência do que antes, provavelmente você está ficando mais forte. Se você nunca falha, evita desafios e se sente geralmente desesperado e desanimado, provavelmente está ficando mais fraco.

Aprendi a primeira manifestação dessa regra muito cedo, quando me preparava para um determinado teste de matemática. Percebi que minha pontuação diminuía a cada teste prático, e ao analisar as folhas de respostas que havia elaborado, notei que estava marcando as respostas corretas para depois apagá-las. Então, disse a mim mesmo: ‘Tudo bem, desta vez vou confiar na intuição e agir por instinto’. Minha pontuação disparou acima do que havia sido no início, e no teste real foi ainda mais elevada. Foi assim que compreendi que duvidar de si mesmo nem sempre o torna mais forte, especialmente se isso interferir na sua capacidade de se orientar por boas informações, como suas intuições matemáticas. (Mas ainda precisei do teste para me conscientizar disso!)

A falta de confiança não é um pecado exclusivo dos racionalistas, mas é um perigo particular ao qual a busca pela racionalidade pode conduzir. É um erro que impede a obtenção da experiência adicional necessária para corrigi-lo.

Dado que a falta de confiança parece ser bastante comum entre os aspirantes a racionalistas que conheço, embora seja consideravelmente menos frequente entre os racionalistas que se tornaram modelos famosos, eu a consideraria como o terceiro dos três pecados que afligem os racionalistas.

333 — Vá em frente e crie a arte!

Nos últimos meses, discutí aspectos da racionalidade. Abordei como desvendar problemas que se tornaram confusos, e como fazer a distinção entre raciocínio legítimo e falacioso, e a disposição para fortalecer-se, que o motiva a abordagem em vez da evasão. Abordei a realização do impossível.

Todas essas são técnicas desenvolvidas em meus próprios projetos, razão pela qual há tanto sobre reducionismo cognitivo. No entanto, ao aplicá-las, a experiência pode variar. Aqueles que se perguntam 'Mas para que serve isso?' talvez devessem reler alguns ensaios anteriores. Compreender, por exemplo, a falácia da conjunção e como identificá-la em uma discussão, raramente parece esotérico. Entender por que o ceticismo motivado é prejudicial pode ser a diferença crucial entre uma pessoa inteligente que permanece inteligente e uma que se torna estúpida. As espirais de declínio emocional engolem muitos que estão desprevenidos...

No entanto, acredito que a 'arte da racionalidade' tem mais lacunas do que preenchimentos - vencer a acrasia e coordenar grupos são déficits que sinto intensamente. Minha ênfase tem sido mais na racionalidade epistêmica do que na instrumental, em geral. E, além disso, há treinamento, ensino, verificação e a transformação em uma ciência experimental legítima. Generalizando mais, a construção da Arte poderia envolver desenvolvimento de literatura introdutória aprimorada, criação de slogans eficazes para relações-públicas, estabelecimento de uma causa comum com outras submetas do Iluminismo, além da análise e abordagem do desequilíbrio de gênero...

Mas esses pequenos fragmentos de racionalidade que compartilhei... Espero... talvez...

Suspeito - poderia até chamar de intuição - que existe uma barreira inicial na jornada da racionalidade. No começo, por padrão, você tem tão pouco para construir que nem percebe a existência de uma Arte a ser descoberta. Se começar a sentir que mais é possível, pode se perder instantaneamente. Conforme observado por David Stove, muitos 'grandes pensadores' da filosofia, como Hegel, são objetos de piedade. [1] Isso é o que geralmente ocorre quando alguém se propõe a desenvolver a arte de pensar; respostas falsas são desenvolvidas.

Quando você tenta aprimorar parte da arte humana de pensar... está fazendo algo não muito diferente do que eu fazia na Inteligência Artificial. Será tentado por explicações errôneas da mente, relatos falsos de causalidade, palavras sagradas misteriosas e a ilusão incrível que resolve tudo.

Não é que os métodos específicos, tanto epistêmicos quanto de detecção de falsificações, que utilizo sejam perfeitos para cada problema específico; mas parecem úteis para distinguir entre sistemas de pensamento bons e ruins.

Espero que alguém que se aprofunde na parte da Arte que descrevi aqui não cometa erros de imediato e de forma automática ao começar a se questionar: 'Como as pessoas deveriam pensar para resolver o novo problema X com o qual estou trabalhando?' Eles não se desviarão imediatamente; não criarão coisas aleatórias; poderão ser conduzidos a consultar a literatura em psicologia experimental; não mergulharão automaticamente em uma espiral de desânimo em torno de sua Ideia Brilhante; terão alguma compreensão do que distingue uma explicação falsa de uma explicação real. Serão submetidos a um teste de resistência.

É possivelmente esse tipo de barreira que impede as pessoas de iniciar o desenvolvimento de uma arte da racionalidade, caso ainda não sejam racionais.

Então, em vez disso, elas... saem e inventam a psicanálise freudiana. Ou uma nova religião. Ou algo

assim. Isso é o que geralmente acontece quando as pessoas começam a refletir sobre o ato de pensar.

Espero que a parte da Arte que delineei, por mais incompleta que seja, consiga superar essa barreira inicial – proporcionar às pessoas uma base sobre a qual construir; dar-lhes uma compreensão de que existe uma Arte e de como ela deve ser desenvolvida; e proporcionar-lhes, pelo menos, um teste de resistência antes que se percam instantaneamente.

Esse é o meu sonho - que essa arte aparentemente altamente especializada de lidar com perguntas confusas possa ser parte do que é necessário, desde o início, para concluir o restante.

Deixo essa tarefa para você. Provavelmente, de qualquer maneira. Não faço promessas sobre para onde minha atenção pode se voltar no futuro. Mas você sabe, há outras coisas que preciso fazer. Mesmo que eu desenvolva mais Arte por acaso, pode ser que não tenha tempo para escrever nada a respeito.

Além de tudo o que mencionei sobre respostas equivocadas e armadilhas, há duas questões que gostaria que você considerasse.

A primeira – baseei minha Arte em diversas fontes. Li autores variados, explorei diferentes experimentos e utilizei analogias de diversas áreas. Para desenvolver sua própria contribuição para a Arte, será necessário recorrer a múltiplas fontes. Aconselho contra depender exclusivamente de um único autor, embora possa haver um site central onde compartilhe links e artigos que julgue realmente importantes. Uma Arte em crescimento necessitará de insights de várias fontes. Até onde sei, não há ciência genuína que derive sua força de uma única pessoa. Essa ideia parece ser estritamente uma expressão idiomática dos cultos. Na verdadeira ciência, pode haver heróis, até mesmo heróis solitários e desafiadores, mas não apenas um.

A segunda – desenvolvi minha Arte enquanto tentava realizar algo específico que motivasse todos os meus esforços. Pode ser que eu esteja sendo idealista demais – talvez esteja imaginando demais como o mundo deveria funcionar –, mas ainda assim, acredito que não se pode criar a Arte apenas sentando e pensando consigo mesmo: “Agora, como posso superar essa acrasia?” Desenvolverá o restante da Arte ao tentar realizar algo. Talvez até mesmo – se não estou generalizando demais a partir da minha própria experiência –, enfrentar uma tarefa suficientemente difícil para forçar a revisão do seu entendimento anterior e obrigá-lo a reinventar algumas ideias. Contudo, posso estar equivocado, e a próxima fase do trabalho pode ser conduzida por meio de pesquisa direta e específica sobre “racionalidade”, sem a necessidade de uma aplicação considerada mais importante.

Uma tentativa passada minha de descrever esse princípio, em termos de manter uma identidade secreta ou de realizar um trabalho diário que não ensine racionalidade, foi prontamente rejeitada pelo meu público. Talvez “sair de casa” seja uma descrição mais apropriada? Parece-me uma ideia bastante sensata e saudável. Mesmo assim, posso estar enganado. Veremos de onde surgirão as próximas peças da Arte.

Tenho me esforçado muito para transmitir, compartilhar um pouco daquela experiência singular que vivenciei e que considero tão valiosa. Não tenho certeza se já expressei o ritmo central em palavras. Talvez você consiga captá-lo ao ouvir as notas. Posso proferir estas palavras, mas não a regra que as originou ou a regra subjacente; apenas podemos esperar que, ao utilizar as ideias, uma máquina semelhante possa nascer dentro de você. Lembre-se de que todos os esforços humanos para aprender arcanos geralmente resultam em senhas, hinos e afirmações efêmeras.

Tenho me dedicado há muito tempo a transmitir minha Arte. Principalmente sem êxito, antes deste esforço atual. Anteriormente, minhas tentativas eram passageiras e obtinham, no máximo, o sucesso que mereciam. Era como lançar pedras em um lago, gerando algumas ondulações que logo desapareciam... Desta vez, devolvi um pouco e atirei uma pedra grande. O tempo dirá se foi suficientemente grande – se perturbei alguém profundamente a ponto das ondas do impacto continuarem por conta própria. O tempo dirá se criei algo que se mova autonomamente.

Quero que as pessoas partam, mas também que retornem. Ou talvez até que avancem e permaneçam simultaneamente, afinal, isso é a Internet e podemos escapar impunes desse tipo de coisa. Recentemente, aprendi algumas lições interessantes no *Less Wrong*, e se a motivação constante ao longo dos anos for um problema, conversar com outras pessoas (ou perceber que outros também estão tentando) geralmente ajuda.

Enfim, se de alguma maneira influenciei você, espero que siga em frente, enfrente desafios, vá além do conforto da sua poltrona e crie uma nova Arte. Então, recordando de onde veio, volte para compartilhar o que aprendeu com os outros.

Referências

[1] David Charles Stove, The Plato Cult and Other Philosophical Follies (Cambridge University Press, 1991).