O SÉCULO + IMPORTANTE

HOLDEN KARNOFSKY

O Século Mais Importante

por Holden Karnofsky

Traduzido por Mariana Hungria

Brasil, 2024

Sumário

Resumo

O século mais importante

A série de postagens de blog intitulada **"O século mais importante"** argumenta que **o século XXI pode ser o mais importante de toda a história da humanidade**. Isso se deve ao desenvolvimento de sistemas de IA avançados que podem acelerar dramaticamente o avanço científico e tecnológico, nos impulsionando mais rapidamente do que muitos imaginam para um futuro profundamente desconhecido.

Você pode acessar os **principais destaques da série** por meio de:

- Um resumo de algumas páginas (abaixo)
- Discussões sobre a série em The Ezra Klein Show (NYT, 90 minutos) ou no podcast 80,000 Hours (2 horas)

Você pode **ler toda a série** como:

- Uma **série de postagens de blog**: sugiro começar pelo Roteiro. (Cada artigo contém links para o próximo no final.) Este é o formato original, onde será mais fácil navegar, ver gráficos em tamanho real, etc.
- Uma **série em áudio**, disponível na maioria das plataformas de podcast (Spotify, Stitcher, Apple Podcasts, etc.)
- Um único PDF para impressão.
- Para Kindle, você pode comprar uma versão no formato Kindle por $0,99 (o preço mínimo permitido) ou baixar este arquivo AZW3 gratuitamente (veja instruções para colocar no seu Kindle). Há também um arquivo ePub gratuito para outros leitores.

A série em poucas palavras

Passei a maior parte da minha carreira buscando maneiras de fazer o máximo de bem possível, por unidade de dinheiro ou tempo. Trabalhei para encontrar instituições de caridade apoiadas por evidências, voltadas para a saúde global e o desenvolvimento (co-fundando a GiveWell), e mais tarde me envolvi em filantropia que assume mais riscos (co-fundando a Open Philanthropy).

Nos últimos anos – graças ao diálogo com a comunidade do altruísmo eficaz e às extensas pesquisas realizadas pela equipe de investigações de visões de mundo da Open Philantropy – me convenci de que a humanidade enfrenta grandes riscos e oportunidades neste século. Compreender melhor e nos preparar para esses riscos e oportunidades é no que estou focado agora.

Este artigo resume uma série de postagens sobre por que acredito que estamos vivendo o **século mais importante de toda a história da humanidade**. Ele oferece um resumo curto, postagens-chave e, às vezes, gráficos principais para os seguintes cinco pontos básicos:

- **O futuro de longo prazo será radicalmente diferente.** Avanços tecnológicos suficientes podem levar a uma civilização duradoura, que se espalharia por toda a galáxia e poderia ser uma utopia, distopia ou algo entre os dois.
- **O futuro de longo prazo pode chegar muito mais rápido do que pensamos,** devido a uma possível explosão de produtividade impulsionada pela IA.
- O tipo relevante de **IA parece que será desenvolvido neste século** – o que torna este o século que iniciará, e terá a oportunidade de moldar uma civilização galáctica.
- Essas alegações parecem "audaciosas" demais para serem levadas a sério. Mas há muitas razões para acreditar que **vivemos em tempos audaciosos e devemos estar preparados para qualquer coisa.**
- Nós, as pessoas que vivem neste século, temos a chance de ter um grande impacto em um número gigantesco de pessoas no futuro – se conseguirmos entender a situação o suficiente para encontrarmos ações úteis. Mas no momento, **não estamos preparados para isso.**

Esta tese tem uma sensação excêntrica, de ficção científica. Está muito longe de onde eu esperava chegar quando comecei a tentar fazer o máximo de bem possível.

Mas uma parte da mentalidade que desenvolvi por meio da GiveWell e da Open Philanthropy é estar aberto a possibilidades estranhas, ao mesmo tempo que as examino criticamente com o máximo rigor possível. E depois de muito tempo investido examinando essa tese, acho que é provável o suficiente para o mundo precisar urgentemente prestar mais atenção nela.

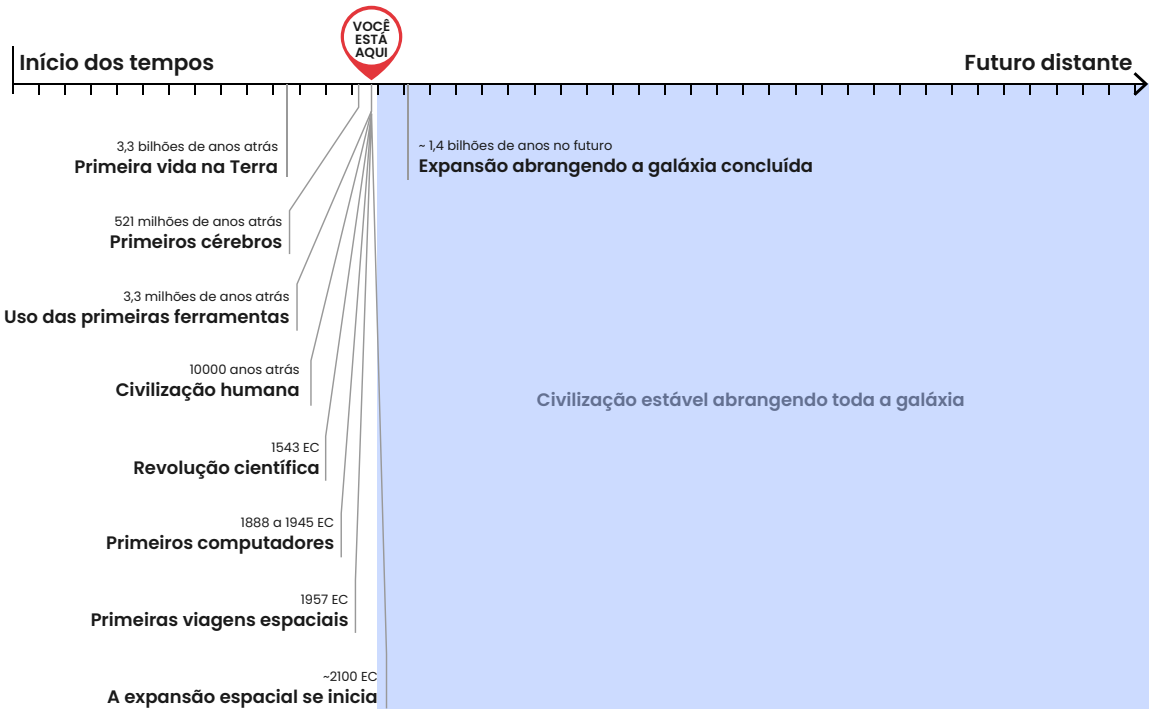
Ao escrever sobre isso, espero chamar mais atenção para essa ideia ou ganhar mais oportunidades de ser criticado e mudar de ideia.

Vivemos em tempos audaciosos e devemos estar preparados para qualquer coisa

Muitas pessoas acham que a afirmação sobre o "século mais importante" é audaciosa demais: um futuro radical com IA avançada e civilização se espalhando pela galáxia pode acontecer *eventualmente*, mas seria mais como daqui a 500, 1.000 ou 10.000 anos. (Não neste século.)

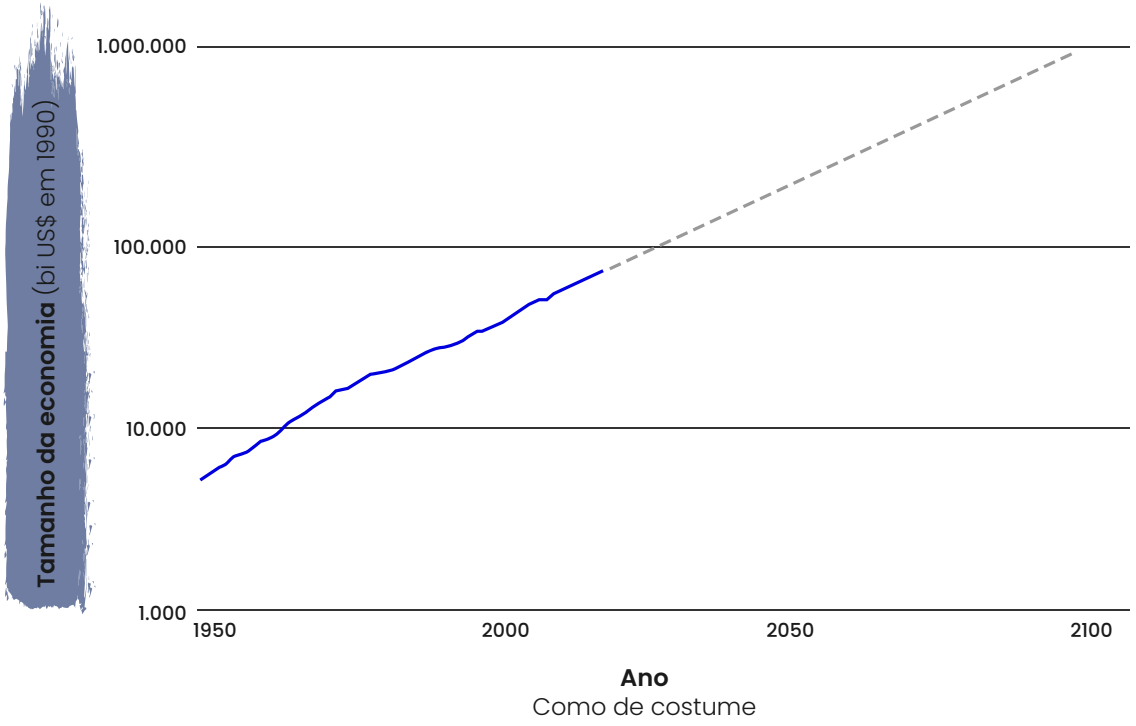
Esses prazos mais longos nos colocariam em uma posição *menos* audaciosa do que se já estivermos vivendo no "século mais importante". Mas, em termos gerais, **mesmo que a expan-**

são pela galáxia comece daqui a 100.000 anos, ainda assim viveríamos em uma era extraordinária – a pequena fração de tempo durante a qual a galáxia passará de quase sem vida para amplamente povoada. Isso significa que, de um número impressionante de pessoas que existirão, estamos entre as primeiras. E que, de centenas de bilhões de estrelas na nossa galáxia, a nossa produzirá os seres que a preencherão.

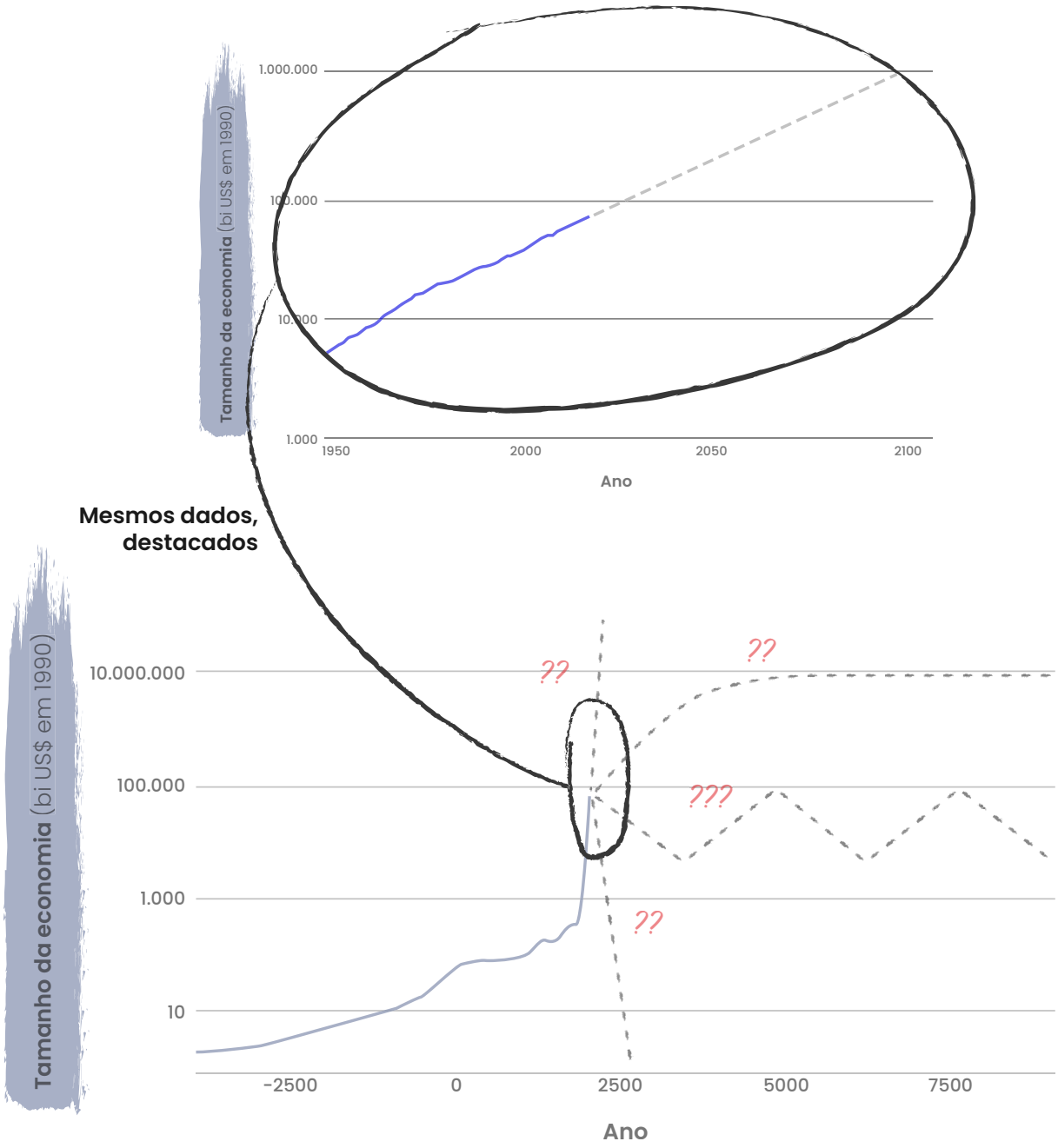


Mais em [Todas as visões possíveis sobre o futuro da humanidade são audaciosas]

Quando nos aproximamos, vemos que vivemos em um século especial, não apenas em uma era especial. Podemos perceber isso ao observar a rapidez com que a economia está crescendo. Não *parece* que algo especial esteja acontecendo, porque, durante toda a vida de qualquer um de nós, a economia mundial cresceu alguns por cento ao ano:



No entanto, quando nos afastamos para olhar para a história num contexto maior, vemos uma imagem de um passado instável e de um futuro incerto:

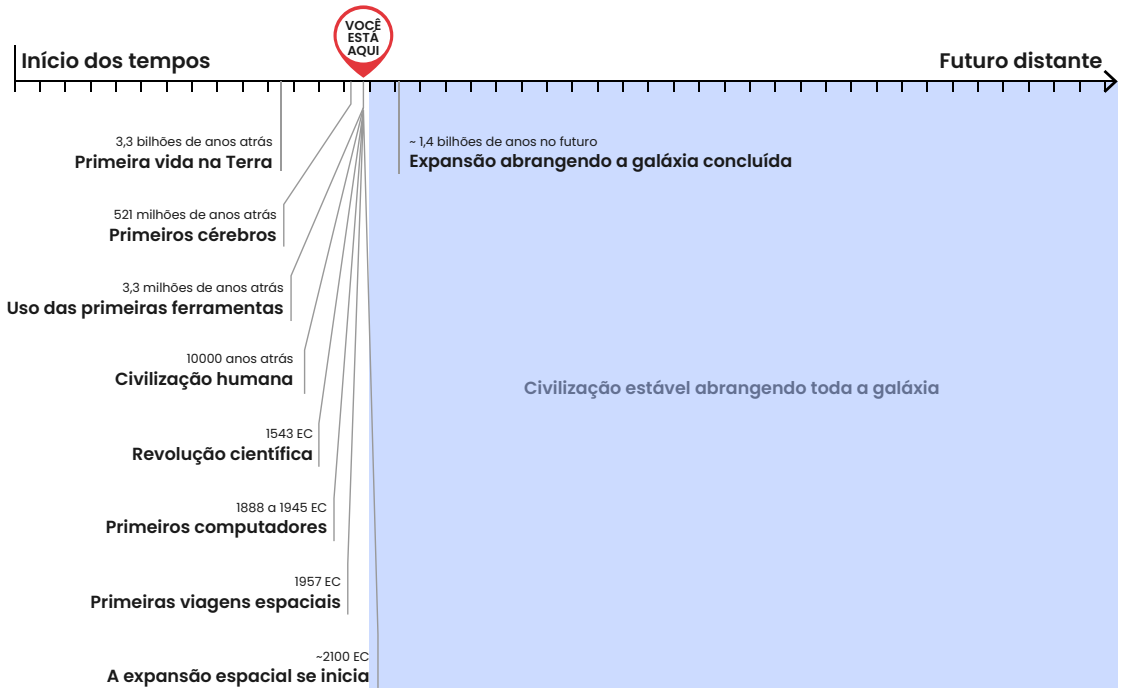


Mais em [Isto não pode continuar]

Estamos vivendo o período de crescimento mais rápido da história. Esta taxa de crescimento não tem sido assim há muito tempo e não pode continuar indefinidamente (não há átomos suficientes na galáxia para sustentar esta taxa de crescimento nem mesmo por mais 10 000 anos). E se conseguirmos uma aceleração maior nesta taxa de crescimento — alinhada com a aceleração histórica — poderemos alcançar os limites daquilo que é possível mais rapidamente: durante este século.

Para recapitular:

- Os últimos milhões de anos — com o início da nossa espécie — têm tido mais acontecimentos do que os vários bilhões anteriores.
- Os últimos cem anos têm tido mais acontecimentos do que os vários milhões anteriores.
- Se tivermos outro acelerador (como eu acho que a IA pode ser), as próximas décadas podem ser as que terão mais acontecimentos de todas.



Mais informações sobre estas cronologias em Todas as visões possíveis sobre o futuro da humanidade são audáciosas, Isto não pdoe continuar, e Prevendo a IA Transformadora: o Método das "Âncoras Biológicas", respectivamente.

Dado o momento em que vivemos, precisamos estar abertos às maneiras pelas quais o mundo pode mudar rápida e radicalmente. Idealmente, deveríamos estar um pouco super atentos a essas possibilidades, assim como priorizamos a segurança ao dirigir. Mas hoje, tais possibilidades recebem pouca atenção.

Artigos principais:

- Todas as visões possíveis sobre o futuro da humanidade são audáciosas
- Isto não pdoe continuar

O futuro de longo prazo é radicalmente desconhecido

A tecnologia tende a aumentar o controle das pessoas sobre o ambiente. Para um exemplo concreto e fácil de visualizar sobre como as coisas poderiam ser, se a tecnologia avançar o suficiente, podemos imaginar uma tecnologia como "pessoas digitais": pessoas totalmente conscientes "feitas de software", que habitam ambientes virtuais de tal forma que podem experimentar qualquer coisa e serem copiadas, rodadas em velocidades diferentes e até mesmo "reinicializadas".

Um mundo de pessoas digitais poderia ser radicalmente distópico (ambientes virtuais usados para consolidar o poder absoluto de algumas pessoas sobre outras) ou utópico (sem doenças,

pobreza material ou violência não consensual, e com muito mais sabedoria e autocompreensão do que é possível hoje). De qualquer forma, pessoas digitais poderiam permitir que uma civilização se espalhasse pela galáxia e durasse por muito tempo.

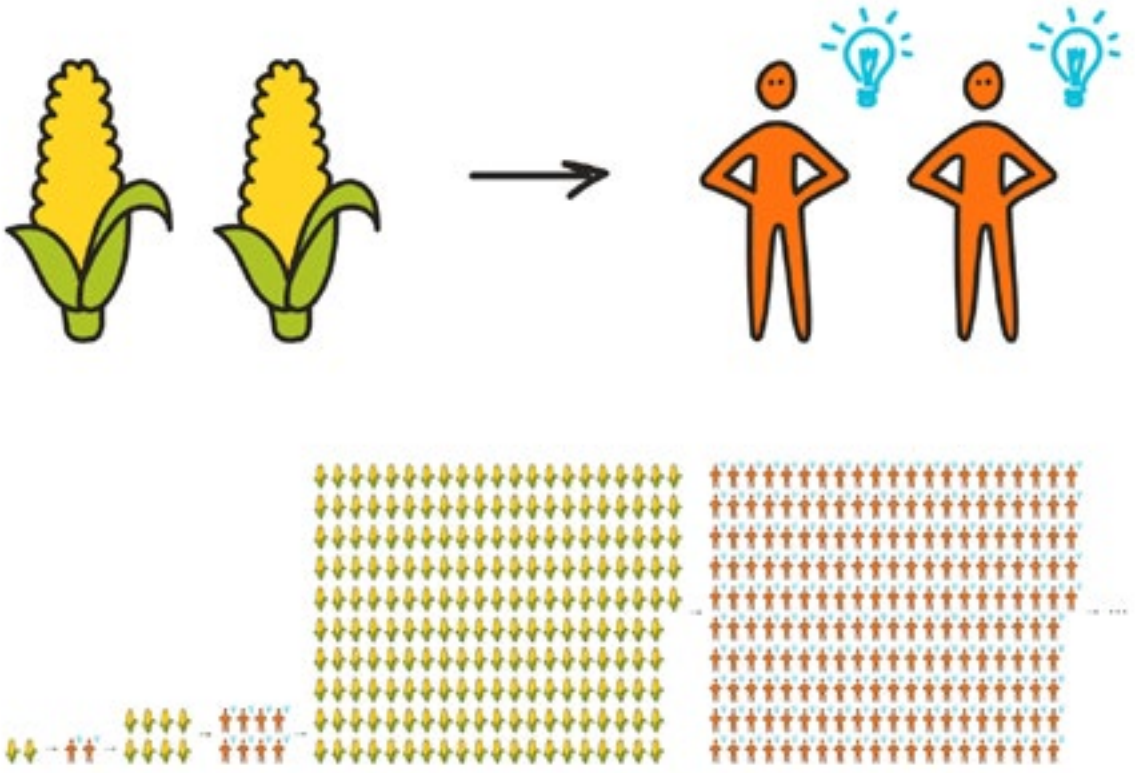
Muitas pessoas acreditam que esse tipo de grande civilização futura estável está no horizonte eventualmente (seja por meio de pessoas digitais ou outras tecnologias que aumentem o controle sobre o ambiente), mas não se preocupam em discutir isso por parecer algo muito distante.

Artigo principal: Pessoas digitais seriam mais importantes ainda

O futuro de longo prazo pode chegar muito mais rápido do que pensamos

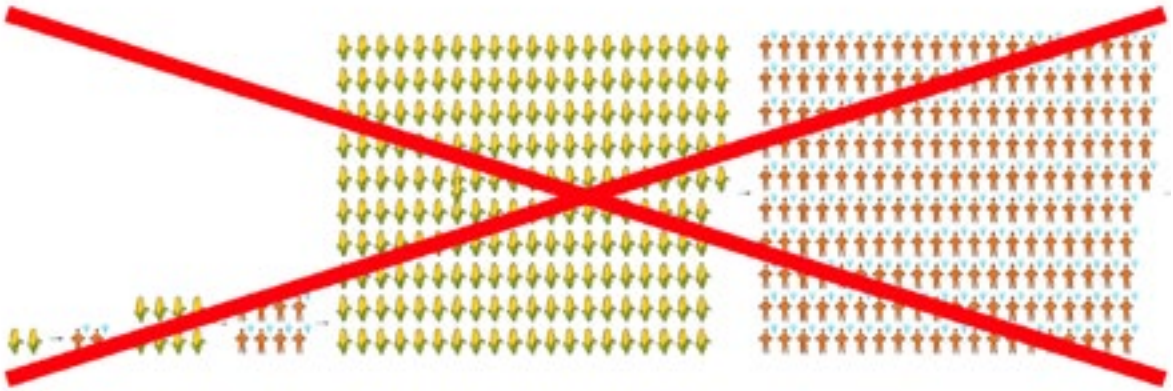
Modelos de crescimento econômico padrão implicam que **qualquer tecnologia que automatize totalmente a inovação causaria uma "singularidade econômica"**: a produtividade iria para o infinito neste século. Isso ocorreria porque criaria um poderoso ciclo de feedback: mais recursos → mais ideias e inovações → mais recursos → mais ideias e inovações...

Esse ciclo não seria sem precedentes. Eu acho que, de certa forma, esse é o modo "padrão" de funcionamento da economia – ao longo de grande parte da história econômica até alguns séculos atrás.



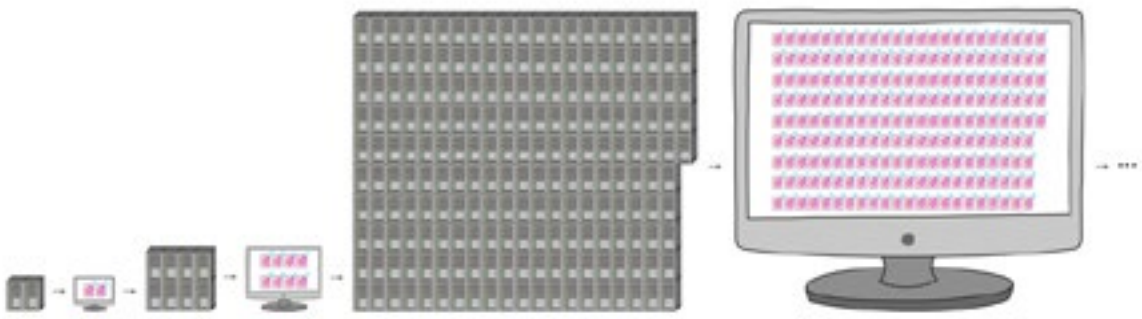
História econômica: mais produção → mais pessoas → mais ideias → mais produção...

Mas na "transição demográfica" de alguns séculos atrás, o passo "mais produção → mais pessoas" desse ciclo foi interrompido. O crescimento populacional se estabilizou, e mais produção levou a pessoas que se tornaram mais ricas, em vez de, simplesmente, mais pessoas:



Economia atual: mais produção → pessoas mais ricas → mesmo ritmo de ideias...

O ciclo de retroalimentação poderia voltar, se alguma outra tecnologia restaurasse a dinâmica de “mais produção → mais ideias”. Uma dessas tecnologias poderia ser o tipo certo de IA: o que chamo de PASTA, ou Processo para Automatizar o Avanço Científico e Tecnológico.



Futuro possível: mais produção → mais IAs → mais ideias → mais produção...

Isso significa que **nosso futuro radical de longo prazo poderia chegar muito rapidamente** depois que o PASTA for desenvolvido (se for desenvolvido).

Também significa que, se os sistemas PASTA forem *desalinhados* – perseguindo objetivos incompatíveis com os humanos – as coisas podem rapidamente sair de controle.

Artigos principais:

- O duplicador: a clonagem instantânea faria a economia explodir
- Prevendo a IA transformadora, parte 1: Qual tipo de IA?
- Porque o alinhamento da IA pode difícil com o Deep Learning moderno

O PASTA parece que será desenvolvido neste século

Não é controverso dizer que um sistema de IA altamente geral, como o PASTA, seria significativo. A questão é: quando (se algum dia) tal coisa existirá?

Nos últimos anos, uma equipe da *Open Philanthropy* investigou essa questão sob várias perspectivas.

Um dos métodos de previsão observa que:

- Nenhum modelo de IA até hoje foi nem mesmo 1% tão "grande" (em termos de cálculos realizados) quanto um cérebro humano, e até recentemente isso não seria viável financeiramente – mas isso mudará relativamente logo.
- Até o final deste século, será financeiramente possível treinar modelos de IA enormes muitas vezes; treinar modelos do tamanho do cérebro humano em tarefas extremamente difíceis e caras; e até talvez realizar tantos cálculos quanto foram feitos "pela evolução" (por todos os cérebros animais na história até hoje).

As previsões desse método estão alinhadas com a pesquisa mais recente dos pesquisadores de IA: algo como PASTA é mais provável do que não é, de existir neste século.

Vários outros ângulos também foram examinados.

Um desafio para essas previsões: **não há um "campo de previsão de IA"** e nenhum consenso de especialistas comparável ao que existe em torno das mudanças climáticas.

É difícil ter confiança quando as discussões em torno desses tópicos são pequenas e limitadas. Mas acho que devemos levar a sério a hipótese do «século mais importante» com base no que sabemos agora, até que e a menos que, um «campo de previsão de IA» se desenvolva.

Artigos principais:

- Cronologia da IA: onde os argumentos e os especialistas se posicionam (recapitula os outros e discute como devemos raciocinar sobre tópicos como esse, onde não está claro quem são os "especialistas")
- Prevendo a IA transformadora: Qual é o ônus da prova?
- Estamos "Tendendo em direção" à IA transformadora?
- Prevendo a IA Transformadora: o método das "Âncoras Biológicas"

Não estamos prontos para isso

Quando falo sobre estar no "século mais importante", não me refiro apenas a eventos significativos que vão ocorrer. Quero dizer que nós, as pessoas que vivem neste século, temos a chance de ter um impacto enorme em um número gigantesco de pessoas no futuro – se conseguirmos entender a situação o suficiente para encontrar ações úteis.

Mas esse é um grande "se". Muitas das ações que podemos realizar podem tanto melhorar quanto piorar as coisas (e é difícil dizer qual delas).

Ao confrontar a hipótese do "século mais importante", minha atitude não corresponde às atitudes familiares de "empolgação e movimento" ou "medo e evasão". Em vez disso, sinto uma **mistura estranha de intensidade, urgência, confusão e hesitação.** Estou olhando para algo maior do que eu jamais esperava enfrentar, me sentindo despreparado e ignorante sobre o que fazer em seguida.

Situação	**Reação adequada (na minha opinião)**
"Isso pode ser uma empresa bilionária!"	"Uhul, vamos nessa!"
"Este pode ser o século mais importante!"	"... Oh ... uau ... eu não sei o que dizer e estou meio que com vontade de vomitar ... preciso sentar e pensar sobre isso."

Com isso em mente, em vez de fazer uma "chamada à ação", faço uma chamada à vigilância:

- Se você for convencido pelos argumentos desta série, não se apresse para "fazer algo" e depois siga em frente.
- Em vez disso, tome as ações robustamente boas que puder hoje e, caso contrário, se coloque em uma posição melhor para tomar ações importantes quando a hora chegar.
- Para aqueles que buscam uma ação rápida que torne mais provável a ação futura, veja esta seção de "Uma chamada à vigilância."

Artigos principais:

- Como aproveitar o melhor do século mais importante
- Uma chamada à vigilância

Uma metáfora para o meu estado mental é que parece que o mundo é um conjunto de pessoas em um avião decolando em alta velocidade pela pista:

E sempre que leio comentários sobre o que está acontecendo no mundo, as pessoas estão discutindo como ajustar o cinto de segurança da forma mais confortável possível, dado que o usar faz parte da vida, ou dizendo que os melhores momentos da vida são sentar com sua família e assistir às linhas brancas passarem voando, ou discutindo de quem é a culpa por haver um rugido de fundo que dificulta ouvir uns aos outros.

Não sei para onde realmente estamos indo, nem o que podemos fazer a respeito. Mas me sinto bastante seguro ao dizer que, como civilização, não estamos prontos para o que está por vir, e precisamos começar, levando isso mais a sério.

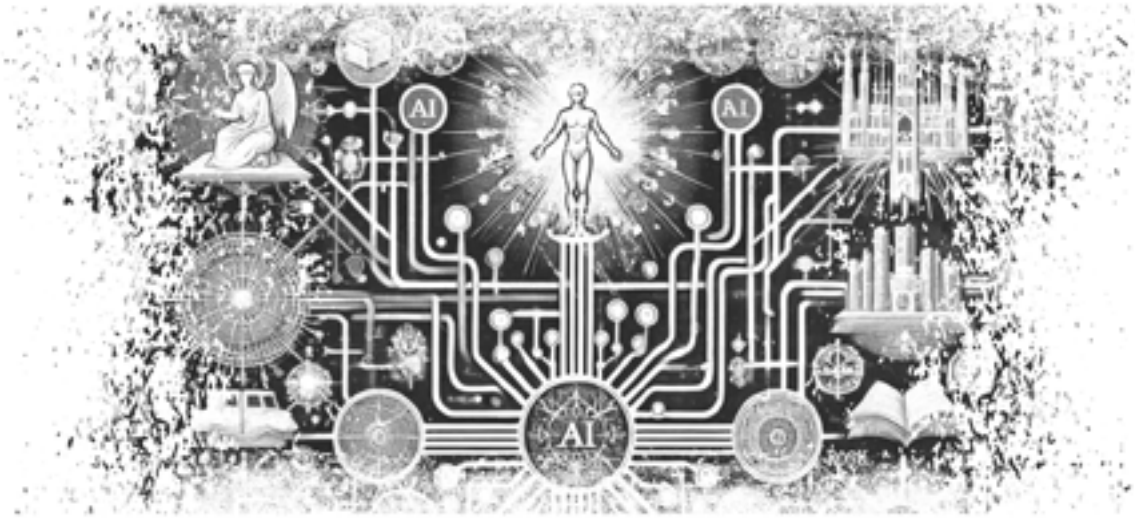
Agradecimentos

Eu não faço praticamente nenhuma reivindicação de originalidade. A grande maioria das afirmações, observações e percepções desta série veio de alguma combinação de:

- Anos de discussões com outras pessoas, particularmente nas comunidades do altruísmo eficaz e dos racionalistas. É difícil rastrear ideias específicas com pessoas específicas nesse contexto, mas sei que uma grande parte do meu pensamento vem, pelo menos proximamente, de Carl Shulman, Dario Amodei e Paul Christiano, e que o trabalho de Nick Bostrom e Eliezer Yudkowsky foi muito influente de modo geral. (Também entendo que futuristas e transhumanistas anteriores influenciaram essas pessoas e comunidades, embora eu não tenha me envolvido diretamente com seus trabalhos.)
- Análises detalhadas realizadas pela equipe de investigações de visões de mundo da Open Philanthropy, composta por Ajeya Cotra e Tom Davidson (especialmente), bem como Nick Beckstead, Joe Carlsmith e David Roodman. Também me baseei fortemente em relatórios de Katja Grace e Luke Muehlhauser.

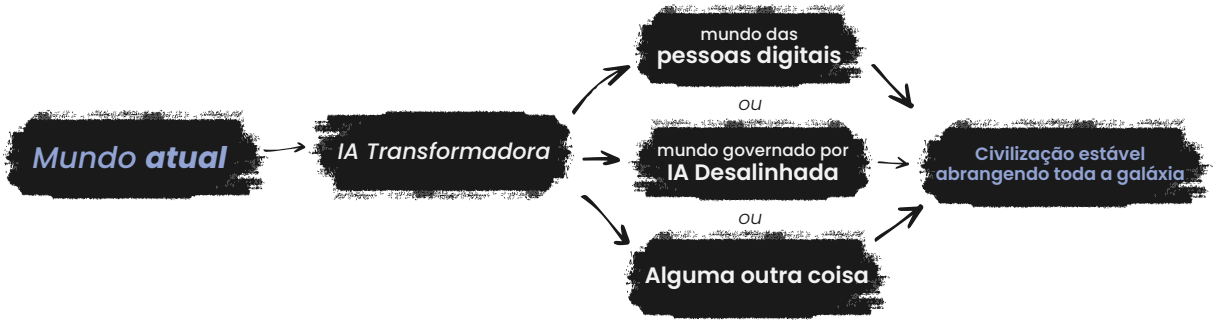
Além disso, agradeço a:

- Ajeya Cotra, María Gutiérrez Rojas e Ludwig Schubert pela ajuda com as visualizações.
- Várias pessoas pelo feedback em rascunhos anteriores:
    - Minha irmã Daliya Karnofsky, minha esposa Daniela Amodei e Elie Hassenfeld: agradecimentos especiais por lerem os primeiros (e menos legíveis) rascunhos e frequentemente dado feedback detalhado em várias iterações.
    - Pessoas que serviram como “leitores beta” e deram grandes quantidades de feedback, sobre particularmente o que estava ou não fazendo sentido para elas: Alexander Berger, Damon Binder, Lukas Gloor, Derek Hopf, Mike Levine, Eli Nathan, Sella Nevo, Julian Sancton, Simon Shifrin, Tracy Williams. (Além de várias outras pessoas já mencionadas acima.)



Roteiro

Este é um resumo de como cada artigo da série "O século mais importante" se relaciona com o argumento geral. Acho útil ler este roteiro antes da série completa para ter uma noção de onde cada artigo se encaixa.



Acho que temos boas razões para acreditar que **o século XXI pode ser o século mais importante para a humanidade.** Acredito que a maneira provável disso acontecer seria por meio do desenvolvimento de sistemas avançados de Inteligência Artificial, que levariam a um crescimento econômico explosivo e ao avanço científico, conduzindo-nos mais rapidamente do que a maioria das pessoas imagina a um futuro profundamente desconhecido.

Um pouco mais especificamente,[1] acredito que há uma boa chance de que:

1. Durante o século atual, desenvolveremos tecnologias que nos farão transitar para um estado no qual os humanos, tais como os conhecemos, não serão mais a força principal nos eventos mundiais. Esta é nossa última chance de moldar como essa transição acontecerá.

2. Independentemente de qual seja a força principal nos eventos mundiais (talvez pessoas digitais, Inteligência Artificial desalinhada ou qualquer outra coisa), ela criará civilizações altamente estáveis que povoarão toda a nossa galáxia por bilhões de anos. A transição que está ocorrendo neste século poderá dar forma a tudo isso.

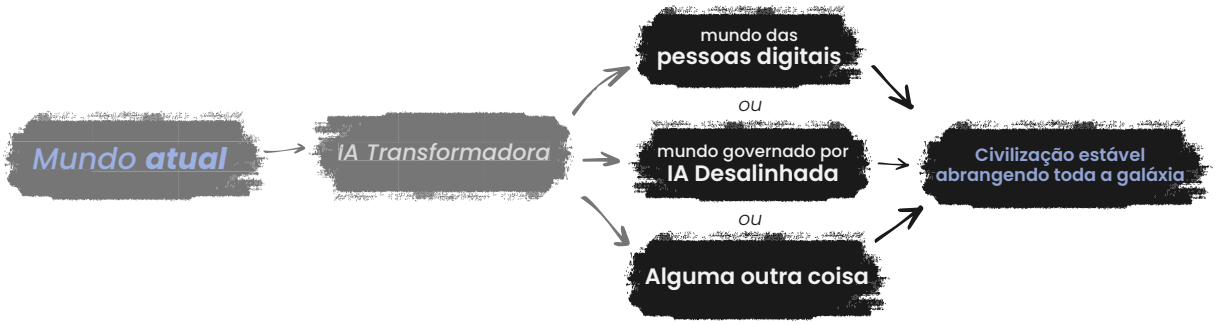
Acho que não está claro se isso seria bom ou ruim. O importante é que a transição poderia seguir vários caminhos diferentes e nós temos a possibilidade de influenciar seu resultado.

Acredito que a possibilidade acima não recebe atenção, discussão ou investimento suficiente, principalmente de pessoas cujo objetivo é melhorar o mundo. Ao escrever sobre o assunto,

gostaria de ajudar a mudar isso ou ter mais oportunidades de ser desafiado a mudar de ideia.

Esta postagem serve como um resumo/roteiro para uma série de 11 postagens discutindo esses argumentos (e as próprias postagens são frequentemente resumos de análises mais longas realizadas por outras pessoas).

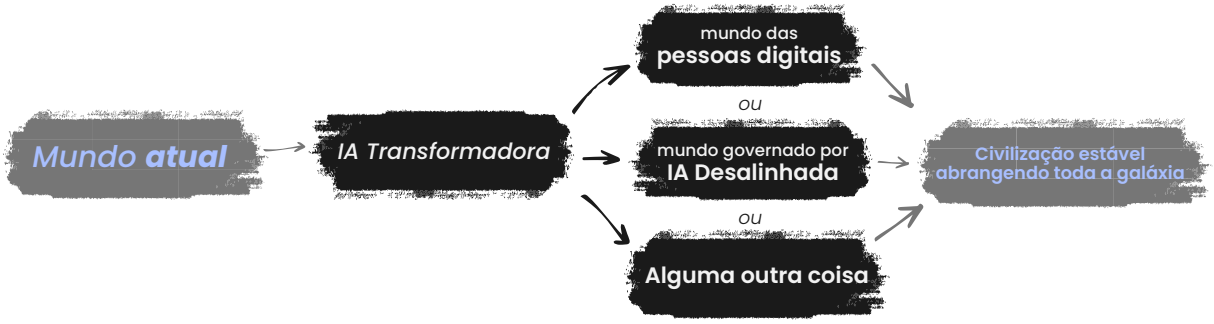
Nossa era extremamente importante



Todas as visões possíveis sobre o futuro da humanidade são audaciosas* discute duas observações simples — (a) é provável que, *eventualmente*, consigamos nos espalhar pela galáxia e (b) não parece que outras formas de vida já o tenham feito. Essas observações são o suficiente para comprovar que vivemos em um momento extremamente relevante. Ilustro isso com uma cronologia da galáxia.

O Duplicador explica o mecanismo básico pelo qual o "eventualmente" descrito acima poderia se tornar "em breve": a capacidade de "copiar mentes humanas" poderia levar a uma explosão de produtividade. Este texto fornece o contexto para os próximos artigos. **Pessoas digitais seriam mais importantes ainda** discute como tecnologias aparentemente alcançáveis — em particular, **upload de** mentes — poderiam levar a uma produtividade sem precedentes, controle do ambiente e muito mais. O resultado poderia ser uma civilização estável abrangendo toda a galáxia, profundamente desconhecida do ponto de vista atual.

O potencial do nosso século para a aceleração



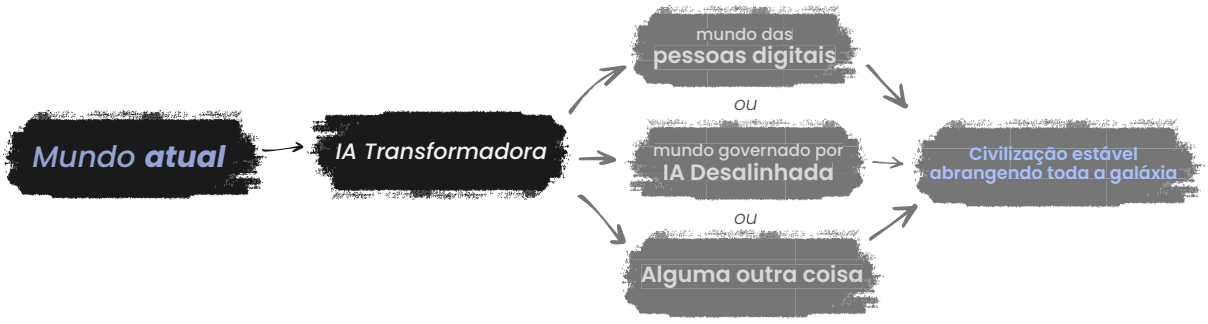
Isto não pode continuar analisa o crescimento econômico e o avanço científico ao longo da história humana. Nas últimas gerações, o crescimento tem sido bastante estável. Mas quando distanciamos a visão para um período mais longo, parece que o crescimento acelerou bastante recentemente; está perto de seu ápice histórico; e não poderá manter essa velocidade por muito mais tempo (não há átomos suficientes na galáxia para sustentar essa taxa de crescimento por mais 10.000 anos).

Os tempos em que vivemos são inusitados e instáveis. Em vez de planejar mais do mesmo, devemos antecipar a estagnação (desaceleração do crescimento e do avanço científico), a explosão (maior aceleração) ou o colapso econômico.

Prevendo a Inteligência Artificial transformadora, parte 1: qual tipo de IA? Introduz a possibilidade de sistemas de Inteligência Artificial que automatizam o avanço científico e tecnológico, o que poderia causar produtividade explosiva. Argumento que tais sistemas seriam "transformadores" no sentido de nos levar a um futuro novo e qualitativamente desconhecido.

Porque o alinhamento da Inteligência Artificial pode ser difícil com o Deep Learning moderno (postagem de convidado) entra em mais detalhes sobre porque os sistemas avançados de Inteligência Artificial podem estar "desalinhados", com consequências potencialmente catastróficas.

Prevendo a Inteligência Artificial transformadora neste século



Prevendo a Inteligência Artificial transformadora: qual é o ônus da prova? Argumenta que não devemos ter um "ônus da prova" muito alto por acreditar que a Inteligência Artificial transformadora poderia ser desenvolvida neste século, em parte porque nosso século já é especial de muitas maneiras que podemos constatar sem uma análise detalhada da IA.

Prevendo a Inteligência Artificial transformadora: estamos "tendendo em direção" à Inteligência Artificial transformadora? Discute a estrutura básica da previsão da Inteligência Artificial transformadora, os problemas com a tentativa de efetuar previsões com base nas tendências da "imponência da IA" e o estado de opinião dos pesquisadores de Inteligência Artificial sobre as cronologias da Inteligência Artificial transformadora.

Prevendo a Inteligência Artificial transformadora: o método das "âncoras biológicas" em poucas palavras resume o **método das âncoras biológicas** para previsão de IA. Este método é o principal fator em minhas previsões específicas.

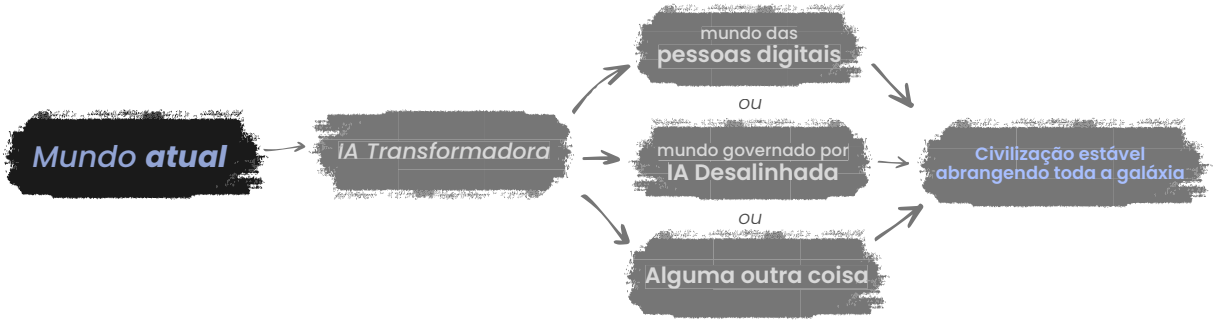
Prevejo mais de 10% de probabilidade da Inteligência Artificial transformadora ser desenvolvida dentro de 15 anos (até 2036); aproximadamente 50% de ser desenvolvida em 40 anos (até 2060); e aproximadamente 2/3 de ser desenvolvida neste século (até 2100).

Cronologias da IA: quais são os argumentos e como os especialistas se posicionam, resume brevemente o estado dos argumentos e aborda a questão: "Como se posicionam os especialistas sobre tudo isso?"

As afirmações que estou fazendo não *contradizem* um determinado consenso de especialistas, nem são *corroboradas* por ele (embora a maioria dos principais relatórios que cito tenham sido revisados por especialistas externos). Elas são, ao contrário, afirmações sobre tópicos que simplesmente não têm "uma área do conhecimento" com especialistas dedicados a estudá-los.

Algumas pessoas podem optar por ignorar quaisquer afirmações que não sejam ativamente corroboradas por um consenso robusto de especialistas; mas não acredito ser isso o que deveríamos fazer aqui.

Implicações



'O século mais importante'

Como aproveitar o melhor do século mais importante? discute visões diferentes e contrastantes de como ajudar o século mais importante a ser da melhor maneira possível para a humanidade — e lista "ações vigorosamente úteis" que, independentemente disso, parecem valer a pena serem tomadas.

Chamada à vigilância está no lugar de uma "chamada à ação" para esta série. Dada toda a incerteza que enfrentamos, não acredito que as pessoas devam se apressar para "fazer alguma coisa" e depois seguir adiante. Em vez disso, elas devem realizar as ações vigorosamente boas que elas possam realizar hoje, e, além disso, colocarem-se em uma posição melhor para realizar ações importantes quando a hora chegar.

Algumas postagens complementares que desenvolvem os argumentos abordados são:

- **Alguns detalhes adicionais sobre o que quero dizer com "o século mais importante"**
- **Uma nota sobre o crescimento econômico histórico**: como o argumento do "século mais importante" é afetado se nossa visão da história econômica de longo prazo mudar.
- **Mais sobre "múltiplas economias de tamanho mundial por átomo"**. Uma continuação de "Isto não pode continuar" para os céticos.
- **Ponto fraco de "O século mais importante": automação completa** (reconhece que eu poderia ter feito mais para abordar a questão de quão completa a automação de Inteligência Artificial deve ser para acarretar as consequências que discuto e acrescenta um pouco mais sobre esse assunto)
- **Ponto fraco de "O século mais importante": *lock-in*** (reconhece que eu poderia ter feito mais para abordar como a Inteligência Artificial poderia levar ao "*lock-in*"(ou aprisionamento em português) do futuro de longo prazo e acrescenta um pouco mais sobre esse assunto)
- **"Âncoras biológicas" é sobre delimitar e, não, especificar, cronologias da IA**: mais sobre como usei o método das "âncoras biológicas"; este artigo é voltado para leitores céticos.

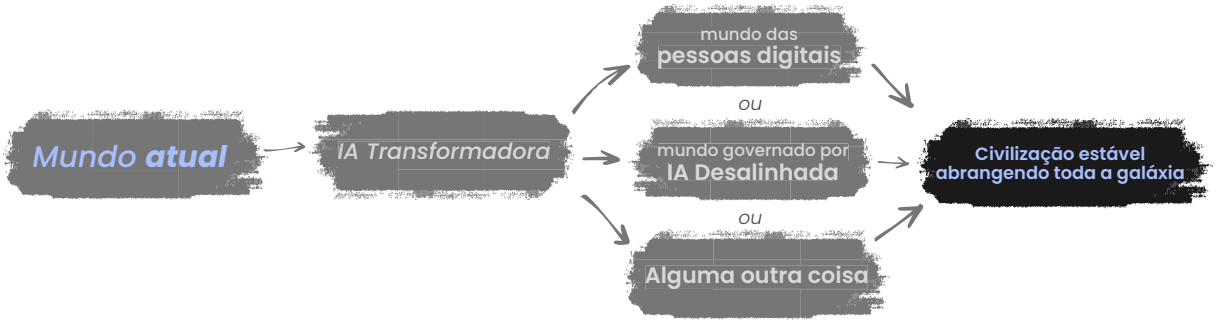
Listei algumas fontes importantes para esta série em um só lugar **aqui**, para os interessados em se aprofundar muito mais no tema.

Notas

[1]Para uma elaboração mais detalhada do que quero dizer com “o século mais importante”, veja aqui (provavelmente não será de interesse para a maioria dos leitores).

*Nota do tradutor: o termo utilizado pelo autor no original em inglês é “*wild*”. Embora uma tradução literal possível seria “selvagem”, esta nos parece inapropriada para o significado que o autor deseja transmitir. Então, sem haver uma palavra diretamente equivalente no português em todas as suas conotações, aqui utilizaremos o termo “audacioso”. Outras possíveis traduções consideradas incluem “alucinante”, “incerto”, “desafiador”, etc.

Todas as visões possíveis sobre o futuro da humanidade são audaciosas



- Em uma série de postagens, começando com esta, argumentarei que no século XXI nossa civilização poderia desenvolver tecnologias que permitirão uma rápida expansão em toda a nossa galáxia, atualmente vazia. E, consequentemente, que **este século poderia determinar todo o futuro da galáxia por dezenas de bilhões de anos, ou mais.**

- Essa visão parece "audaciosa": deveríamos analisar atentamente qualquer visão de que vivemos em um momento tão especial assim. Ilustro isso com uma cronologia da galáxia. (Pessoalmente, essa "audácia" é provavelmente a maior razão pela qual fui cético por muitos anos em relação aos argumentos apresentados nesta série. Tais alegações sobre a importância dos tempos em que vivemos parecem "audaciosas" o suficiente para serem suspeitas.)

- Mas não acredito que seja realmente possível não ter uma visão "audaciosa" sobre este tema. Discuto duas alternativas à minha visão: uma visão "conservadora" que pensa que as tecnologias que estou descrevendo são possíveis, mas que elas levarão muito mais tempo do que eu penso para se desenvolverem, e uma visão "cética" que pensa que a expansão em escala galáctica nunca acontecerá. Cada uma dessas visões parecem "audaciosas" à sua própria maneira.

- No final das contas, conforme sugerido pelo **Paradoxo de Fermi**, nossa espécie parece estar simplesmente em uma situação "audaciosa".

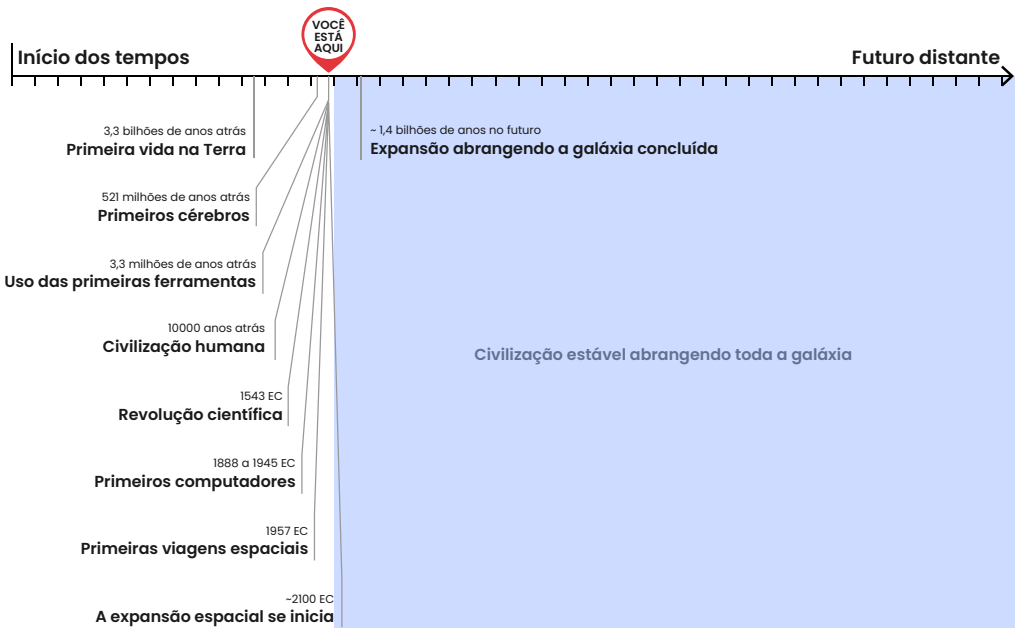
Minha visão

Este é o primeiro de uma série de artigos sobre a hipótese de que vivemos no século mais importante para a humanidade.

Nesta série, argumentarei que há uma boa chance de uma explosão de produtividade ocorrer até 2100, o que acarretaria rapidamente o que poderíamos chamar de uma civilização "tecnologicamente madura"[2] . Isso significaria que:

- Conseguiríamos enviar espaçonaves por toda a galáxia e além.
- Essas espaçonaves conseguiriam extrair materiais, construir robôs,computadores e assentamentos robustos e duradouros em outros planetas, aproveitando a energia das estrelas e sustentando inúmeras pessoas (e/ou nossos "descendentes digitais").
- Veja ***Eternity in Six Hours* (A eternidade em seis horas)** para uma discussão fascinante e curta, embora técnica, sobre o que seria necessário para isto acontecer. Também discutirei em artigos futuros (já disponíveis aqui e aqui), que existe uma probabilidade de "aprisionamento de valor": quem estiver à frente do processo de expansão espacial poderá determinar que tipo de pessoas estarão no comando dos assentamentos e que tipo de valores sociais eles terão, de uma forma estável por muitos bilhões de anos[3].

Se isso acabar acontecendo, podemos pensar sobre a história da nossa galáxia[4] dessa forma. Marquei as principais conquistas da humanidade entre os pontos "ausência de vida" até "vida inteligente que constrói seus próprios computadores e viaja pelo espaço".



Agradecimentos a Ludwig Schubert pelo gráfico. Muitas datas são altamente aproximadas e/ou propensas a críticas e/ou apenas extraídas da Wikipedia (fontes aqui), mas mudanças plausíveis não mudariam o quadro geral. Os aproximadamente 1,4 bilhões de anos para completar a expansão espacial são baseados na distância até a borda externa da Via Láctea, dividida pela velocidade de uma nave espacial rápida feita pelo homem (detalhes na planilha recém-vinculada); Na minha opinião, é provável que seja uma superestimativa de quanto tempo levaria para se expandir por toda a galáxia. Consulte a nota de rodapé para saber porque não usei um eixo logarítmico.[5]

Que loucura!!! Na minha opinião, há uma probabilidade razoável de que estamos vivendo bem no início da pequena fração de tempo durante a qual a galáxia passará de quase sem vida para amplamente povoada. Que, de um número impressionante de pessoas que existirão, estamos entre os primeiros. E que, de centenas de bilhões de estrelas em nossa galáxia, a nossa produzirá os seres que a preencherão.

Sei o que estão pensando: “As chances de que viveríamos em um tempo tão importante parecem infinitesimais; as chances de que *Holden* esteja tendo delírios de grandeza (em nome de toda a Terra, mas ainda assim delírios), parecem muito maiores[6]”.

Mas:

A visão “conservadora”

Digamos que você concorde comigo sobre para onde a humanidade poderia estar se dirigindo — que eventualmente teremos a tecnologia para criar assentamentos robustos e estáveis em toda a nossa galáxia e além. Mas você acha que demorará muito mais do que estou dizendo.

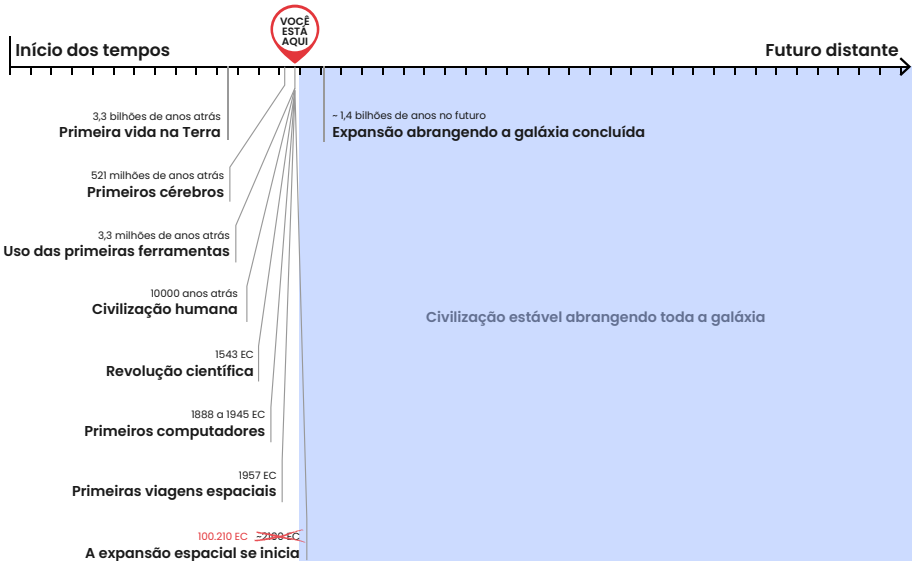
Uma parte fundamental da minha visão (sobre a qual escreverei mais adiante) é que, neste século, poderíamos desenvolver uma Inteligência Artificial avançada o suficiente para iniciar uma explosão de produtividade. Digamos que você não acredite nisso.

- Você acha que estou subestimando os limites fundamentais dos sistemas de IA até o momento.
- Você acha que precisaremos de um número enorme de novos avanços científicos para construir IAs que realmente raciocinem de forma tão eficaz quanto os humanos.
- E mesmo quando o fizermos, a expansão por toda a galáxia será um caminho ainda mais longo.

Você não acha que nada disso acontecerá neste século — você pensa, em vez disso, que isso levará algo como **500 anos para acontecer.** Isso é de 5 a 10 vezes o tempo que já passou desde que começamos a construir computadores. É mais tempo do que se passou desde que Isaac Newton fez a primeira tentativa crível de formular as Leis da Física. É quase o mesmo tempo que se passou desde o início da Revolução Científica.

Na verdade, não; sejamos ainda mais conservadores. Você acha que nosso progresso econômico e científico estagnará. As civilizações de hoje vão desmoronar e muitas outras civilizações vão se findar e se erguer. Claro, *eventualmente* conseguiremos nos expandir por toda a galáxia. Mas isso levará cerca de **100.000 anos.** Isso é 10 vezes a quantidade de tempo que se passou desde que a civilização humana começou no Levante.

Aqui está a sua versão da cronologia:



A diferença entre a sua cronologia e a minha não chega nem a um píxel, então ela não aparece no gráfico. No final das contas, essa visão "conservadora" e a minha são as mesmas.

É verdade que a visão "conservadora" não tem a mesma urgência para a nossa geração em particular. Mas ela ainda nos coloca entre uma pequena proporção de pessoas em um período incrivelmente importante.

E ainda levanta questões sobre se as coisas que fazemos para melhorar o mundo — mesmo que estas ações tenham apenas um impacto minúsculo no mundo daqui a 100.000 anos — elas poderiam ser amplificadas em um grau galáctico-histórico-excepcional.

A visão cética

A "visão cética" seria, essencialmente, a de que a humanidade (ou algum descendente da humanidade, inclusive digital) nunca se espalhará pela galáxia. Há muitas razões pelas quais isso pode não acontecer:

- Talvez algo como viagens espaciais —e/ou a criação de robôs de mineração, painéis solares, etc. em outros planetas — seja efetivamente impossível de realizar, de modo que mesmo outros 100.000 anos de civilização humana não chegarão a esse ponto.[7]
- Ou talvez, por algum motivo, mesmo se for tecnologicamente viável, isso não acontecerá (porque ninguém queira fazer, porque quem não quer, bloqueia quem quer, etc.)
- Talvez seja possível nos expandirmos por toda a galáxia, mas não seja possível mantermos uma presença em muitos planetas por bilhões de anos, por alguma razão.
- Quem sabe a humanidade esteja destinada a se destruir antes de chegar a esse estágio.
    - Porém observe que, se a maneira como nos destruiremos for por meio da Inteligência Artificial desalinhada,[8] seria possível para a mesma construir sua própria tecnologia e se espalhar por toda a galáxia, o que parece estar alinhado com o pensamento das seções acima. Na verdade, isso ressalta que como lidaremos com a Inteligência Artificial neste século poderá ter ramificações por muitos bilhões de anos. Portanto, a humanidade teria que ser extinta de alguma forma que não deixasse nenhuma outra vida inteligente (ou máquinas inteligentes) para trás.
- Talvez uma espécie extraterrestre se espalhe pela galáxia antes de nós (ou mais ou menos na mesma época).
    - No entanto, observe que isso não parece ter acontecido nos aproximadamente 13,77 bilhões de anos desde o início do universo e, segundo as seções acima, faltam apenas cerca de 1,5 bilhão de anos para nos espalharmos pela galáxia.
- Talvez alguma espécie extraterrestre *já tenha* efetivamente se espalhado por nossa galáxia e, por algum motivo, não a vejamos. Talvez eles estejam escondendo sua presença deliberadamente, enquanto estão prontos para impedir que nos espalhemos muito longe.
    - Isso implicaria que eles decidiram não extrair energia de nenhuma das estrelas que podemos ver, pelo menos não de uma maneira que podemos perceber. No que lhe concerne, isso também implicaria que eles estão se abstendo de minerar energia, que poderiam usar para o que quisessem,[9] inclusive para se defenderem de espécies como a nossa.
- Talvez tudo isso seja um sonho. Ou, uma **simulação**.
- Talvez seja outra coisa em que eu não esteja pensando.

É um número razoável de possibilidades, embora muitas delas pareçam bastante "audaciosas" à sua própria maneira. Coletivamente, eu diria que elas somam mais de 50% de probabilidade, mas eu me sentiria muito estranho afirmando que elas são esmagadoramente prováveis no seu conjunto.

"Vamos eventualmente criar assentamentos robustos e estáveis em toda a nossa galáxia e além." No final das contas, aceitar um argumento *contra* pensar que algo assim é pelo menos razoavelmente provável é muito difícil para mim. Parece que dizer "de jeito nenhum" a essa afirmação exigiria uma confiança "audaciosa" sobre os limites da tecnologia e/ou as escolhas de longo prazo que as pessoas farão e/ou a inevitabilidade da extinção humana e/ou algo sobre alienígenas ou simulações.

Imagino que essa afirmação seja intuitiva para muitos leitores, mas não para todos. Defendê-la em profundidade não está nos meus planos no momento, mas repensarei se houver demanda o suficiente para isso.

Porque todas as visões possíveis são audaciosas: o paradoxo de Fermi

Estou afirmando que seria "audacioso" pensar que, basicamente, temos a garantia de nunca nos espalharemos pela galáxia, mas também que é "audacioso" pensar que temos uma probabilidade razoável de nos espalharmos por toda a galáxia.

Em outras palavras, estou chamando todas as crenças possíveis sobre esse tópico de "audaciosas". Isso porque acredito que estamos em uma situação "audaciosa".

Aqui estão algumas situações alternativas que poderiam ter acontecido conosco, que eu não consideraria tão audaciosas assim:

- Poderíamos viver em uma galáxia predominantemente povoada, seja por nossa espécie ou por várias espécies extraterrestres. Estaríamos em alguma região do espaço densamente povoada, cercada por planetas povoados. Talvez ao ler a história de nossa civilização saberíamos (pela História e pela falta de estrelas vazias) que não fomos alguns dos primeiros seres vivos com oportunidades incomuns à nossa frente.
- Poderíamos viver em um mundo onde o tipo de tecnologia que venho discutindo aqui nunca pareceria possível. Não teríamos nenhuma esperança de fazer viagens espaciais, estudar com sucesso nossos próprios cérebros ou construir nossos próprios computadores. Talvez pudéssemos, de alguma forma, detectar vida em outros planetas, mas, se o fizéssemos, veríamos que eles também carecem desse tipo de tecnologia.

Mas a expansão espacial parece viável *e* nossa galáxia está vazia. Essas duas coisas parecem estar em tensão. Uma tensão semelhante — a questão de porque não vemos sinais de extraterrestres, apesar de a galáxia ter tantas estrelas possíveis das quais eles poderiam emergir — é frequentemente discutida sob o título de **Paradoxo de Fermi**.

A Wikipedia tem uma lista de **possíveis soluções** do paradoxo de Fermi. Muitas delas correspondem às possibilidades da visão cética que listei acima. Algumas parecem menos relevantes para este artigo. Por exemplo, existem várias razões pelas quais os extraterrestres podem estar presentes, mas não detectados. Mas acredito que qualquer mundo onde os extraterrestres não impeçam nossa espécie de expandir-se em escala galáctica acaba sendo "audacioso", mesmo que os extraterrestres já estejam lá.

Minha opinião atual é que a melhor análise do Paradoxo de Fermi disponível hoje favorece a explicação de que **a vida inteligente é extremamente rara:** algo sobre o surgimento da vida primeiramente, ou a evolução dos cérebros, é muito improvável que tenha acontecido em mui-

tas (ou quaisquer) outras partes da galáxia.[10]

Isso implicaria que **os passos mais difíceis e improváveis no caminho para a expansão em escala galáctica são os passos que nossa espécie *já deu***. E isso implica que vivemos em uma época estranha: extremamente recente na história de uma estrela extremamente incomum.

Se começássemos a encontrar sinais de vida inteligente em outro lugar da galáxia, eu consideraria isso uma grande atualização da minha visão "audaciosa" atual. Isso implicaria que o que quer que tenha impedido a expansão de outras espécies em toda a galáxia também nos impedirá.

Este ponto pálido azul pode ser incrivelmente importante

Descrevendo a Terra como um pequeno ponto em uma **foto do espaço**, Ann Druyan e Carl Sagan **escreveram**:

Este é um sentimento um tanto comum - quando você recua e pensa em nossas vidas no contexto de bilhões de anos e bilhões de estrelas, você vê quão insignificantes são todas as coisas com as quais nos preocupamos hoje.

Mas aqui estou apresentando o argumento oposto.

Parece que nosso "pequeno ponto" tem uma chance real de ser a origem de uma civilização em escala galáctica. Parece absurdo, até delirante acreditar nessa possibilidade. Mas, dadas as nossas observações, parece igualmente estranho rejeitá-la.

E se isso estiver certo, as escolhas feitas nos próximos 100.000 anos — ou mesmo neste século — poderão determinar se essa civilização em escala galáctica existirá e quais valores ela terá, em bilhões de estrelas e bilhões de anos por vir.

Então, quando olho para a vasta extensão do espaço, não penso: "Ah, no final nada disso importa". Penso: "Bem, *parte* do que fazemos provavelmente não importa. Mas *parte* do que fazemos pode ser mais importante do que qualquer outra coisa… Seria ótimo se pudéssemos prestar atenção nisso."

Notas

[2] ou **Kardashev Type III** (Kardashev Tipo III).

[3] Se conseguirmos criar **uploads de mentes**, ou simulações computacionais detalhadas de pessoas tão conscientes quanto nós, seria possível colocá-las em ambientes virtuais que reinicializariam automaticamente ou se "corrigiriam" sempre que a sociedade desejasse mudá-los de alguma maneira; por exemplo, se uma determinada religião se tornar dominante ou perder o domínio. Isto pode dar aos designers desses "ambientes virtuais" a habilidade de "aprisionar" religiões particulares, governantes específicos, etc. Discutirei mais sobre isso em artigos futuros (já disponíveis aqui e aqui).

[4] Concentrei-me na "galáxia" de forma um tanto arbitrária. Espalhar-se por todo o universo acessível levaria muito mais tempo do que se espalhar por toda a galáxia, e até que o façamos, ainda é imaginável que algumas espécies de fora de nossa galáxia perturbarão a "civilização estável em escala galáctica". Mas acredito que levar isso em consideração adicionaria bastante complexidade sem alterar o quadro geral. Posso abordar isso em algum artigo futuro.

[5]Uma versão logarítmica não aparenta ser menos estranha, porque as distâncias entre os marcos temporais "intermediários" são minúsculas em comparação com os espaços de tempo antes e depois desses marcos. Mais fundamentalmente, estou falando sobre como é notável estar no [pequeno número] de anos mais importantes de [um grande número] de anos —isso é melhor apresentado usando um eixo linear. Muitas vezes, gráficos de aparência estranha parecem mais razoáveis com eixos logarítmicos, mas, neste caso, acredito que o gráfico parece estranho por a situação ser estranha. Provavelmente, a versão menos estranha deste gráfico teria o eixo x como sendo a distância registrada a partir do ano 2100, mas isso seria uma premissa e tanto para um gráfico — basicamente isso seria como assumir como certo o meu argumento de que esse parece ser um período muito especial.

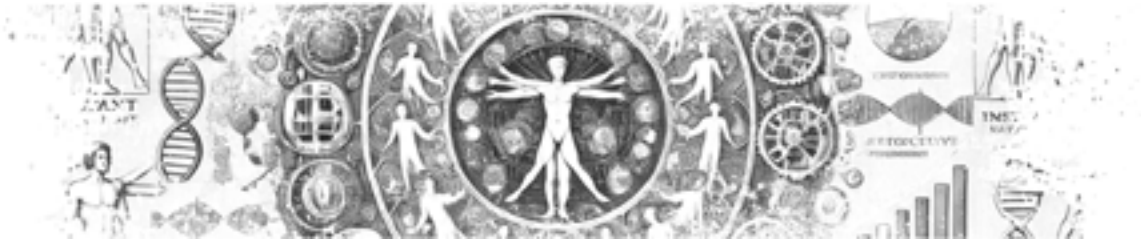
[6]Esse é exatamente o tipo de pensamento que me manteve cético por muitos anos em relação aos argumentos que apresentarei no restante desta série sobre os potenciais impactos e o momento das tecnologias avançadas. Lidar diretamente com o quão "audaciosa" nossa situação parece ser tem sido, inegavelmente, fundamental para mim.

[7]Espalhar-se pela galáxia certamente seria mais difícil se nada como o **upload de mentes** pudesse ser feito; discuto o upload de mentes em um artigo separado. Este é parte do motivo pelo qual acredito que assentamentos espaciais futuros poderiam ter um aprisionamento de valor, como discutido acima. Eu acharia uma visão de que o "upload de mentes é impossível" como sendo "audaciosa" à sua própria maneira, porque ela implica que os cérebros humanos são tão especiais que simplesmente não há como, nunca, replicar digitalmente o que eles estão fazendo. (Obrigado a David Roodman por este ponto.)

[8]Ou seja, uma Inteligência Artificial avançada que persegue objetivos próprios e que não são compatíveis com a existência humana. Escreverei mais sobre essa ideia. As discussões existentes sobre isso incluem os livros ***Superintelligence* (Superinteligência)**, ***Human Compatible*, *life 3.0* (Compatível com o humano, vida 3.0)**, e ***The Alignment Problem* (O problema do alinhamento)**. A apresentação mais curta e acessível que conheço é ***The case for taking AI seriouslyas a threat to humanity* (O caso de levar a sério a Inteligência Artificial como sendo uma ameaça a humanidade)**(Artigo na Vox por Kelsey Piper). Este **relatório sobre o risco existencial de uma Inteligência Artificial que visa poder**, escrito por Joe Carlsmith da *Open Philanthropy*, estabelece um conjunto detalhado de premissas que coletivamente implicariam que o problema é sério.

[9]Obrigado a Carl Shulman por este argumento.

[10]Veja **https://arxiv.org/pdf/1806.02404.pdf**



O duplicador: a clonagem instantânea faria a economia explodir

Esta é a segunda postagem de uma série que explica minha visão de que poderíamos estar no século mais importante de todos os tempos. Aqui está o roteiro para esta série.

- O primeiro artigo desta série discute nossa era incomum, que pode estar muito próxima da transição entre uma civilização terrestre e uma civilização estável em toda a galáxia.
- Artigos futuros discutirão como "pessoas digitais" - e/ou Inteligência Artificial avançada - podem ser a chave para essa transição.
- Este artigo explora uma dinâmica particularmente importante que pode levar pessoas digitais ou Inteligência Artificial avançada a provocar um aumento explosivo na produtividade.

Exploro a questão de como o mundo mudaria se as pessoas pudessem ser "copiadas". Argumento que isso poderia levar a um crescimento econômico e produtividade sem precedentes. Mais tarde, descreverei como pessoas digitais ou Inteligência Artificial avançada poderiam causar uma explosão de crescimento/produtividade.

Quando algumas pessoas imaginam o futuro, elas imaginam o tipo de coisa que você vê nos filmes de ficção científica. Mas esses futuros de ficção científica parecem muito modestos em comparação com o futuro que eu vislumbro.

Na ficção científica, o futuro difere do presente principalmente por meio de:

- Edifícios brilhantes, dispositivos eletrônicos e hologramas.
- Robôs fazendo muitas das coisas que os humanos fazem hoje.
- Medicina avançada.
- Transporte aprimorado, de *hoverboards* a carros voadores, viagens espaciais e teletransporte.

Mas, fundamentalmente, nesses futuros existem os mesmos tipos de pessoas que vemos hoje, com os mesmos tipos de personalidade, objetivos, relacionamentos e preocupações.

O futuro que imagino é enormemente maior, mais rápido, mais estranho e muito, muito melhor ou muito, muito pior se comparado ao dia de hoje. Ele também acontecerá potencialmente muito *mais cedo* do que os futuros da ficção científica:[11] acredito que tecnologias específicas e aparentemente alcançáveis podem nos levar até lá rapidamente.

Essas tecnologias poderiam incluir "pessoas digitais" ou formas específicas de IA avançada cada uma das quais discutirei em um artigo futuro.

Por enquanto, quero me concentrar em apenas um aspecto do que esse tipo de tecnologia permitiria: a capacidade de fazer cópias instantâneas de pessoas (ou de entidades com características semelhantes). A teoria econômica — e a história — sugerem que essa capacidade, por si só, poderia levar a níveis de crescimento econômico e produtividade sem precedentes (na história ou até em filmes de ficção científica).

Isso ocorre por meio de um processo de retroalimentação autorreforçada na qual a inovação leva a mais produtividade, levando a mais "cópias" de pessoas, que, em contrapartida, criam mais inovação e aumentam a produtividade, o que, por sua vez…

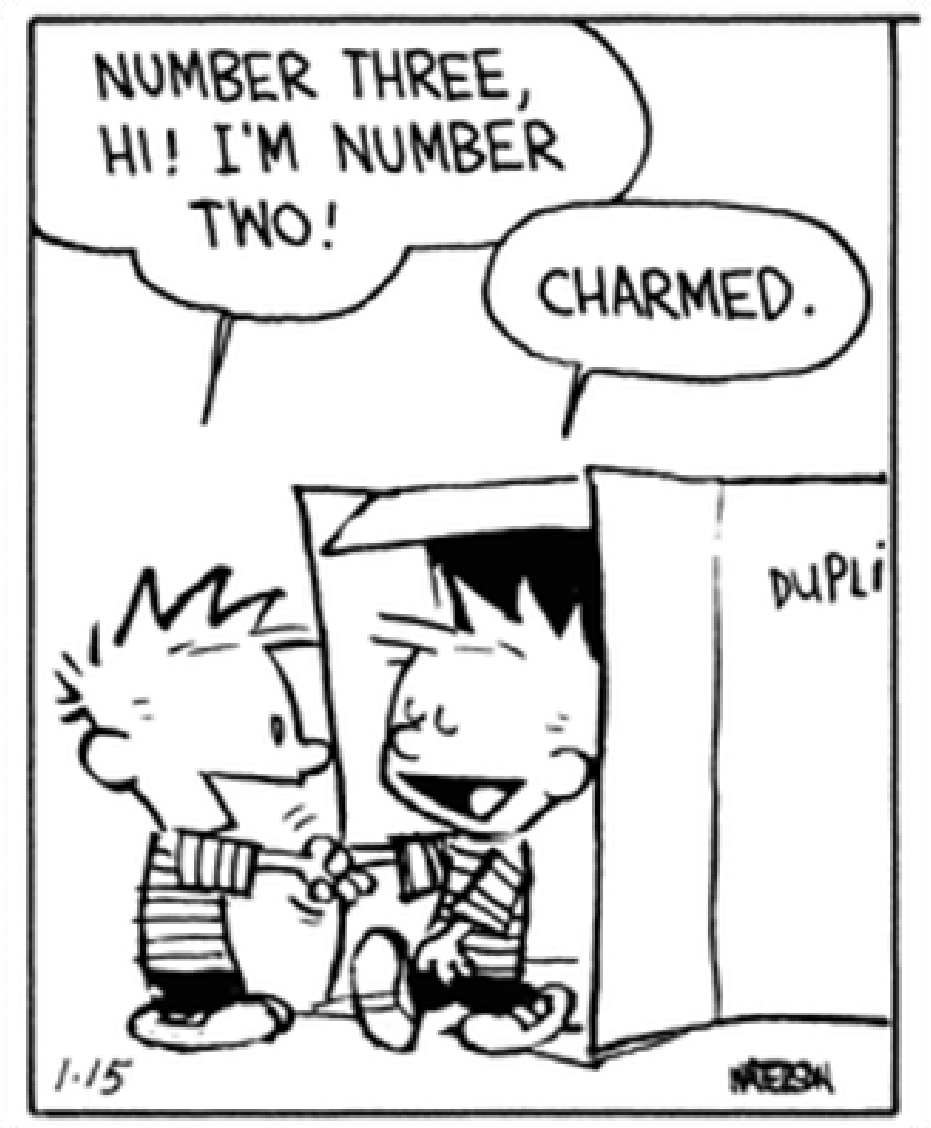
Nesta postagem, em vez de discutir diretamente pessoas digitais ou Inteligência Artificial avançada, manterei as coisas relativamente simples e discutirei uma tecnologia hipotética diferente: o **Duplicador de Calvin e Haroldo**, que simplesmente copia pessoas.

Como o Duplicador funciona

O Duplicador é retratado **nesta série de quadrinhos**. Sua principal característica é fazer uma cópia instantânea de uma pessoa:

Um Calvin entra e dois Calvins idênticos saem.

Isso é muito diferente da versão comum (e mais realista) de "clonagem", na qual o clone de uma pessoa tem o mesmo DNA, mas tem que começar como um bebê e levar anos para tornar-se um adulto.[12]



Para detalhar isso um pouco, presumirei que:

- O Duplicador permite que qualquer pessoa crie rapidamente uma cópia de si mesma. Essa cópia pode ser criada a partir do estado atual e condição mental do original ou de um estado anterior (por exemplo, eu poderia fazer uma réplica do "Holden de 1º de janeiro de 2015").[13] Ao contrário de muitos filmes de ficção científica, as cópias funcionariam normalmente (elas não são más, sem alma, decadentes ou qualquer coisa semelhante).
- O Duplicador pode ser usado para fazer um número ilimitado de cópias, embora cada uma tenha um custo perceptível de produção (elas não são gratuitas).[14]

Impacto na produtividade

Parece que grande parte da economia atual gira em torno de tentar aproveitar ao máximo o "capital humano escasso". Ou seja:

- Algumas pessoas são "escassas" ou "em demanda". Exemplos extremos incluem Barack Obama, Sundar Pichai, Beyoncé Knowles e Jennifer Doudna.[15] Essas pessoas possuem alguma combinação de habilidades, experiência, conhecimento, relacionamentos, reputação, etc., que tornam muito difícil para outras pessoas fazerem o que elas fazem. Exemplos menos extremos seriam quaisquer pessoas que desempenhem um papel crucial em uma organização, sejam difíceis de substituir e, que sejam frequentemente bem pagas.
- Essas pessoas acabam sobrecarregadas, com muito mais demandas de tempo do que podem cumprir. Exércitos de outras pessoas acabam se dedicando a otimizar o tempo delas e a trabalhar de acordo com suas agendas.

O Duplicador removeria esses gargalos. Por exemplo:

- Cópias de Sundar Pichai poderiam operar em todos os níveis da Google, equipadas com sua habilidade de se comunicar facilmente com o CEO e tomar decisões como ele faria. Essas cópias também poderiam abrir novas empresas.
- Cópias do presidente dos Estados Unidos poderiam se encontrar pessoalmente com qualquer eleitor que desejasse entrevistar o presidente, bem como com qualquer congressista ou potencial nomeado, ou conselheiro que o presidente não tivesse tempo de encontrar. Elas poderiam estudar profundamente as principais questões domésticas e internacionais e relatar suas conclusões ao presidente "original".
- Cópias de Beyoncé poderiam produzir quantos álbuns o mercado demandasse. Elas poderiam estudar profundamente e se especializar em diferentes gêneros musicais. Poderiam até tentar viver estilos de vida variados para obter experiências distintas, que serviriam de base para álbuns diversos, mantendo ainda a estética e criatividade pessoal de Beyoncé. Provavelmente, haveria pelo menos uma cópia de Beyoncé cujo trabalho musical seria considerado superior ao da original; essa então poderia se replicar ainda mais.
- Cópias de Jennifer Doudna poderiam investigar qualquer uma das ideias e experimentos que a original não tem tempo de analisar, além de explorar os muitos campos de estudo nos quais ela não teve oportunidade de se especializar. Poderiam existir cópias de Jennifer Doudna atuando na Física, Química e Ciência da Computação, bem como na Biologia, cada uma colaborando com muitas outras cópias de Jennifer Doudna.

A capacidade de fazer cópias para fins *temporários* — e utilizá-las em velocidades diferentes —pode aumentar a eficiência, como discutirei em um artigo futuro sobre pessoas digitais.

Crescimento explosivo

OK, o Duplicador tornaria a economia mais produtiva - mas *quanto* mais produtiva? Para responder, resumirei brevemente o argumento que se pode chamar de **"o crescimento populacional é o gargalo para o crescimento econômico explosivo".** Recomendo ler mais sobre esse ponto de vista nos links a seguir, todos os quais acho fascinantes:

- ***The Year The Singularity Was Cancelled* (O ano em que a singularidade foi cancelada)** (*Slate Star Codex* —razoavelmente acessível se você tiver familiaridade básica com **crescimento econômico**)
- ***Modeling the Human Trajectory* (Modelando a trajetória humana)** (David Roodman da *Open Philanthropy* — postagem de blog razoavelmente acessível, com link para um relatório técnico denso)
- ***Could Advanced AI Drive Explosive Economic Growth?* (Poderia a Inteligência Artificial avançada impulsionar o crescimento econômico explosivo?)** (Tom Davidson da *Open Philanthropy* —postagem de blog acessível, com link para um relatório técnico denso) Segue a continuação meu resumo aproximado.

Nos modelos econômicos padrão, o tamanho total da economia (sua produção total, ou seja, quanto "material" ela cria) é uma função de:

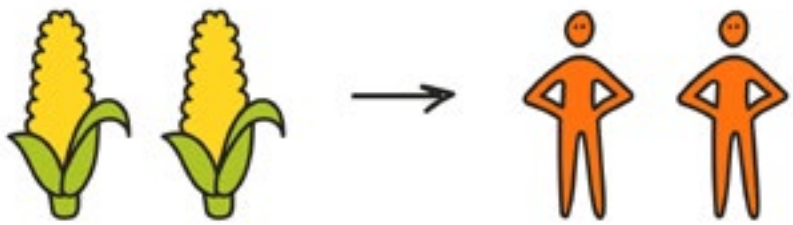
- Quanto "trabalho" total (pessoas trabalhando) existe na economia;
- Quanto "capital" (por exemplo, máquinas e fontes de energia – basicamente tudo exceto trabalho) existe na economia; Quão alta é a produtividade, ou seja, quanto é criado para uma dada quantidade de trabalho e capital. Isso às vezes é chamado de "tecnologia".

Ou seja, a economia cresce quando (a) há mais mão de obra disponível, ou (b) mais capital (~tudo menos trabalho) disponível, ou quando (c) a produtividade ("produção por unidade de trabalho/capital") aumenta.

A população total (número de pessoas) afeta tanto o trabalho quanto a produtividade, porque as pessoas podem ter ideias que aumentam a produtividade.

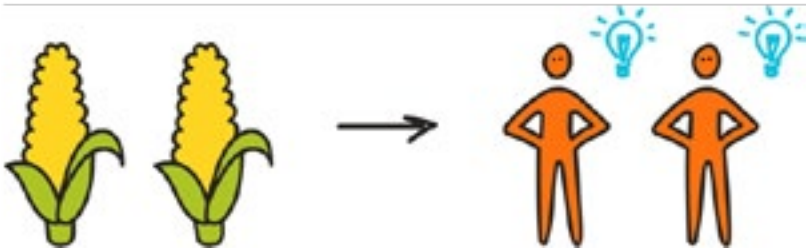
Uma maneira pela qual as coisas poderiam teoricamente acontecer em uma economia seria a seguinte:

A economia começa com algum conjunto de recursos (capital) sustentando algum conjunto de pessoas (população).

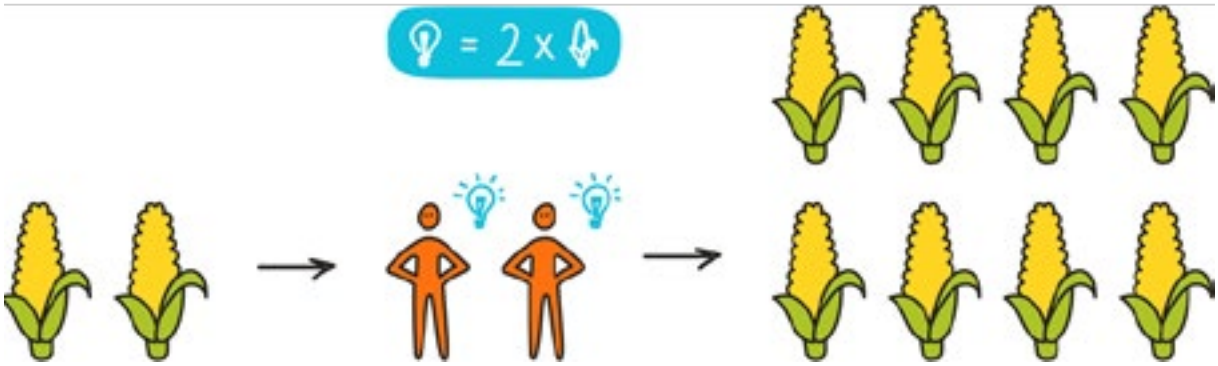


Obrigado a María Gutiérrez Rojas por estes gráficos

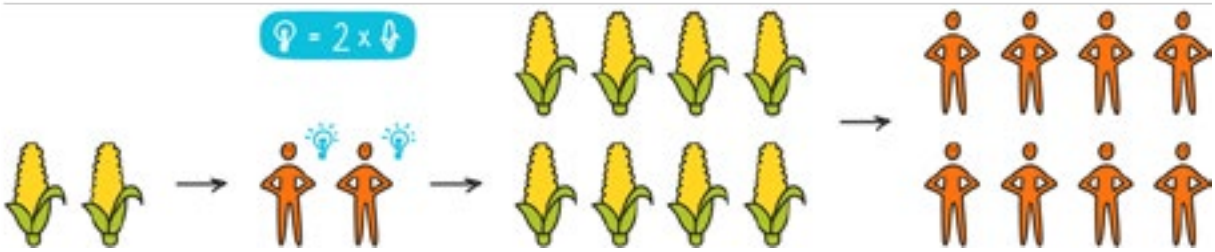
Este conjunto de pessoas têm novas ideias e inovações.



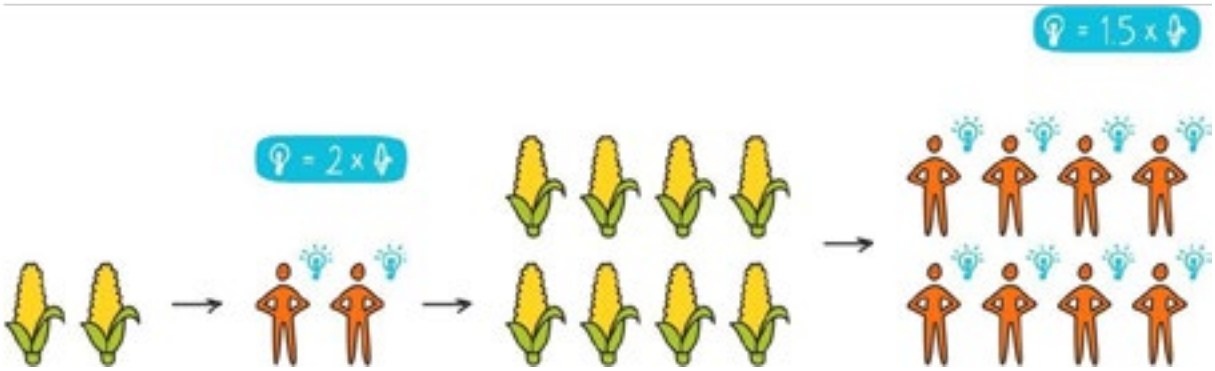
Isto leva a uma certa quantidade de aumento de produtividade, o que significa haver mais produção econômica total[16]



Isso significa que as pessoas podem se dar ao luxo de ter mais filhos. Elas assim o fazem, e a população cresce mais rapidamente.



Por causa desse crescimento populacional, a economia tem novas ideias e inovações *mais rapidamente* do que antes (já que mais pessoas significam mais ideias novas).[17]



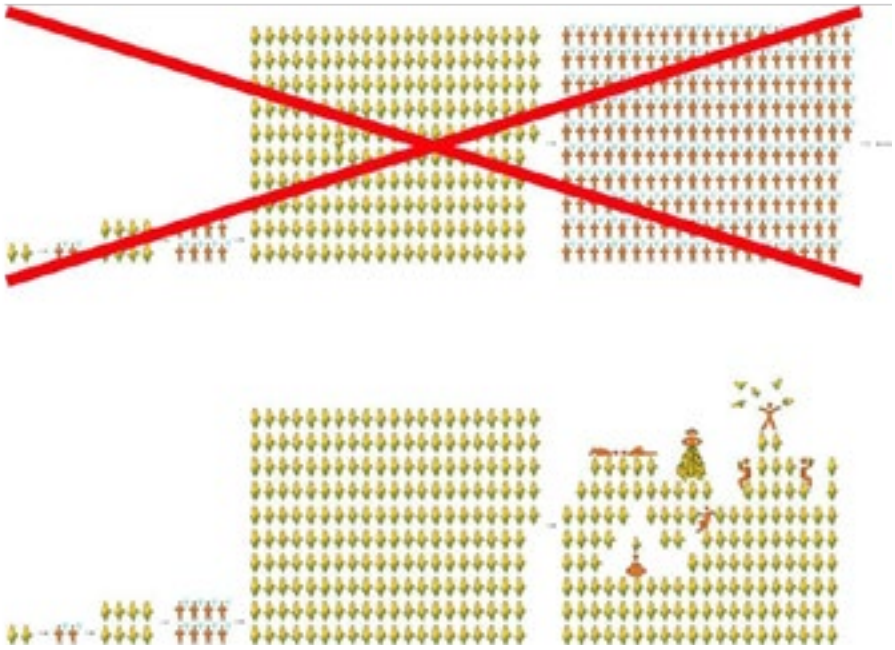
Isso leva a uma produção econômica maior ainda e a um crescimento populacional mais rápido ainda, em um ciclo de autorreforço: *mais ideias → mais produção → mais pessoas → mais ideias → ....*

Quando você incorpora essa retroalimentação completa em modelos de crescimento econômico,[18] eles preveem que (sob suposições plausíveis) a economia mundial terá um **crescimento acelerado**.[19] "Crescimento acelerado" é uma dinâmica bastante "explosiva" na qual a economia pode ir de pequena a extremamente grande com uma velocidade desorientadora.

O padrão de crescimento previsto por esses modelos parece se encaixar razoavelmente bem aos dados da economia mundial nos últimos 5000 anos; veja ***Modeling the Human Trajectory* (Modelando a trajetória humana)**, embora haja um **debate** em aberto sobre este argumento. Discuto como o debate poderia mudar minhas conclusões aqui. **No entanto, nas últimas centenas de anos, o crescimento não acelerou; ele tem sido "constante"** (uma dinâmica menos explosiva) ao redor **do nível de crescimento que temos hoje.**

Por que o crescimento acelerado migrou para o crescimento constante?

Essa mudança coincidiu com a **transição demográfica**. Na transição demográfica **deixou de ser o caso de que ter mais produção - > ter mais filhos. Em vez disso, mais produção significou apenas que as pessoas se tornaram mais ricas, e, na verdade, tiveram menos filhos à medida que sua riqueza aumentava.** Isso quebrou o ciclo de autorreforço descrito acima.

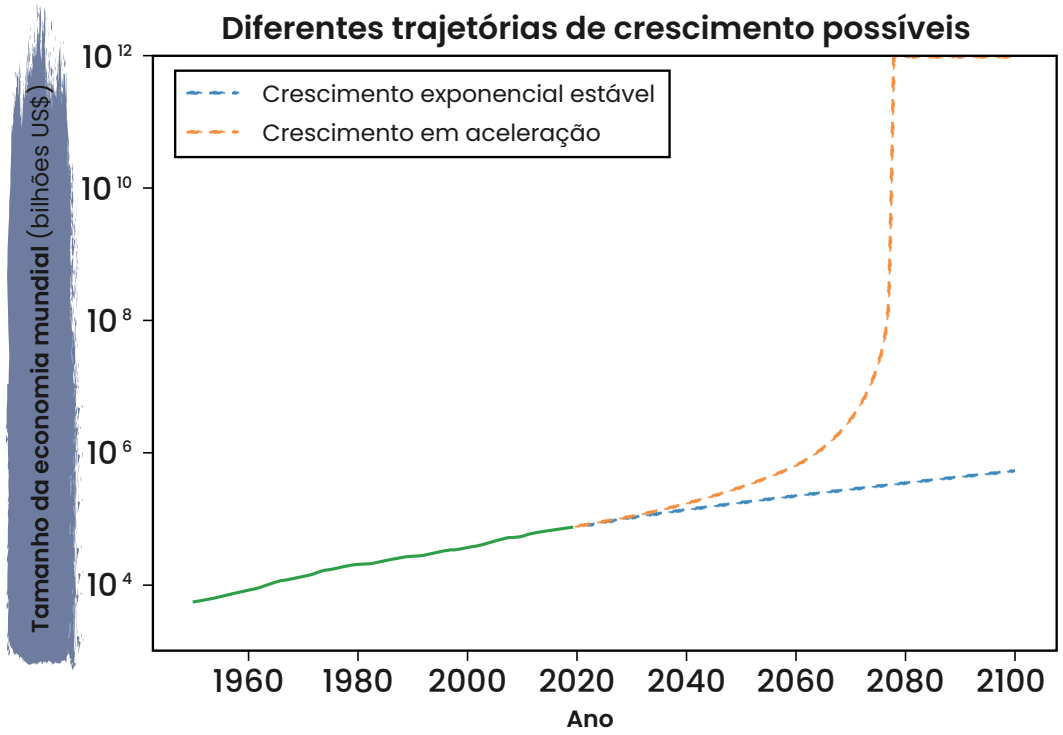


Transição demográfica

Criar filhos é um investimento massivo (de tempo e energia pessoal, não apenas "capital"), e os filhos demoram muito para crescer. Ao mudar o que é necessário para aumentar a população, o Duplicador pode restaurar a retroalimentação acelerada.

Período	Retroalimentação?	Padrão de crescimento
Antes da transição demográfica	Sim: mais ideias → mais produção → mais pessoas → mais ideias →	**Crescimento acelerado** (a economia pode passar de pequena para grande muito rapidamente)
Desde a transição demográfica	Não: mais ideias → mais produção → pessoas mais ricas	**Crescimento constante** (menos explosivo)
Com o Duplicador	Sim: mais ideias → mais produção → mais pessoas → mais ideias →	**Crescimento acelerado**

Esta figura de **Could Advanced AI Drive Explosive Economic Growth? (A Inteligência Artificial avançada poderia impulsionar o crescimento econômico explosivo?)** ilustra como as próximas décadas podem parecer diferentes com crescimento exponencial constante em comparação com o crescimento acelerado:



Para ver exemplos de números mais detalhados (mas simplificados) que demonstram o crescimento explosivo, consulte a nota de rodapé.[20]

Se quiséssemos adivinhar o que um Duplicador poderia fazer na realidade, poderíamos imaginar que ele causaria um retorno ao tipo de aceleração que a economia mundial teve historicamente. Isso significaria, vagamente, com base em ***Modeling the Human Trajectory (Modelando a trajetória humana)***, que **a economia atingiria tamanho *infinito* em algum momento no próximo século.**[21]

Claro, isso não acontecerá —em algum momento o tamanho da economia será limitado pelos recursos naturais básicos, como o número de átomos ou a quantidade de energia disponível na galáxia. Mas em algum ponto entre o momento que vivemos e ficar sem espaço/átomos/energia/alguma coisa, poderíamos facilmente ter níveis de crescimento econômico que seriam massivamente mais rápidos do que qualquer coisa na história.

Nos últimos 100 anos, a economia dobrou de tamanho a cada poucas décadas. Com um Duplicador, ela poderia dobrar de tamanho a cada ano ou mês, na direção de atingir os seus limites.

Dependendo de como as coisas acontecessem, essa produtividade poderia resultar no fim da escassez e da necessidade material, ou então em uma corrida distópica entre pessoas diferentes criando o máximo possível de cópias de si mesmas na esperança de dominar a população; ou em muitos outros cenários intermediários e alternativos.

Conclusão

Acho que o Duplicador seria uma tecnologia mais poderosa do que as "dobras espaciais" da Jornada nas Estrelas, *tricorders*, armas de laser[22] ou até teletransportadores. As mentes são a fonte de inovação que pode criar todas essas outras coisas. Portanto, conseguir duplicá-las com baixo custo seria uma situação extraordinária.

Uma tecnologia mais difícil de intuir, mas mais poderosa ainda, seriam as **pessoas digitais**, a capacidade de rodar simulações detalhadas de pessoas[23] em um computador.

Essas pessoas simuladas poderiam ser copiadas no estilo do Duplicador e também poderiam ser aceleradas, desaceleradas e redefinidas, com ambientes virtuais totalmente controlados.

Acho que esse tipo de tecnologia provavelmente será possível e espero que um mundo onde ela exista seja ainda mais "audacioso" do que um mundo com o Duplicador. Detalharei isso no próximo artigo.

Notas

[11]Por exemplo, o capitão Kirk de **Jornada nas Estrelas** assume o comando da nave *Enterprise* em meados dos anos 2200. Acho que poderíamos antes de 2100 viver em um mundo muito mais avançado e transformado do que o de Jornada nas Estrelas.

[12]**Exemplo**

[13]Não é bem assim que funciona nos quadrinhos, mas é como funcionará aqui.

[14]O Duplicador da história em quadrinhos queima depois de algumas cópias, mas ele é apenas um protótipo.

[15]Bióloga que co-inventou o CRISPR e ganhou o Prêmio Nobel em 2020.

[16]Cada ideia dobra a quantidade de milho.

[17]Uma população em crescimento mais rápido não significa *necessariamente* um avanço tecnológico mais rápido. Pode haver "retornos decrescentes": as primeiras ideias são mais fáceis de ter do que as seguintes, portanto, mesmo que o esforço para encontrar novas ideias aumente, novas ideias são encontradas mais lentamente. ***Are Ideas Getting HarderTo Find?*** **[As ideias estão ficando mais difíceis de encontrar?]**é um artigo bem conhecido sobre este tópico. População maior é igual a progresso tecnológico mais rápido se a população estiver crescendo mais rápido do que a dificuldade de ter novas ideias. Essa dinâmica é retratada de forma simplificada no gráfico: inicialmente as pessoas têm ideias que levam à duplicação da produção de milho, mas depois as ideias levam apenas a um aumento de 1,5 vezes na produção de milho.

[18]É crucial incluir a etapa "mais produção → mais pessoas", que com frequência não existe obrigatoriamente e não descreve o mundo de hoje, mas poderia descrever um mundo com O Duplicador. É usual que os modelos de crescimento incorporem as outras partes da retroalimentação: mais pessoas -→ mais ideias -→ mais produção.

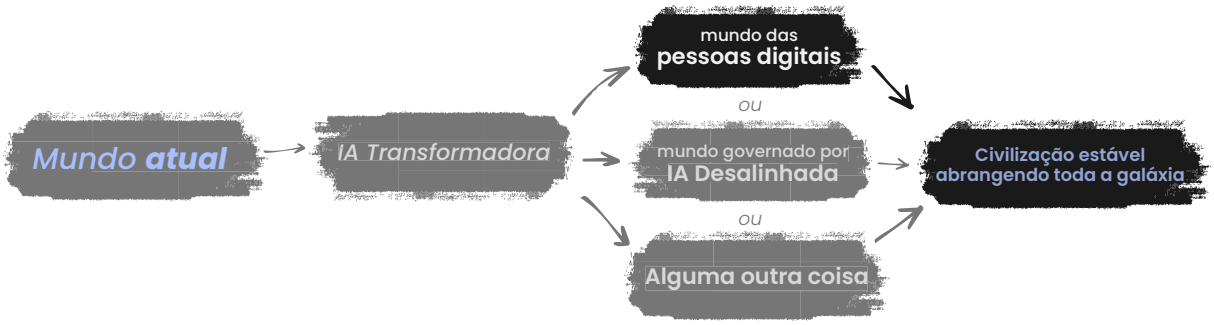
[19]Esta premissa é defendida em detalhes em Could Advanced AI Drive Explosive Economic Growth (Poderia a Inteligência Artificial avançada impulsionar o crescimento econômico explosivo?) 20Veja as notas de fim (**1**)

[21]Como mencionado acima, há um **debate** em aberto sobre se o crescimento econômico passado realmente segue o padrão descrito em ***Modeling the Human Trajectory*** **(Modelando a trajetória humana).** Discuto como o debate poderia mudar minhas conclusões aqui; de qualquer forma, acredito que há um caso suficientemente crível para defender a possibilidade do crescimento explosivo neste século.

[22]Honestamente, nunca consegui descobrir a razão dessas armas serem melhores do que as normais.

[23]Ou de algum tipo de entidade que seja adequadamente descrita como um "descendente" de pessoas, como discutirei no artigo sobre pessoas digitais.

Pessoas digitais seriam mais importantes ainda (Introdução)



Esta é a terceira postagem de uma série que apresenta a minha visão de que poderíamos estar no século mais importante de todos os tempos. (Aqui está o roteiro para esta série.)

- O primeiro artigo desta série discute a nossa era incomum, que pode estar muito próxima da transição entre uma civilização terráquea e uma civilização estável em toda a galáxia.
- Este artigo discute "pessoas digitais", uma categoria de tecnologia que poderia ser chave para esta transição, e que teria impactos ainda maiores do que o hipotético **Duplicador** discutido anteriormente.
- Muitas das ideias discutidas aqui aparecem em obras de ficção científica ou de não-ficção especulativa, mas não conheço outro artigo que exponha, brevemente, a ideia básica de pessoas digitais e as principais razões pelas quais um mundo de pessoas digitais seria tão diferente do mundo atual.
- A ideia das pessoas digitais fornece uma maneira concreta de imaginar como o tipo certo de tecnologia (que acredito ser quase certamente viável) poderia mudar o mundo radicalmente, de modo que "os humanos como os conhecemos" não seriam mais a força principal dos eventos mundiais

Será importante ter esse cenário em mente, porque argumentarei que os avanços na Inteligência Artificial neste século poderiam levar rapidamente ao desenvolvimento de pessoas digitais ou a uma tecnologia importante semelhante.

O potencial transformador de algo como pessoas digitais, combinado com a rapidez com que a IA poderia dar a isso corrobora a ideia de que poderíamos estar no século mais importante.

Introdução

Anteriormente, escrevi:

Quando algumas pessoas imaginam o futuro, elas imaginam o tipo de coisa que você vê nos filmes de ficção científica. Mas esses futuros de ficção científica parecem muito modestos em comparação com o futuro que eu vislumbro.

O futuro que imagino é enormemente maior, mais rápido, mais estranho e muito, muito melhor ou muito, muito pior comparado ao dia de hoje. Ele também chegará potencialmente muito mais cedo do que os futuros da ficção científica: acredito que tecnologias específicas e aparentemente alcançáveis poderiam nos levar até esse futuro rapidamente.

Este artigo é sobre **pessoas digitais**, um exemplo[24] de tecnologia que poderia nos conduzir a um futuro extremamente grande, rápido e estranho.

Para ter uma ideia do que seriam as pessoas digitais, imagine uma simulação computacional de uma pessoa específica, num ambiente virtual. Por exemplo, uma simulação de você mesmo, que reage a todos os "eventos virtuais" — fome virtual, clima virtual, um computador virtual com uma caixa de entrada de e-mail — exatamente como você o faria. Como no filme ***Matrix*** (Veja a nota de rodapé.)[25] Explico isso mais detalhadamente na seção de perguntas frequentes.

O cenário central em que vou me concentrar é o de pessoas digitais que se parecem com nós mesmos, talvez criadas por meio de **upload de mentes** (simulação de cérebros humanos). No entanto, pode-se também imaginar entidades diferentes de nós de várias maneiras, mas que ainda podem ser consideradas como "descendentes" da humanidade; essas também seriam pessoas digitais. Mais sobre a minha escolha de termos nas Perguntas Frequentes.

A cultura popular sobre esse tipo de tópico tende a se concentrar na perspectiva da **imortalidade digital**: pessoas evitando a morte assumindo uma forma digital, que pode ser copiada da mesma forma que você faz backup de seus dados. Mas considero que isso é pouco em comparação com outros impactos potenciais das pessoas digitais, em particular:

- **Produtividade.** As pessoas digitais poderiam ser copiadas, assim como podemos facilmente fazer cópias de qualquer software hoje. Elas também poderiam ter uma velocidade de processamento mental muito mais rápida do que a dos humanos. Por causa disso, as pessoas digitais poderiam ter efeitos comparáveis aos do Duplicador, mas mais ainda: elas poderiam impulsionar níveis sem precedentes, na história ou em filmes de ficção científica, de crescimento econômico e produtividade.

- **Ciências sociais.** Hoje vemos muito progresso na compreensão das leis científicas e no desenvolvimento de novas tecnologias interessantes, mas não tanto progresso na compreensão da natureza humana e do comportamento humano. Pessoas digitais mudariam fundamentalmente essa dinâmica: as pessoas poderiam fazer cópias de si mesmas (incluindo cópias aceleradas e temporárias) para explorar como diferentes escolhas, estilos de vida e ambientes as afetariam. A comparação de cópias seria instrutiva de uma forma que a ciência social atual raramente é.

- **Controle do ambiente.** As pessoas digitais experimentariam qualquer mundo que elas (ou o controlador de seu ambiente virtual) quisessem. Supondo que as pessoas digitais teriam uma experiência consciente verdadeira (uma suposição discutida na seção de Perguntas Frequentes), isso poderia ser algo bom; deveria ser possível eliminar doenças, pobreza material e violência não consensual para pessoas digitais). Ou, também, poderia ser algo ruim; se os direitos humanos não forem protegidos, pessoas digitais poderiam estar sujeitas a níveis assustadores de controle).

- **Expansão espacial.** A população de pessoas digitais poderia se tornar incrivelmente grande, e os computadores que as executassem poderiam estar distribuídos por toda a nossa galáxia e além. As pessoas digitais poderiam existir em qualquer lugar onde os computadores pudessem funcionar — assim, viver em assentamentos espaciais poderia ser menos difícil para as pessoas digitais do que para os humanos biológicos.

Aprisionamento/Lock-in. No mundo de hoje, estamos acostumados com a ideia de que o futuro é imprevisível e incontrolável. Regimes políticos, ideologias e culturas mudam e evoluem. Mas uma comunidade, cidade ou nação de pessoas digitais poderia ser muito mais estável.

- Pessoas digitais não precisariam morrer ou envelhecer.
- Quem criasse um "ambiente virtual" contendo uma comunidade de pessoas digitais poderia ter um controle bastante duradouro sobre como seria essa comunidade. Por exemplo, elas poderiam criar um software para redefinir a comunidade (tanto o ambiente virtual quanto as pessoas existentes nele) para um estado anterior se certas coisas mudassem — como quem estivesse no poder ou qual religião fosse dominante.
- Considero isso um pensamento perturbador, pois poderia permitir um autoritarismo duradouro, embora também pudesse permitir coisas como a proteção permanente de determinados direitos humanos.

Acho que esses efeitos (elaborados abaixo) poderiam ser uma coisa ótima ou péssima. Dependendo de como fossem os primeiros anos com as pessoas digitais, isso poderia determinar irreversivelmente como seriam esses efeitos.

Acho que consequências semelhantes surgiriam de qualquer tecnologia que permitisse

1. o controle extremo sobre nossas experiências e ambiente; (b) a duplicação de mentes humanas.

Isso significa que existem potencialmente **muitas maneiras do futuro se tornar tão alucinante quanto o que esboço aqui.** Discuto pessoas digitais porque isso nos fornece uma maneira particularmente fácil de imaginar as consequências de (a) e (b): trata-se essencialmente de transferir o alicerce mais importante do nosso mundo (mentes humanas) para um domínio (software) onde estamos acostumados com a ideia de ter um enorme controle para programar qualquer comportamento que quisermos.

Grande parte deste artigo é inspirado em ***The age of Em* (A era de Em)**, um livro incomum e fascinante. Ele tenta descrever um mundo hipotético de pessoas digitais (especificamente, upload de mentes) com muitos detalhes, mas (ao contrário da ficção científica) também visa a precisão preditiva em vez do entretenimento.

Em muitas partes do livro, acho isso excessivamente específico e, no geral, não espero que o mundo que ele descreve acabe tendo muito em comum com um mundo real repleto de pessoas digitais. No entanto, ele tem várias seções que, a meu ver, ilustram o quão poderosa e radical uma tecnologia como as pessoas digitais poderia ser.

Abaixo:

- Descreverei a ideia básica de pessoas digitais, com links com a seção de Perguntas Frequentes sobre a ideia.
- Examinarei as possíveis implicações listadas acima sobre as pessoas digitais.

Este é um artigo que pessoas diferentes podem querer ler em ordens diferentes.

Aqui está um guia geral para o artigo e a seção de Perguntas Frequentes:

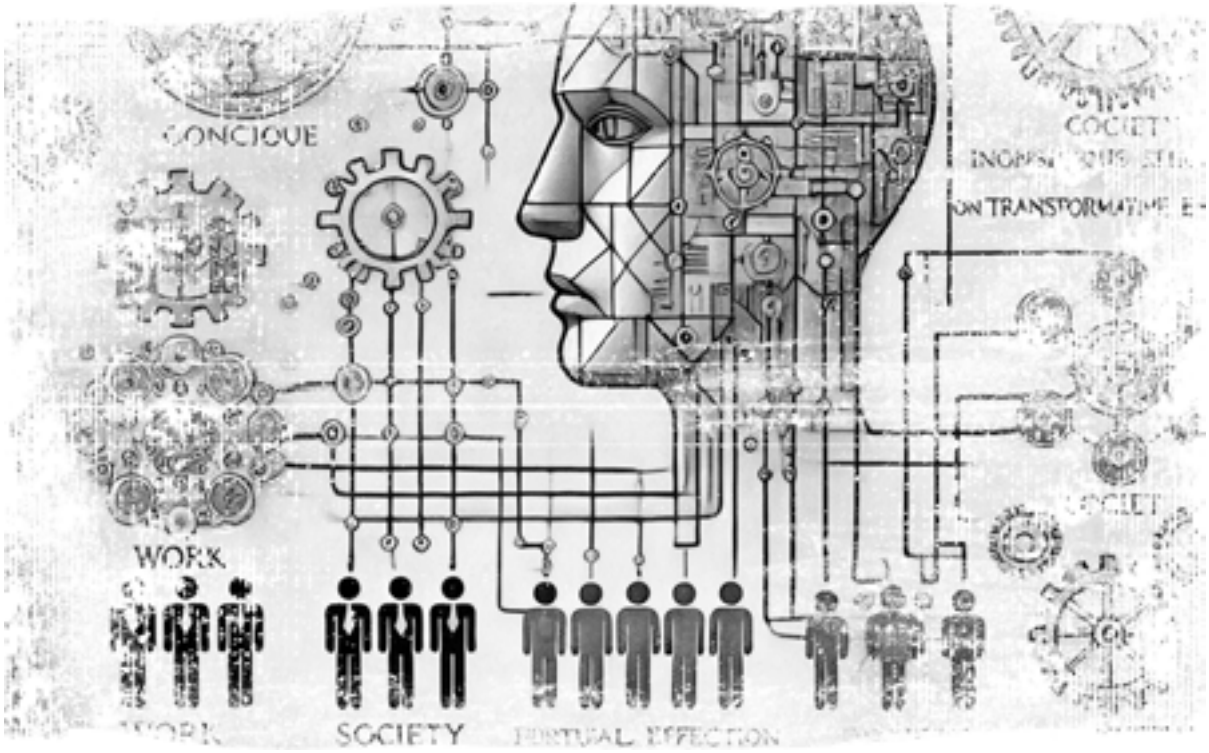
	Humanos normais	Pessoas digitais
Possível atualmente	✓	✗
Provavelmente possível algum dia	✓	✓
Podem interagir com o mundo real, fazer a maioria dos trabalhos	✓	✓
Conscientes, deveriam ter direitos humanos	✓	✓
Facilmente duplicados, como no **o Duplicador**	✗	✓
Podem ser acelerados	✗	✓
Podem efetuar **"cópias temporárias"** que são executadas rapidamente e depois se aposentam, sendo executadas em velocidade lenta	✗	✓
Produtividade e ciência social: poderiam causar crescimento econômico sem precedentes, produtividade e conhecimento da natureza e comportamento humanos	✗	✓
Controle do ambiente: podem ter suas experiências alteradas de qualquer maneira	✗	✓
Aprisionamento/Lock-in: poderiam viver em civilizações altamente estáveis sem envelhecimento ou morte, e "reinicializações digitais" interrompendo certas mudanças	✗	✓
Expansão espacial: podem viver confortavelmente em qualquer lugar que os computadores possam rodar, portanto, altamente adequadas para expansão em toda a galáxia	✗	✓
Bom ou mau?	Fora do escopo deste artigo	Pode ser ótimo ou péssimo

Este artigo se concentra em como as pessoas digitais poderiam mudar o mundo. Presumirei principalmente que **as pessoas digitais são como nós, exceto que podem ser facilmente copiadas, executadas em velocidades diferentes e incorporadas em ambientes virtuais.** Em particular, suporei que as pessoas digitais são conscientes, têm direitos humanos e podem fazer a maioria das coisas que os humanos podem, inclusive interagir com o mundo real. Acredito que **muitos leitores terão problemas para se envolver com isso até que tenham mais respostas para algumas perguntas básicas sobre as pessoas digitais.** Portanto, incentivo os leitores a clicar em quaisquer perguntas que pareçam úteis na seção de Perguntas Frequentes, ou apenas leiam a seção diretamente. Se você estiver lendo a versão "e-book" ou "PDF consolidado" desta série, a seção de Perguntas Frequentes será a próxima seção, seguida pelo restante deste artigo. Provavelmente faria sentido dar uma olhada no índice da seção e depois seguir adiante, dependendo se alguma das perguntas te parecer interessante ou importante.

Notas

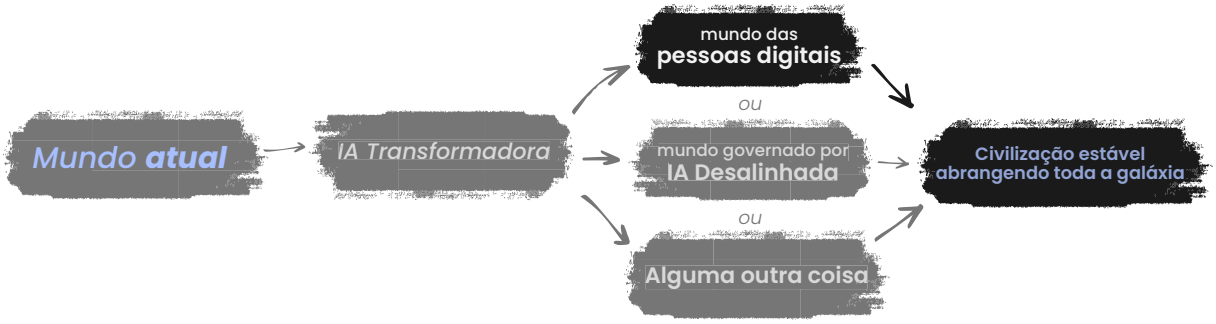
[24]O melhor exemplo em que posso pensar, mas certamente não é o único.

[25]O filme *Matrix* oferece uma intuição decente para a ideia, com sua realidade virtual totalmente imersiva, mas, ao contrário dos heróis de *Matrix*, uma pessoa digital não precisa estar conectada a nenhuma pessoa física —elas podem existir como software puro. Os agentes são mais parecidos com pessoas digitais do que os heróis. De fato, um deles **copia a si mesmo** repetidamente.



Pessoas digitais: Perguntas frequentes

Artigo que acompanha "Pessoas digitais seriam mais importantes ainda."



Este artigo acompanha Pessoas digitais seriam mais importantes ainda, o terceiro de uma série de postagens sobre a possibilidade de estarmos no **século mais importante da história da humanidade.**

Este artigo discute questões básicas sobre "pessoas digitais"; por exemplo, simulações computacionais extremamente detalhadas e realistas de pessoas específicas. Esta é uma tecnologia hipotética (mas, acredito, realista) que poderia ser a chave para uma transição para uma civilização estável em toda a galáxia. (Outro artigo descreve as *consequências* de tal tecnologia; este artigo se concentra em questões básicas sobre como isso pode funcionar).

Será importante ter esse cenário em mente, porque argumentarei que os avanços da Inteligência Artificial neste século poderiam levar rapidamente ao desenvolvimento de pessoas digitais ou a uma tecnologia importante semelhante.

O potencial transformador de algo como pessoas digitais, combinado com a rapidez com que a Inteligência Artificial levaria a isso, forma o argumento de que poderíamos estar no século mais importante.

Esta tabela (também no capítulo anterior) serve como um resumo dos dois artigos juntos:

	Humanos normais	Pessoas digitais
Possível atualmente	✓	✗
Provavelmente possível algum dia	✓	✓
Podem interagir com o mundo real, fazer a maioria dos trabalhos	✓	✓
Conscientes, deveriam ter direitos humanos	✓	✓
Facilmente duplicados, como no **o Duplicador**	✗	✓
Podem ser acelerados	✗	✓
Podem efetuar **"cópias temporárias"** que são executadas rapidamente e depois se aposentam, sendo executadas em velocidade lenta	✗	✓
Produtividade e ciência social: poderiam causar crescimento econômico sem precedentes, produtividade e conhecimento da natureza e comportamento humanos	✗	✓
Controle do ambiente: podem ter suas experiências alteradas de qualquer maneira	✗	✓
Aprisionamento/Lock-in: poderiam viver em civilizações altamente estáveis sem envelhecimento ou morte, e "reinicializações digitais" interrompendo certas mudanças	✗	✓
Expansão espacial: podem viver confortavelmente em qualquer lugar que os computadores possam rodar, portanto, altamente adequadas para expansão em toda a galáxia	✗	✓
Bom ou mau?	Fora do escopo deste artigo	Pode ser ótimo ou péssimo

Índice para as Perguntas Frequentes

Noções Básicas

Noções básicas sobre pessoas digitais

Para ter a ideia de pessoas digitais, imagine uma simulação computacional de uma pessoa específica, num ambiente virtual. Por exemplo, uma simulação sua que reagisse a todos os "eventos virtuais" (fome virtual, clima virtual, um computador virtual com uma caixa de mensagens) exatamente como você faria.

O filme *Matrix* oferece uma intuição adequada para a ideia com sua realidade virtual totalmente imersiva. Mas, ao contrário dos heróis de *Matrix*, uma pessoa digital não precisa estar conectada a nenhuma pessoa física — elas podem existir como software puro.[26]

Como outros softwares, as pessoas digitais podem ser copiadas (como no Duplicador) e executadas em velocidades diferentes. E seus ambientes virtuais não teriam que obedecer às regras do mundo real —eles poderiam funcionar da maneira que os designers de ambiente quisessem. Essas propriedades gerariam a maioria das consequências das quais falo no artigo principal.

Estou achando isso difícil de imaginar. Você poderia fazer uma analogia?

Hoje não existe nada que se pareça muito com uma pessoa digital, mas para começar a abordar a ideia, considere esta pessoa simulada:

Esse é o lendário jogador de futebol americano Jerry Rice, retratado no videogame ***Madden NFL 98***. Ele representa provavelmente o melhor que alguém naquela época (1997) poderia fazer para simular o verdadeiro Jerry Rice, no contexto de um jogo de futebol americano.

A ideia é que esse personagem de videogame corra, salte, faça recepções, deixe a bola cair e responda aos desarmes o mais próximo possível de como o verdadeiro Jerry Rice faria em situações análogas. (Pelo menos, é isso que ele faz quando o jogador de videogame não o está controlando explicitamente.) A simulação é uma versão muito rudimentar, simplificada e restrita a jogos de futebol americano da realidade. Com o passar dos anos, os videogames avançaram e suas simulações de Jerry Rice — assim como o resto dos jogadores, o campo de futebol, etc. — tornaram-se cada vez mais realistas:[27]

OK, a última é uma foto do verdadeiro Jerry Rice. Contudo, imagine que os desenvolvedores de jogos continuassem a criar suas versões de Jerry Rice cada vez mais realistas e que o universo do jogo se ampliasse,[28] a tal ponto, que o Jerry Rice simulado pudesse dar entrevistas a jornalistas virtuais, brincar com seus filhos virtuais, declarar impostos online e fazer tudo o que o verdadeiro Jerry Rice faria.

Nesse caso, o Jerry Rice simulado teria uma mente que funcionaria exatamente como a do verdadeiro Jerry Rice. Seria uma versão "pessoa digital" de Jerry Rice.

Agora imagine que alguém fizesse o mesmo para todas as pessoas, e você imaginará um mundo de pessoas digitais.

As pessoas digitais poderiam interagir com o mundo real? Por exemplo, uma empresa real poderia contratar uma pessoa digital para trabalhar nela?

A resposta é sim para ambas as perguntas.

- Uma pessoa digital poderia ser conectada a um corpo de robô. As câmeras alimentariam sinais de luz para a mente da pessoa digital e os microfones, os sinais sonoros; a pessoa digital enviaria sinais para, por exemplo, mover a mão, que iriam para o robô. Os humanos geralmente podem aprender a controlar implantes dessa maneira, então parece muito provável que as pessoas digitais também poderiam aprender a pilotar robôs.

- As pessoas digitais poderiam habitar um "escritório" virtual com um monitor virtual exibindo seu navegador da Web, um teclado virtual no qual digitariam, etc. Elas usariam essa configuração para enviar informações pela Internet, assim como os humanos biológicos o fazem (e como fazem os bots de hoje). Assim, elas responderiam e-mails, escreveriam e enviariam memorandos, postariam nas redes sociais e realizariam outros "trabalhos remotos" normalmente, sem precisar de nenhum "corpo" do mundo real.

O escritório virtual não precisaria ser como o mundo real em todos os seus detalhes — um ambiente virtual bastante simples com um "computador virtual" básico seria suficiente para uma pessoa digital realizar a maioria do "trabalho remoto".

- Elas também poderiam fazer chamadas telefônicas e de vídeo com humanos biológicos, transmitindo seu "rosto/voz virtual" para o humano biológico do outro lado da linha.

No geral, você poderia ter o mesmo relacionamento com uma pessoa digital que você teria com qualquer pessoa que você nunca conheceu pessoalmente.

Humanos e pessoas digitais

As pessoas digitais seriam conscientes? Elas teriam direitos humanos?

Imagine uma cópia digital detalhada de você, capaz de enviar e receber sinais de e para um corpo virtual em um mundo virtual. Essa pessoa digital envia sinais instruindo o corpo virtual a colocar a mão em um fogão virtual. Como resultado, a pessoa digital recebe sinais que simulam a sensação de queimadura na mão. Ela processa esses sinais e, em resposta, envia outros sinais para a boca, fazendo-a gritar "Ai!", e para a mão, ordenando que se afaste do fogão virtual.

Essa pessoa digital sente dor? Elas são realmente "conscientes" ou "sensíveis" ou "vivas?" Da mesma forma, devemos considerar a experiência da queimadura como um evento infeliz, que gostaríamos que tivesse sido evitado para que elas não tivessem que passar por isso?

Esta não é uma questão da Física ou Biologia, mas sim da Filosofia. E uma resposta completa está fora do escopo deste artigo.

Acredito que **simulações suficientemente detalhadas e precisas de humanos seriam conscientes, no mesmo grau e pelas mesmas razões que os humanos são conscientes.**[29]

É difícil colocar uma probabilidade nisso quando não está totalmente claro o que a afirmação significa, mas acredito que esta seja a melhor conclusão disponível, dado o estado da Filosofia da Mente acadêmica. Presumo que essa visão seja bastante comum, embora não universal, entre os filósofos da mente.[30]

Darei uma explicação abreviada do porquê, por meio de alguns experimentos mentais.

Experimento mental 1. Imagine que alguém, de alguma forma, substituísse um dos meus neurônios por um "neurônio digital": um dispositivo elétrico, feito do material de que os computadores atuais são feitos, em vez do material de que meus neurônios são feitos. E, que gravasse informações de entrada de outros neurônios (talvez usando uma câmera para monitorar os vários sinais que eles estavam enviando) e, então, enviasse informações de saída para eles exatamente no mesmo padrão do neurônio antigo.

Se alguém fizesse isso comigo, eu não me comportaria de maneira diferente, nem teria como "perceber" a diferença.

Agora imagine que alguém fizesse o mesmo com todos os outros neurônios do meu cérebro, um por um — de modo que meu cérebro tivesse apenas "neurônios digitais" conectados mutuamente, recebendo sinais de entrada de meus olhos/ouvidos/, etc. e enviando sinais de saída para meus braços/pés/, etc.

Eu ainda não me comportaria de maneira diferente de forma alguma, ou de alguma maneira "perceberia" isso.

Enquanto você estivesse trocando todos os neurônios, eu não notaria nenhuma diminuição na clareza dos meus pensamentos. Hipótese: se eu percebesse essa diminuição, essa "percepção" me afetaria de maneiras que poderiam alterar meu comportamento. Por exemplo, eu poderia notar a perda de clareza nos meus pensamentos. Mas já especificamos que os sinais de entrada e saída do meu cérebro não se alterariam, o que significa que o meu comportamento não mudaria.

Agora, imagine que alguém removesse o conjunto de "neurônios digitais" interconectados da minha cabeça e alimentasse sinais de entrada e saída similares diretamente (em vez de através de meus olhos/ouvidos/, etc.). Esta seria uma versão digital de mim: uma simulação do meu cérebro, rodando em um computador. E que, em nenhum momento, eu tivesse notado qualquer alteração — nenhuma diminuição de consciência, nenhum sentimento silenciado, etc.

Experimento mental 2. Imagine que eu tivesse uma conversa com uma cópia digital de mim mesmo — uma simulação extremamente detalhada de mim, que reagisse a todas as situações, exatamente como eu o faria.

Se eu perguntasse à minha cópia digital se ela era consciente, ela insistiria que sim (assim como eu o faria em resposta à mesma pergunta). Se eu explicasse e demonstrasse sua situação (por exemplo, que ela é uma pessoa "virtual") e perguntasse se ela ainda achava que era consciente, ela continuaria a insistir que sim (assim como eu o faria, se passasse pela experiência de ser informado que eu estava sendo simulado em algum computador — algo que minhas observações atuais não podem descartar).

Duvido que haja qualquer argumento que pudesse convencer meu interlocutor digital de que ele não era consciente. Se um processo de raciocínio, funcionando exatamente como o meu, com acesso a todos os mesmos fatos aos quais tenho acesso, está convencido de que "o Holden-digital é consciente", que base racional eu teria para pensar que isso está errado?

Considerações gerais:

- Imagino que seja o que for a consciência, ela é a causa de coisas como "eu digo que sou consciente" e também a fonte de minhas observações sobre minha própria experiência consciente. O fato de meu cérebro ser feito de neurônios (em oposição a chips de computador ou qualquer outra coisa) não é nada que não desempenhe nenhum papel em minha propensão a dizer que sou consciente, ou nas observações que faço sobre minha própria experiência consciente: se meu cérebro fosse um computador em vez de um conjunto de neurônios, enviando os mesmos sinais de saída, eu expressaria as mesmas crenças e observações sobre minha própria experiência consciente.

- A causa de minhas afirmações sobre a consciência e a fonte de minhas observações sobre minha própria consciência não tem relação com o *material de que meu cérebro é feito; em vez disso, é algo que tem relação com os padrões de processamento de informações que meu cérebro executa.* Um computador executando os mesmos padrões de processamento de informações teria, portanto, tantos motivos para se considerar consciente quanto eu tenho.

- Finalmente, meu entendimento ao conversar com físicos é que muitos deles acreditam que existe algum sentido importante no argumento de que "o universo só pode ser entendido fundamentalmente como padrões de processamento de informações" e que a distinção entre, por exemplo, neurônios e processadores de computador parece improvável de ter algo "profundo" a seu respeito.[31]

Para mais detalhes sobre este tópico, veja:

- Seção 9 de ***The Singularity: A Philosophical Analysis*** **(Singularidade: uma análise filosófica)** de David Chalmers. Um raciocínio semelhante aparece na parte III do livro de Chalmers ***The Conscious Mind*** **(A mente consciente)**.

- ***Zombies Redacted*** **(Zumbis - Editado)** de Eliezer Yudkowsky. Este é mais informal e menos acadêmico, e seus argumentos são mais parecidos com os que fiz acima.

Digamos que você esteja errado e as pessoas digitais não pudessem ser conscientes. Como isso afetaria os seus pontos de vista sobre como elas poderiam mudar o mundo?

Digamos que pudéssemos fazer duplicatas digitais dos humanos atuais, mas que elas não fossem conscientes. Nesse caso:

- Elas ainda seriam extremamente produtivas em comparação com os humanos biológicos. E estudá-las esclareceria fatos sobre a natureza e o comportamento humanos. Então, as seções sobre Produtividade e Ciência Social ficariam praticamente inalteradas.

- Elas ainda acreditariam ser conscientes (já que acreditamos ser, e elas seriam simulações de nós). Elas ainda poderiam procurar expandir-se pelo espaço e estabelecer comunidades estáveis/"aprisionadas" para preservar os valores com os quais se importam.

- Devido à sua produtividade e grandes números, eu presumiria que a população de pessoas digitais determinaria como seria o futuro a longo prazo da galáxia — inclusive para os humanos biológicos.

- Os riscos gerais seriam menores, se o grande número de pessoas digitais em toda a galáxia e as experiências virtuais que elas tivessem "não importassem". Mas os riscos seriam ainda maiores, já que a maneira como as pessoas digitais configurassem a galáxia determinaria como seria a vida humana.

Viabilidade

As pessoas digitais são possíveis?

Elas certamente não são possíveis hoje. Não temos ideia de como criar um software que "responda" a dados de vídeo e áudio (por exemplo, enviando os mesmos sinais para falar, mover, etc.) da maneira que um ser humano em particular o faria.

E não podemos simplesmente copiar e simular cérebros humanos, porque relativamente pouco se sabe sobre o que o cérebro humano faz. Os neurocientistas têm uma capacidade muito limitada de fazer observações sobre isso.[32] (Podemos fazer um bom trabalho simulando alguns dos principais *sinais de entrada* do cérebro - as câmeras parecem captar imagens tão bem quanto os olhos humanos, e os microfones parecem capturar o som tão bem quanto os ouvidos humanos.[33])

As pessoas digitais são uma tecnologia hipotética e podemos um dia descobrir que elas são impossíveis. Mas, pelo que sei, não há nenhuma razão atual para acreditar que elas sejam impossíveis.

Eu pessoalmente apostaria que elas eventualmente serão possíveis — pelo menos através do upload de mentes (escaneamento e simulação de cérebros humanos).[34]

Acho que é uma questão de (a) a neurociência avançar até o ponto em que possamos observar e caracterizar minuciosamente os principais detalhes do que os cérebros humanos estão fazendo — um caminho potencialmente muito longo, mas não sem fim; (b) desenvolver um software que simule esses detalhes importantes; (c) executar a simulação de software em um computador; (d) fornecer um corpo virtual e um ambiente virtual "suficientemente bons", o que poderia ser bastante simples (possibilitando, por exemplo, falar, ler e digitar, o que seria um avanço significativo). Acho que (a) esta é a parte difícil e imaginaria que isso (c) poderia ser feito até mesmo no hardware de um computador atual.[35]

Não detalharei isso neste artigo, mas posso fazê-lo no futuro, se houver interesse.

Em quanto tempo as pessoas digitais seriam possíveis?

Não acredito que tenhamos uma boa maneira de prever quando os neurocientistas entenderão o cérebro bem o suficiente para começar a fazer o upload de mentes — além de dizer que não parecemos estar nem perto disso hoje.

A razão pela qual acho que as pessoas digitais poderiam surgir nas próximas décadas é diferente: acredito que poderíamos inventar outra coisa (principalmente inteligência artificial avançada) que aceleraria drasticamente a pesquisa científica. Se isso acontecer, veremos todos os tipos de novas tecnologias que mudam o mundo emergindo rapidamente — incluindo pessoas digitais.

Também acredito que pensar em pessoas digitais ajuda a formar intuições sobre o quão produtiva e poderosa a Inteligência Artificial avançada poderia ser (discutirei isso em um artigo futuro).

Outras dúvidas

Estou tendo problemas para imaginar um mundo de pessoas digitais — como a tecnologia poderia ser introduzida, como elas interagiriam conosco, etc. Você poderia traçar um cenário detalhado de como seria a transição do mundo de hoje para um mundo cheio de pessoas digitais?

Darei um exemplo de como as coisas poderiam acontecer. É um pouco enviesado para o lado otimista, então isso não se tornaria imediatamente uma distopia. E está enviesado para o lado mais "conhecido": não exploro todas as possíveis consequências radicais das pessoas digitais.

Nada mais no artigo depende de que esta história seja precisa; o único objetivo é tornar um pouco mais fácil imaginar este mundo e pensar sobre as motivações das pessoas nele.

Então, imagine que:

Um dia, uma tecnologia de upload de mentes funcionais estivesse disponível. Para simplificar, suponhamos que o preço fosse modesto desde o início.[36] O que isso significaria: quem quisesse poderia ter seu cérebro escaneado, criando uma "cópia digital" de si mesmo.

Algumas dezenas de milhares de pessoas criariam "cópias digitais" de si mesmas. Portanto, agora existiriam dezenas de milhares de pessoas digitais vivendo em um ambiente virtual simples, composto por prédios de escritórios, apartamentos e parques.

Inicialmente, cada pessoa digital pensaria como a pessoa não digital da qual foi copiada, embora, com o passar do tempo, suas experiências de vida e estilos de pensamento divergissem.

Cada pessoa digital conseguiria projetar seu próprio "corpo virtual" que a representaria no ambiente. (Isso é um pouco como escolher um avatar - os corpos precisariam estar em uma faixa normal de altura, peso, força, etc., mas seriam bastante personalizáveis.)

O servidor de computador que rodaria todas as pessoas digitais e o ambiente virtual nos quais elas habitassem seria de propriedade privada. No entanto, graças à regulamentação presciente, as próprias pessoas digitais seriam consideradas pessoas com plenos direitos legais (não propriedade de seus criadores ou da empresa de servidores). Elas fariam suas próprias escolhas, sujeitas à lei, e contariam com algumas proteções iniciais básicas, como:

- Para continuarem existindo, o proprietário do servidor onde estivessem instaladas deveria optar por executá-las. No entanto, cada pessoa digital inicialmente deveria ter um contrato pré-pago de longo prazo com qualquer empresa de servidor que as estivesse executando no início, para que pudessem ter certeza de existir por um longo tempo. Digamos, pelo menos 100 anos a partir da data de nascimento de sua cópia biológica — se assim o desejassem.

- Elas deveriam ser completamente informadas sobre sua situação como pessoa digital, receber outras informações sobre o que estivesse acontecendo, poder entrar em contato com pessoas-chave, etc. Da mesma forma, inicialmente apenas pessoas com 18 anos ou mais poderiam ser copiadas digitalmente, embora pessoas digitais posteriores pudessem ter seus próprios "filhos digitais" — veja abaixo.

- Seu ambiente virtual deveria atender a certos critérios inicialmente (por exemplo, nenhuma violência ou sofrimento infligido a elas, ampla oferta de comida e água virtual). Elas teriam sua própria conta bancária com algum dinheiro para começar e poderiam ganhar mais, assim como as pessoas biológicas (por exemplo, trabalhando para alguma empresa).

- O proprietário do servidor não poderia fazer nenhuma alteração significativa em seu ambiente virtual sem seu consentimento (além de parar de executá-las, o que poderia ser realizado após o término do contrato, após algumas décadas). Pessoas digitais poderiam solicitar e oferecer dinheiro para mudanças em seu ambiente virtual (embora qualquer outra pessoa digital afetada também precisasse dar seu consentimento).

- O proprietário do servidor deveria interromper a execução de quaisquer pessoas digitais que solicitassem a interrupção das suas existências.

As pessoas digitais estabeleceriam relações profissionais e pessoais entre si. Elas também estabeleceriam relacionamentos pessoais e profissionais com humanos biológicos, com quem se comunicariam por e-mail, bate-papo por vídeo, etc.

- Poderiam trabalhar para a primeira empresa a oferecer cópias digitais de seres humanos, fazendo pesquisas sobre como tornar as futuras pessoas digitais mais baratas.
- Elas poderiam manter contato com a pessoa biológica a partir da qual foram copiadas, trocando e-mails sobre suas vidas pessoais.
- Elas estariam quase certamente interessadas em garantir que nenhum ser humano biológico interferisse em seu servidor de maneiras indesejadas, desligando-as, por exemplo.

Algumas pessoas digitais se apaixonariam e se casariam. Um casal poderia "ter filhos" criando uma pessoa digital cuja mente seria um híbrido de suas duas mentes. Inicialmente (sujeito a proteções contra abuso infantil), elas poderiam decidir como seu filho apareceria no ambiente virtual e até mesmo fazer alguns ajustes, como "Quando o cérebro da criança enviasse um sinal para fazer cocô, um arco-íris apareceria". A criança conquistaria direitos à medida que envelhecesse, como acontece com os humanos biológicos.

Pessoas digitais também poderiam se copiar, desde que cumprissem os requisitos para novas pessoas digitais (garantia de poder viver por um tempo razoavelmente longo, etc.) As cópias teriam seus próprios direitos e não deveriam nada aos seus criadores.

A população de pessoas digitais cresceria, por meio de pessoas que se copiassem e tivessem filhos. Eventualmente (talvez rapidamente, como discutido abaixo), haveria muito mais pessoas digitais do que humanos biológicos. Ainda assim, algumas pessoas digitais trabalhariam, empregariam ou teriam relações pessoais (via e-mail, bate-papo por vídeo, etc.) com humanos biológicos.

- Muitas pessoas digitais trabalhariam para possibilitar um maior crescimento populacional — tornando mais barato administrar pessoas digitais, construindo mais computadores (no mundo "real"), encontrando novas fontes de matérias-primas e energia para computadores (também no mundo "real"), etc.
- Muitas outras pessoas digitais trabalhariam na criação de ambientes virtuais cada vez mais criativos, alguns baseados em locais reais, outros mais exóticos (física alterada, etc.) Alguns ambientes virtuais seriam projetados para serem habitados, enquanto outros seriam projetados para serem visitados para recreação. O acesso seria vendido para pessoas digitais que desejassem ser transferidas para esses ambientes.

Então, as pessoas digitais estariam trabalhando, se divertindo, se encontrando, se reproduzindo, etc. Nesses aspectos, suas vidas teriam bastante em comum com as nossas.

- Como nós, elas teriam algum incentivo para trabalhar por dinheiro. Elas precisariam pagar pelos custos do servidor se quisessem continuar existindo por mais tempo do que o seu contrato inicial, ou se quisessem se copiar ou ter filhos (elas precisariam comprar contratos longos de servidores para quaisquer novas pessoas digitais), ou se quisessem participar de vários ambientes e atividades recreativas.
- Ao contrário de nós, elas poderiam fazer coisas como copiar a si mesmas, rodar em diferentes velocidades, mudar seus corpos virtuais, entrar em ambientes virtuais exóticos (por exemplo, com gravidade zero), etc.

Os reguladores prescientes criariam maneiras para grandes grupos de pessoas digitais formarem seus próprios estados e civilizações virtuais, que poderiam definir e alterar seus próprios regulamentos.

Alternativas distópicas. Um mundo de pessoas digitais poderia rapidamente se tornar distópico se houvesse uma regulamentação pior ou nenhuma regulamentação. Por exemplo, imagine se a regra fosse "Quem for o dono do servidor poderá executar o que quiser nele".

Em seguida, as pessoas poderiam fazer cópias digitais de si mesmas nas quais realizariam experimentos, seriam forçadas a trabalhar e até mesmo teriam seu código aberto, para que qualquer pessoa com um servidor pudesse fazer cópias e abusar delas. **Este conto muito curto** (recomendado, mas arrepiante) dá uma ideia de como isso poderia ser.

Existem outras maneiras (mais graduais) de um mundo de pessoas digitais se tornar distópico, conforme descrito **aqui** (autoritarismo incontestável) e no Duplicador (pessoas competindo para fazer cópias umas das outras e dominar a população).

E o que os humanos biológicos estariam fazendo enquanto isso? Ao longo desta seção, falei sobre como o mundo seria *para pessoas digitais*, não para humanos biológicos normais. Estou mais focado nisso, porque prevejo que as pessoas digitais rapidamente se tornariam a maioria da população, e acredito que deveríamos nos preocupar com elas tanto quanto nos preocupamos com os humanos biológicos. Mas se você está se perguntando como seriam as coisas para os humanos biológicos, eu suponho que:

- Pessoas digitais, devido ao seu número e velocidade de processamento, se tornariam os atores políticos e militares dominantes no mundo. Elas seriam provavelmente as pessoas que determinariam como seria a vida dos humanos biológicos.
- Haveria um avanço científico e tecnológico muito rápido (como discutido abaixo). Portanto, supondo que as pessoas digitais e os humanos biológicos permanecessem com boas relações, eu suporia que os humanos biológicos teriam acesso a tecnologias muito mais avançadas do que as que temos atualmente. No mínimo, presumiria que isso significaria tecnologias médicas praticamente ilimitadas (incluindo, por exemplo, "curar" o envelhecimento e ter uma expectativa de vida indefinidamente longa).

As pessoas digitais diferem dos uploads de mentes?

Uploads de mentes refere-se à simulação de um cérebro humano em um computador. (Geralmente está implícito que isso não seria literalmente um cérebro isolado, ou seja, incluiria algum tipo de ambiente e corpo virtuais para a pessoa que fosse simulada, ou talvez elas estariam pilotando um robô)

O upload de mentes seria uma forma de pessoa digital, e a maioria deste artigo poderia ter sido escrito sobre uploads de mentes. O upload de mentes é a versão mais fácil de imaginar as pessoas digitais, e eu me concentro nela quando falo sobre o motivo pelo qual acredito que as pessoas digitais algum dia serão possíveis e o motivo pelo qual elas seriam conscientes como nós.

Mas também posso imaginar um futuro de "pessoas digitais" que não sejam derivadas da cópia de cérebros humanos, ou mesmo muito semelhantes aos humanos de hoje. Acho que é razoavelmente provável que quando as pessoas digitais forem possíveis (ou logo depois), elas serão bem diferentes dos humanos de hoje.[37]

A maioria deste artigo se aplicaria a praticamente qualquer entidade digital que:

1. tivesse valor moral e direitos humanos, como pessoas não digitais;
2. pudessem interagir com seus ambientes com habilidade e engenhosidade iguais (ou maiores) que as pessoas de hoje.

Com compreensão suficiente de como (a) e (b) funcionam, deveria ser possível projetar pessoas digitais sem imitar cérebros humanos.

Vou me referir muito a pessoas digitais ao longo desta série para indicar o quão radicalmente diferente o futuro pode ser. Não quero ser interpretado como dizendo que isso envolveria necessariamente a cópia de cérebros humanos reais.

Uma cópia digital minha, seria eu?

Digamos que alguém digitalizasse meu cérebro e criasse uma simulação dele em um computador: uma cópia digital de mim. Isso contaria como "eu"? Devo esperar que essa pessoa digital tenha uma vida boa, tanto quanto espero isso para mim?

Esta é outra questão da filosofia. Minha resposta básica é "Mais ou menos, mas isso não importa muito". Este artigo é sobre como as pessoas digitais poderiam mudar o mundo radicalmente; isso não depende de nós nos identificarmos com nossas próprias cópias digitais.

Isso depende (um pouco) da questão de as pessoas digitais deverem ser consideradas "pessoas completas", ou não, no sentido de que nos preocupamos com elas, queremos que evitem más experiências, etc. **A seção sobre consciência é mais relevante para esta questão.**

Quais outras perguntas posso fazer?

Tantas outras!

Por exemplo:

https://tvtropes.org/pmwiki/pmwiki.php/Analysis/BrainUp- loading

Por que tudo isso importa?

O artigo que acompanha pessoas digitais são mais importantes ainda, explica uma série de maneiras pelas quais as pessoas digitais poderiam levar a um futuro radicalmente desconhecido.

Em outra parte desta série, argumentarei que os avanços da Inteligência Artificial neste século poderiam levar rapidamente ao desenvolvimento de pessoas digitais ou a uma tecnologia importante semelhante. O potencial transformador de algo como pessoas digitais, combinado com a rapidez com que a Inteligência Artificial poderia levar a isso, forma o argumento de que poderíamos estar no século mais importante.

Notas

[26]Os "agentes" são mais parecidos com pessoas digitais. De fato, um deles **copia a si mesmo** extensivamente.

[27]Todas foram tiradas **deste vídeo**, exceto a última.

[28]Os videogames de futebol americano já se expandiram para simular **negociações fora de temporada, contratações e definição de preços de ingressos**.

[29]Também seria possível que existissem "pessoas digitais" conscientes que não se parecessem com os humanos de hoje, mas não vou entrar nisso aqui – vou me concentrar apenas no exemplo concreto de "pessoas digitais" que são versões virtuais dos humanos.

[30]Conforme o ***Phil Papers Surveys***, 56,5% dos filósofos endossam o **fisicalismo**, contra 27,1% que endossam o não-fisicalismo e 16,4% "outros". Acredito que a grande maioria dos filósofos que endossam o **fisicalismo** concordaria que uma simulação suficientemente detalhada de um ser humano seria consciente. Meu entendimento é que o **naturalismo biológico** é uma posição marginal/impopular, e que o fisicalismo + a rejeição do naturalismo biológico implicariam em acreditar que simulações suficientemente detalhadas de humanos seriam conscientes. Também presumo que alguns filósofos que não endossam o fisicalismo ainda acreditariam que tais simulações seriam conscientes. David Chalmers é um exemplo —veja ***The ConsciousMind* (A mente consciente)**. Essas presunções são baseadas apenas em minhas impressões sobre a área de conhecimento.

[31]De um e-mail de um amigo físico: “Acho que muitas pessoas têm a intuição de que a atividade neural real, produzida por reações químicas reais de neurotransmissores reais, e a atividade elétrica real que você pode sentir com a sua mão, tem, de alguma forma, algumas propriedades que o mero código de computador não poderia ter. Mas uma das mensagens avassaladoras da física moderna é que tudo o que existe — partículas, campos, átomos, etc., é melhor pensado em termos de informação e pode simplesmente *ser* informação. O universo talvez possa ser mais bem descrito como uma abstração matemática. As reações químicas não vêm de alguma propriedade essencial dos átomos, mas de interações sutis entre suas camadas de elétrons de valência. Elétrons e prótons não são partículas bem definidas, elas são nuvens abstratas de massa de probabilidade. Até mesmo o conceito de “partículas” é enganoso; parecendo realmente existir são campos quânticos, os quais são as soluções de equações matemáticas abstratas, e alguns de cujos estados são rotulados pelos humanos como “1 partícula” ou “2 partículas”. Para ser um pouco metafórico, somos como pequenas ondulações em vastas ondas matemáticas abstratas, ondulações cujos padrões e dinâmicas acontecem para executar o processamento de informações correspondente ao que chamamos de senciência. Na minha opinião, nossa existência e o substrato em que vivemos já são muito mais estranhos e efêmeros do que qualquer coisa para a qual possamos fazer upload de humanos.”

[32]Para uma ilustração disso, veja este relatório: ***How much computational power does it take to match the human brain?* (Quanto poder computacional é necessário para igualar o cérebro humano?)**. Particularmente a seção ***Uncertainty in neuroscience* (A incerteza na neurociência)**. Até mesmo estimar quantas operações significativas o cérebro humano realiza é, hoje, muito difícil e complicado — muito menos caracterizar quais são essas operações.

[33]Esta afirmação é baseada em minha compreensão da sabedoria convencional, além do fato de que o vídeo e o áudio gravados geralmente parecem bastante realistas, o que implica que a câmera/microfone não deixou de registrar muitas informações importantes sobre sua fonte.

[34]Pressupondo que a tecnologia continuaria avançando, que a espécie não seria extinta, etc.

[35]**Este relatório conclui** que um computador que custa ~ US$ 10.000 hoje tem poder computacional suficiente (10^14 FLOP/s, uma medida de poder computacional) para estar dentro de 1/10 do melhor palpite do autor sobre o que seria necessário para replicar o comportamento de entrada-saída de um cérebro humano (1015 FLOP/s). Se considerarmos a **estimativa alta** do autor, em vez do melhor palpite, é cerca de 10 milhões de vezes mais computação (1022 FLOP/s), o que presumivelmente custaria US $1 trilhão hoje - provavelmente um valor muito alto para valer a pena, mas a computação está ficando cada vez mais barata. É possível que replicar apenas o comportamento de entrada-saída não seja detalhado o suficiente para atingir a “consciência” - embora eu ache que o seria - e, de qualquer forma, seria suficiente para ter consequências na “produtividade e ciências sociais”.

[36]Na verdade, imagino que seria muito caro inicialmente, mas se tornaria mais barato muito rapidamente devido a uma explosão de produtividade, discutida abaixo.

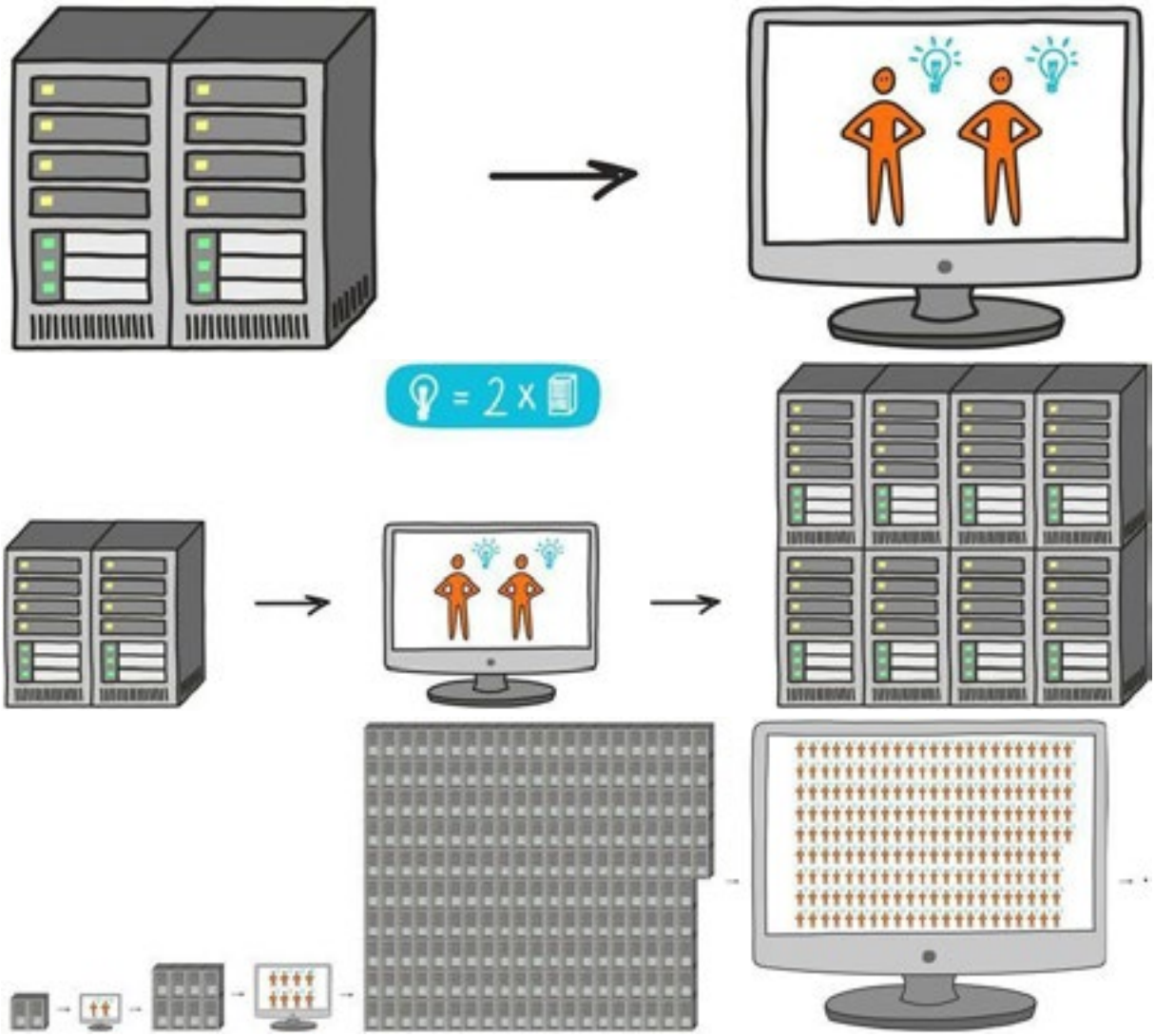
[37]Eu também poderia imaginar um futuro no qual as duas propriedades principais listadas no próximo parágrafo — (a) valor moral e direitos humanos, (b) capacidades de nível humano ou superior — seriam totalmente separadas. Ou seja, poderia haver um mundo cheio de (a) IAs com capacidades de nível humano ou superior, mas sem consciência ou valor moral; (b) entidades digitais com valor moral e experiência consciente, mas muito poucas habilidades em comparação com as IAs e até mesmo em comparação com as pessoas de hoje. A maioria das coisas que digo neste artigo sobre um mundo de “pessoas digitais” se aplicaria a esse mundo; neste caso, você poderia pensar em “pessoas digitais” como “equipes” de IAs e entidades moralmente valiosas, mas com poucas habilidades.

Pessoas digitais seriam mais importantes ainda: Sessão final

Como pessoas digitais poderiam mudar o mundo?

Produtividade

Como qualquer software, as pessoas digitais poderiam ser copiadas instantaneamente e com precisão. O Duplicador argumenta que a capacidade de "copiar as pessoas" poderia levar a uma rápida aceleração do crescimento econômico: "Nos últimos 100 anos, a economia dobrou de tamanho a cada poucas décadas. Com um Duplicador, ela poderia dobrar de tamanho a cada ano ou mês, a caminho de atingir os seus limites."



Obrigado a María Gutiérrez Rojas por estes gráficos, uma variação de um conjunto semelhante de gráficos de **O Duplicador** que ilustra como a duplicação de pessoas poderia causar um crescimento explosivo.

As pessoas digitais poderiam criar um efeito mais dramático do que isso, devido à capacidade de terem seus processamentos acelerados (talvez milhares ou milhões de vezes)[38] , bem como desacelerados (para economizar custos). Isso poderia aumentar a velocidade e a capacidade de coordenação.[39]

Outro fator que poderia aumentar a produtividade: pessoas digitais "temporárias" poderiam concluir uma tarefa e depois se aposentarem para ter uma boa vida virtual, enquanto seriam rodadas em baixa velocidade (e com baixo custo).[40] Isso faria com que algumas pessoas digitais se sentissem confortáveis em copiar a si mesmas para fins temporários.

As pessoas digitais poderiam, por exemplo, copiar a si mesmas centenas de vezes para tentar diferentes abordagens de resolver um problema ou adquirir uma habilidade e, em seguida, manteriam apenas a versão mais bem-sucedida e fariam muitas cópias dessa versão.

Possivelmente, as pessoas digitais seriam uma força econômica *menor* do que **O Duplicador** , já que as pessoas digitais careceriam de corpos humanos. Mas isso parece ser apenas uma consideração menor (detalhes na nota de rodapé).[41]

Ciências sociais

Hoje, vemos muita inovação e progresso impressionantes em algumas áreas e relativamente pouco em outras áreas.

Por exemplo, podemos constantemente comprar computadores mais baratos e mais rápidos e videogames mais realistas, mas não parecemos estar constantemente melhorando em fazer amigos, apaixonar-nos ou encontrar a felicidade.[42] Também não estamos claramente melhorando em coisas como lutar contra o vício e nos comportar como queremos (em reflexão).

Uma maneira de pensar sobre isso é que as *ciências naturais* (por exemplo, física, química, biologia) estão avançando de forma muito mais impressionante do que as *ciências sociais* (por exemplo, economia, psicologia, sociologia). Ou: "Estamos fazendo grandes progressos na compreensão das leis naturais, mas nem tanto na compreensão de nós mesmos."

Pessoas digitais poderiam mudar isso. Elas poderiam abordar o que vejo como talvez a **razão fundamental pela qual a ciência social é tão difícil de aprender: é muito difícil realizar experimentos verdadeiros e fazer comparações precisas.**

Hoje, se quisermos saber se a meditação é útil para as pessoas:

- Podemos comparar pessoas que meditam com pessoas que não o fazem, mas haverá muitas diferenças entre essas pessoas e não podemos isolar o efeito da meditação em si. (Pesquisadores tentam fazer isso com várias técnicas estatísticas, mas elas têm seus próprios problemas.)

- Também poderíamos tentar realizar um experimento no qual as pessoas são designadas aleatoriamente para meditar ou não. Mas precisaríamos que muitas pessoas participassem, todas simultaneamente, e sob as mesmas condições, na esperança de que as diferenças entre meditadores e não meditadores fossem estatisticamente "eliminadas" e captássemos os efeitos da meditação. Atualmente, esses tipos de experimentos — conhecidos como "ensaios controlados randomizados" — são caros, logisticamente desafiadores, demorados e quase sempre terminam com resultados ambíguos e difíceis de interpretar.

Mas em um mundo com pessoas digitais:

- Qualquer um poderia fazer uma cópia de si mesmo para experimentar a meditação, talvez até se dedicando a ela por vários anos (possivelmente aceleradamente).[43] Se gostassem dos resultados, poderiam então meditar por vários anos e garantir que todas as cópias futuras fossem feitas por alguém que tivesse colhido os benefícios da meditação.

- Cientistas sociais poderiam estudar pessoas que haviam tentado coisas assim e procurar padrões, o que seria muito mais esclarecedor do que a pesquisa em ciências sociais tende a ser atualmente. (Eles também poderiam realizar experimentos deliberados, recrutando/pagando pessoas para fazer cópias de si mesmas para experimentar diferentes estilos de vida, cidades, escolas, etc. — estes poderiam ser menores, mais baratos e mais definitivos do que os experimentos de ciências sociais de hoje.[44])

A capacidade de realizar experimentos pode ser boa ou ruim, dependendo da robustez e aplicação da ética científica. Se o consentimento informado não fosse suficientemente protegido, as pessoas digitais poderiam abrir a porta para uma enorme quantidade de abuso potencial; mas se ao contrário, ele fosse protegido, isso poderia possibilitar principalmente o aprendizado.

Pessoas digitais também poderiam permitir:

- **Superação do viés.** As pessoas digitais poderiam fazer cópias de si mesmas (incluindo cópias temporárias e aceleradas) para considerar argumentos apresentados de várias maneiras diferentes, por pessoas diferentes, inclusive com raça e gênero aparentemente diferentes, e ver se as cópias chegariam a conclusões diferentes. Dessa forma, elas poderiam explorar quais vieses cognitivos —de sexismo e racismo

- até pensamento ilusório e ego — afetariam seus julgamentos e trabalhar para melhorar e adaptar esses vieses. (Mesmo que as pessoas não estivessem empolgadas para fazer isso, talvez elas precisem fazê-lo, pois outras pessoas poderiam pedir informações sobre o quão tendenciosas elas são e esperar obter dados claros.)

- **Uma fortuna de reflexão e discussão.** As pessoas digitais poderiam fazer cópias de si mesmas (incluindo cópias temporárias aceleradas) para estudar e discutir, em profundidade, questões específicas de filosofia, questões de psicologia, etc. e, em seguida, resumir suas descobertas para a original.[45] Ao ver como diferentes cópias com diferentes conhecimentos e experiências de vida formaram opiniões diferentes, elas poderiam ter respostas muito mais ponderadas e embasadas do que eu para perguntas como “O que eu quero na vida?”, “Por que eu quero isso?”, “Como posso ser uma pessoa da qual eu me orgulhe?”, etc.

Realidade virtual e controle do ambiente

Como dito acima, as pessoas digitais poderiam viver em “ambientes virtuais”. Para projetar um ambiente virtual, os programadores gerariam sistematicamente o tipo certo de sinais de luz, sinais sonoros, etc. para enviar para uma pessoa digital, para que ela se sentisse como se estivesse “realmente lá”.

Pode-se dizer que o papel histórico da ciência e da tecnologia é dar às pessoas mais controle sobre seu ambiente. E pode-se pensar nas pessoas digitais quase como o apogeu lógico disso: as pessoas digitais experimentariam qualquer mundo que elas (ou o controlador de seu ambiente virtual) quisessem.

Isso pode ser uma coisa muito ruim ou boa:

Uma coisa ruim. Alguém que controlasse o ambiente virtual de uma pessoa digital poderia ter controle quase ilimitado sobre ela.

- Por esse motivo, seria importante que um mundo de pessoas digitais

Por esse motivo, seria importante que um mundo de pessoas digitais incluísse a aplicação efetiva dos direitos humanos básicos para todas as pessoas digitais. (Mais sobre essa ideia nas Perguntas Frequentes.)

Um mundo de pessoas digitais se tornaria rapidamente distópico se as pessoas digitais não tivessem seus direitos protegidos. Por exemplo, imagine se a regra fosse "Quem for dono de um servidor poderá executar o que quiser nele, incluindo cópias digitais de qualquer pessoa". Então, por exemplo, as pessoas poderiam fazer "cópias digitais" de si mesmas, nas quais realizariam experimentos, as a trabalhar e até mesmo, com pessoas digitais de código aberto, qualquer pessoa que executasse um servidor pudesse fazer cópias e abusar delas. **Este conto muito curto** (recomendado, mas arrepiante) dá uma ideia de como isso poderia ser.

Uma coisa boa. Por outro lado, se uma pessoa digital estivesse no controle de seu próprio ambiente (ou outra pessoa estivesse e cuidasse dela), ela poderia se livrar de quaisquer experiências das quais desejasse se livrar, incluindo fome, violência, doenças, outros problemas de saúde, e dor debilitante de qualquer tipo. Em termos gerais, elas poderiam ser "livres de necessidades materiais" - fora da necessidade de recursos de computação para serem executadas.

- Esta é uma grande mudança em relação ao mundo de hoje. Hoje, se você tiver câncer, sofrerá de dor e debilitação, mesmo que todos no mundo desejassem que você não tivesse a doença. As pessoas digitais não precisariam ter câncer se elas e outras pessoas não quisessem que isso acontecesse.
- Em particular, a coerção física num ambiente virtual poderia ser impossibilitada (poderia ser simplesmente impossível transmitir sinais para outra pessoa digital correspondendo, por exemplo, a um soco ou a um tiro).
- As pessoas digitais também poderiam experimentar muitas coisas que não podemos experimentar atualmente. Elas poderiam habitar o corpo de outra pessoa, viajar para o espaço sideral, estar em uma situação "perigosa" sem realmente se colocar em perigo, comer sem se preocupar com as consequências para a saúde, mudar de uma raça ou gênero aparente para outro, etc.

Expansão espacial

Se as pessoas digitais passassem por uma explosão de crescimento econômico como discutido acima, isso poderia vir em conjunto com uma explosão na *população* de pessoas digitais (por razões discutidas em O Duplicador).

Poder-se-ia chegar a um ponto em que elas precisariam construir naves espaciais e deixar o sistema solar para obter energia, metais, etc., suficientes para construir mais computadores e permitir a existência de mais vidas.

Colonizar o espaço poderia ser muito mais fácil para pessoas digitais do que para humanos biológicos. Elas poderiam existir em qualquer lugar onde computadores pudessem funcionar, e os ingredientes básicos necessários para fazer isso — matérias-primas, energia e "imóveis"[46] — fossem superabundantes em toda a galáxia, não apenas na Terra. Por causa disso, a população de pessoas digitais poderia acabar se tornando incrivelmente grande.[47]

Aprisionamento/Lock-in

No mundo de hoje, estamos acostumados com a ideia de que o futuro é imprevisível e incontrolável. Regimes políticos, ideologias e culturas vêm e vão (e evoluem). Alguns são bons e outros são ruins, mas geralmente não parece que nada durará para sempre. Mas comunidades, cidades e nações de pessoas digitais poderiam ser muito mais estáveis.

Primeiro, porque as pessoas digitais não precisariam morrer ou envelhecer fisicamente, e seu ambiente não precisaria se deteriorar ou se exaurir. Contanto que elas pudessem manter seu servidor funcionando, tudo em seu ambiente virtual seria fisicamente capaz de permanecer como está.

Em segundo lugar, porque um ambiente poderia ser projetado para *reforçar* a estabilidade. Por exemplo, imagine se:

- Uma comunidade de pessoas digitais formasse seu próprio governo (isso exigiria dominar ou obter o consentimento de seu governo original).
- O governo se tornasse autoritário e revogasse as proteções básicas dos direitos humanos discutidas nas Perguntas Frequentes.
- Os líderes desejassem ter certeza de que eles — ou talvez sua ideologia — permaneceriam no poder para sempre.
- Eles poderiam revisar o ambiente virtual onde residem, juntamente com todos os outros cidadãos. Através do acesso e reprogramação do código-fonte, ou da operação de robôs que alterem fisicamente o servidor, poderiam garantir que características específicas do ambiente, como a identidade dos governantes, jamais fossem alteradas. Caso uma mudança como essa estivesse prestes a ocorrer, o ambiente virtual poderia simplesmente bloquear a ação ou restaurar um estado anterior.
- Apesar das salvaguardas, o ambiente virtual permaneceria vulnerável a alterações externas, como a destruição do servidor. Contudo, após um longo período de expansão populacional e colonização espacial, o servidor poderia estar tão distante que tais ações se tornariam impraticáveis.

Alternativamente, a "correção digital" poderia ser uma força para o bem se usada com sabedoria. Ela poderia ser usada para garantir que nenhum ditador chegasse ao poder ou que certos direitos humanos básicos fossem sempre protegidos. Se uma civilização se tornasse "madura" o suficiente – por exemplo, justa, equitativa e próspera, com um compromisso com a liberdade e a autodeterminação e uma população universalmente próspera – ela poderia manter essas propriedades por muito tempo.

Não estou ciente de muitas análises aprofundadas da ideia de "aprisionamento/lock-in", mas elaboro essa ideia um pouco mais aqui. (Além disso, **aqui estão algumas observações informais** do físico **Jess Riedel**.)

Esses impactos seriam uma coisa boa ou ruim?

Ao longo deste artigo, imagino que muitos leitores tenham pensado "Isto parece terrível! O autor acha que isso seria bom?" Ou "Parece ótimo! O autor discorda disto?"

Minha opinião sobre um futuro com pessoas digitais é que ele **poderia ser muito bom ou muito ruim, e a sua configuração inicial, seria o que determinaria irreversivelmente essa questão.**

- O apressado aprisionamento/lock-in (como discutido aqui) e/ou a expansão excessivamente rápida pela galáxia (discutido aqui) resultariam em um mundo cheio de pessoas digitais (tão conscientes quanto nós) que seria altamente disfuncional, distópico ou pelo menos aquém de seu potencial.
- Mas condições iniciais aceitavelmente boas (protegendo os direitos humanos básicos para pessoas digitais, no mínimo), além de muita paciência, acúmulo de sabedoria e autoconsciência que não temos atualmente (talvez facilitadas por uma ciência social melhor), levaria a um mundo grande, estável e muito melhor. Provavelmente, seria possível eliminar doenças, pobreza material e violência, e criar uma sociedade muito melhor do que a que temos atualmente.

Notas

[38]Veja o capítulo 6 de *Age of Em (A era de Em)*, começando com “Em relação ao cálculo…”

[39]Por exemplo, quando várias equipes de pessoas digitais precisassem coordenar um projeto, elas poderiam acelerar (ou desacelerar) etapas e equipes específicas para garantir que cada parte do projeto fosse concluída no prazo. Isso permitiria que planos mais complexos e “frágeis” fossem bem-sucedidos. (Este argumento é de *Age of Em (A era de Em)* Capítulo 17, seção “Preparação”.)

[40]Veja *Age of Em (A era de Em)* Capítulo 11, seção “Aposentadoria”.

[41]Veja a nota de fim (**2**).

[42]É discutível se o mundo está ficando um pouco melhor nessas coisas, um pouco pior ou nenhuma das duas coisas. Mas parece bastante claro que o progresso não tem sido tão impressionante quanto na computação.

[43]Por que as cópias cooperariam no experimento? Talvez porque elas simplesmente concordariam com o objetivo do experimento (eu cooperaria certamente com uma cópia de mim mesmo tentando aprender sobre meditação!). Talvez porque seriam pagas (na forma de uma bela aposentadoria após o experimento). Ou porque considerariam a si mesmas e suas cópias (e/ou originais) como a mesma pessoa (ou pelo menos se importassem muito com essas pessoas muito parecidas entre si). Alguns fatores que facilitariam esse tipo de experimentação: (a) as pessoas digitais poderiam examinar seu próprio estado de espírito para ter uma noção das chances de cooperação (já que a cópia teria o mesmo estado de espírito); (b) se apenas um pequeno número de pessoas digitais participasse de experimentos, inúmeras pessoas ainda poderiam aprender com os resultados.

[44]Eu também suporia que elas conseguiriam tentar coisas mais radicais. Por exemplo, no mundo de hoje, é improvável que você possa realizar um experimento aleatório sobre o que aconteceria se as pessoas que vivem atualmente em Nova York simplesmente decidissem se mudar para Chicago. Seria muito difícil encontrar pessoas dispostas a serem designadas aleatoriamente para ficar em Nova York ou se mudar para Chicago. Mas em um mundo de pessoas digitais, os experimentadores poderiam pagar aos nova-iorquinos para fazerem cópias de si mesmos que se mudariam para Chicago. E depois do experimento, cada cópia de Chicago que desejasse ficar em Nova York poderia escolher substituir-se por outra cópia da versão de Nova York. (Este último argumento levanta questões sobre a filosofia da identidade pessoal, mas para fins de ciências sociais, tudo o que importa é que algumas pessoas ficariam felizes em participar de experimentos devido a essa opção, e todos poderiam aprender com os experimentos.)

[45]Veja a nota de rodapé do primeiro argumento sobre porque as cópias das pessoas poderiam cooperar com elas.

[46]E ar para refrigeração.

[47]Veja estimativas em ***Astronomical Waste* (Resíduos astronômicos)**para uma noção aproximada de quão grande os números podem chegar a ser aqui (embora essas estimativas sejam extremamente especulativas).

Isto não pode continuar

Este artigo começa a defender que vivemos em um século notável, não apenas em uma era notável. Artigos anteriores nesta série falaram sobre o futuro estranho que eventualmente poderia estar à nossa frente (talvez daqui a 100 anos, talvez 100.000).

Resumo deste artigo:

- Estamos acostumados com uma economia mundial que cresce alguns por cento ao ano. Tem sido assim por muitas gerações.
- Entretanto, esta é uma situação muito incomum. Quando distanciamos a visão para abranger toda a história, vemos que o crescimento vem acelerando; que está perto de seu ápice histórico; e não poderá manter essa velocidade por muito mais tempo (não há átomos suficientes na galáxia para sustentar essa taxa de crescimento por mais 10.000 anos).
- O mundo não pode simplesmente continuar crescendo nesse ritmo indefinidamente. Devemos estar preparados para outras possibilidades: estagnação (o crescimento desacelera ou termina), explosão (o crescimento acelera ainda mais, antes de atingir seus limites) e colapso (alguns desastres destroem a economia).

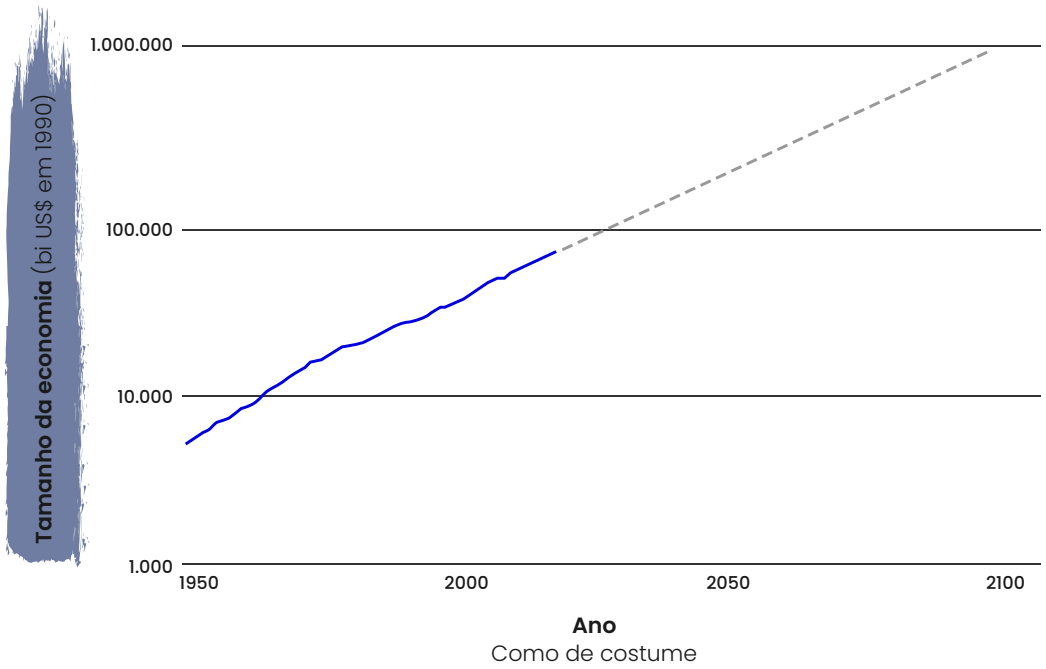
Os tempos em que vivemos são inusitados e instáveis. Não devemos nos surpreender se algo estranho acontecer, como uma explosão no progresso econômico e científico, levando à maturidade tecnológica. Na verdade, tal explosão estaria, de certa forma, seguindo tendências atuais.

Desde que qualquer um de nós consegue se lembrar, a economia mundial cresceu[48] alguns por cento ao ano, em média. Alguns anos apresentam mais ou menos crescimento do que outros anos, mas o crescimento tem sido, em geral, bastante estável.[49] Chamarei isso de mundo **"como de costume"**.

No mundo "como de costume", o mundo está constantemente mudando, e essa mudança é perceptível, mas ela não é avassaladora ou impossível de se acompanhar. Há um fluxo constante de novas oportunidades e novos desafios, mas se você quiser demorar alguns anos para se adaptar a essas mudanças enquanto continua fazendo as coisas do mesmo jeito de sempre, geralmente (pessoalmente), você não terá problemas com isso.

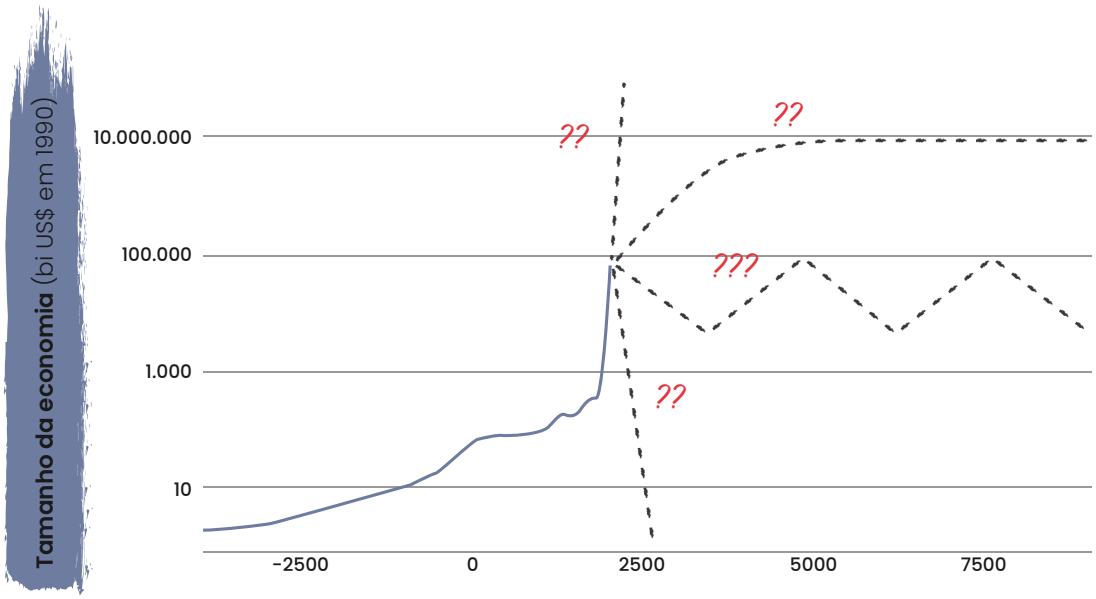
Em termos de vida cotidiana, 2019 foi bastante semelhante a 2018, notavelmente, mas não enormemente diferente de 2010, e enormemente, mas não absurdamente diferente de 1980.[50]

Se isso lhe parece certo, e você está acostumado com isso, e também imagina o futuro como sendo assim, então você vive na mentalidade "como de costume". Quando você pensa sobre o passado e o futuro, provavelmente está pensando em algo mais ou menos assim:



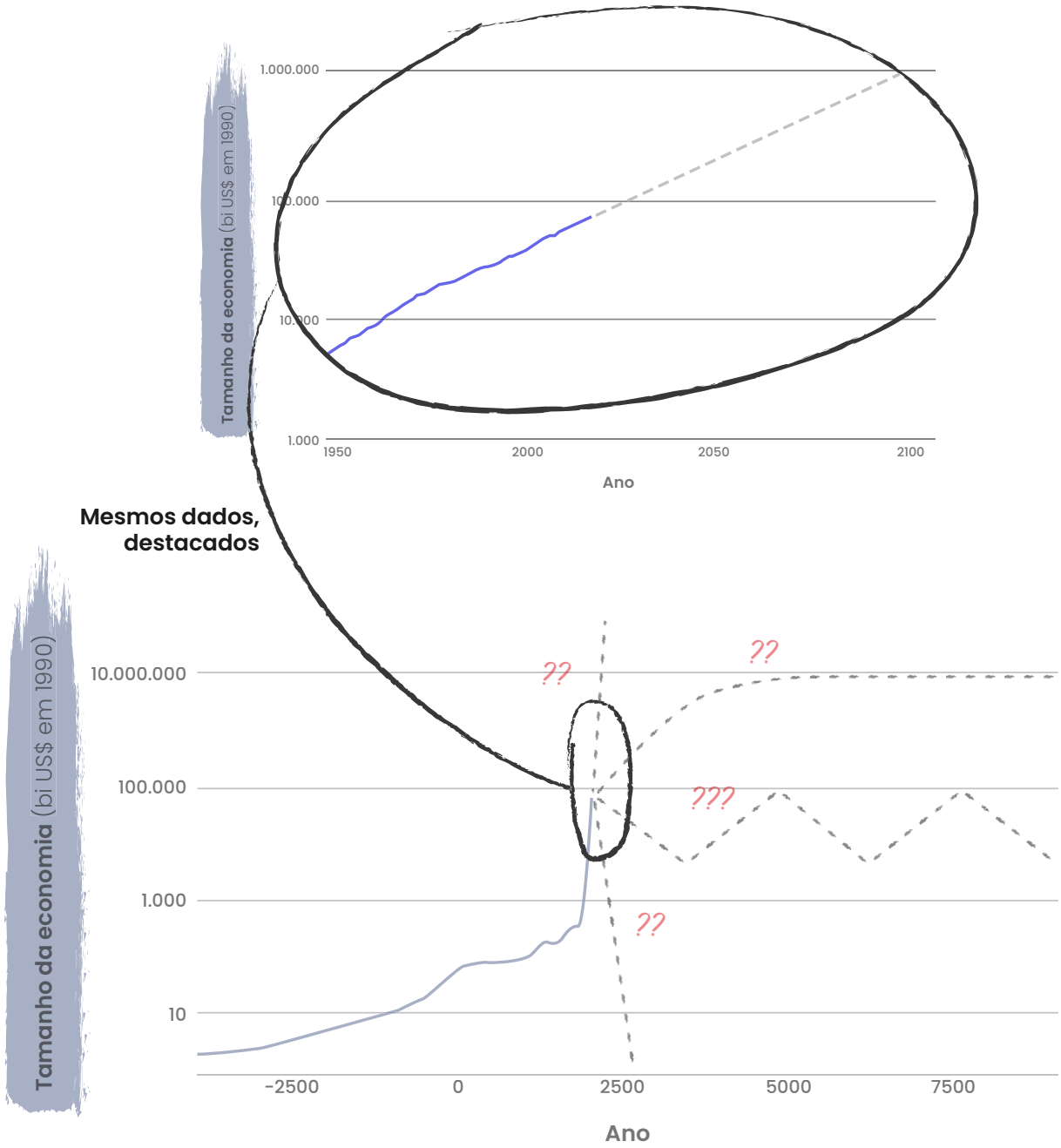
Como de costume

Vivo em uma mentalidade diferente, com um passado mais turbulento e um futuro mais incerto. Chamarei isso de mentalidade **"Isto não pode continuar".** Aqui está a minha versão do gráfico:



Isto não pode continuar

Qual gráfico é o correto? Bem, ambos usam os mesmos dados históricos — só que o gráfico que representa a mentalidade "como de costume" começa em 1950, enquanto o que representa a mentalidade do tipo "isto não pode continuar" começa em 5000 AEC. **O "Isto não pode continuar" é a história completa; o "Como de costume" é uma pequena fatia dela.**



Ter uma economia que cresce alguns por cento ao ano é o que estamos acostumados. Mas, em pleno contexto histórico, crescer alguns por cento ao ano é uma loucura. (É a parte onde a linha azul fica quase vertical.)

Esse crescimento tem durado por mais tempo do que qualquer um de nós consegue se lembrar, mas isso não é muito tempo no final das contas — são apenas algumas centenas de anos, em milhares de anos de civilização humana. É uma aceleração enorme e não pode continuar por muito mais tempo. (Detalharei o que quero dizer com "não pode continuar por muito mais tempo" abaixo.)

O primeiro gráfico sugere regularidade e previsibilidade. O segundo sugere volatilidade e possibilidades futuras dramaticamente diferentes.

Um futuro possível é a **estagnação**: a economia atingirá seu "tamanho máximo" e o crescimento praticamente parará. Estaremos todos preocupados em como dividir os recursos que ainda tivermos, e os dias de uma fartura crescente e uma economia dinâmica terminarão para sempre.

Outra possibilidade é a **explosão**: o crescimento acelerará ainda mais, a ponto de a economia mundial dobrar a cada ano, semana ou hora. Uma tecnologia do tipo do Duplicador (tal como pessoas digitais ou, como discutirei em artigos futuros, Inteligência Artificial avançada) poderia impulsionar um crescimento como esse. Se isso acontecer, tudo mudará muito mais rápido do que os humanos conseguirão processar.

Outro futuro possível é o **colapso**: uma catástrofe global colocará a civilização de joelhos ou acabará com a humanidade completamente, e nunca mais atingiremos o nível atual de crescimento.

Ou talvez outra coisa aconteça.

Por que isto não pode continuar?

Um bom ponto de partida seria **esta análise do blog Overcoming Bias (Superando o viés)**, da qual darei minha própria versão aqui:

- Digamos que a economia mundial esteja crescendo 2% a cada ano.[52] Isso implicaria que a economia estaria dobrando de tamanho a cada 35 anos.[53]
- Se isso se mantiver, daqui a 8.200 anos, a economia terá cerca de $3x10^{70}$ vezes seu tamanho atual.
- Provavelmente há menos de 10^{70} átomos em nossa galáxia,[54] dos quais não poderíamos explorar completamente no prazo de 8200 anos.[55]
- Então, se a economia fosse 3×10^{70} vezes maior que ela é hoje, e só pudesse usar 10^{70} **(ou menos) átomos, precisaríamos sustentar várias economias tão grandes quanto a economia mundial atual inteira *por átomo***

Oito mil e duzentos anos pode soar como bastante tempo, mas é muito menos tempo do que a existência humana. Na verdade, é menos tempo do que a idade da civilização humana (baseada na agricultura).

É plausível que conseguiríamos desenvolver uma tecnologia capaz de sustentar o equivalente ao tamanho de várias civilizações atuais, por átomo disponível? Claro — mas isso exigiria um grau radical de transformação de nossas vidas e sociedades, muito além das mudanças que tivemos ao longo da história humana até hoje. E eu não *apostaria* exatamente que é assim que as coisas vão acontecer nos próximos milhares de anos. (Atualização: para as pessoas que ainda não estão convencidas, expandi esse argumento em outra postagem).

Parece muito provável que "esgotemos" novos *insights* científicos, inovações tecnológicas e recursos, e o regime de "ficar mais rico alguns por cento ao ano" termine. Afinal, esse regime tem apenas algumas centenas de anos. **Esta postagem** faz uma análise semelhante focando na energia em vez da economia. Ela prevê que os limites chegarão mais cedo ainda. Ela presume um crescimento anual de 2,3% no consumo de energia (menos que a taxa histórica dos EUA desde 1600) e estima que isso consumiria tanta energia quanto a produzida por todas as estrelas em nossa galáxia em 2.500 anos.[56]

Explosão e colapso

Então, um futuro possível é a estagnação: o crescimento desacelera gradualmente com o tempo e, eventualmente, temos uma economia sem crescimento. Mas não acredito que este seja o futuro mais provável.

O gráfico acima não mostra o crescimento desacelerando — ele mostra uma aceleração dramática. O que esperaríamos ver se simplesmente projetássemos essa mesma aceleração para a frente?

***Modeling the Human Trajectory* (Modelando a trajetória humana)** (de David Roodman, da *Open Philanthropy*) tenta responder exatamente a essa pergunta, "ajustando uma curva" ao padrão de crescimento econômico passado.[57] Sua extrapolação implica *crescimento infinito* neste século. O crescimento infinito é uma abstração matemática, mas você pode entendê-lo da seguinte forma: "O crescimento acelerará o máximo possível antes de atingir seus limites."

Em O Duplicador, eu resumo uma discussão mais ampla sobre essa possibilidade. A conclusão é que uma explosão de crescimento deveria ser possível, *se* tivéssemos a tecnologia para "copiar" mentes humanas —ou outra coisa que cumprisse o mesmo propósito efetivo, como pessoas digitais ou uma Inteligência Artificial avançada o suficiente.

Em uma explosão de crescimento, a taxa de crescimento anual poderia atingir 100% (com a economia mundial dobrando de tamanho a cada ano) — o que duraria no máximo 250 anos aproximadamente, antes de atingirmos os tipos de limites discutidos acima.[58] Ou poderíamos ter um crescimento ainda mais rápido — com a economia mundial dobrando de tamanho a cada mês (o que conseguiríamos sustentar por no máximo 20 anos antes de atingir os limites),[59] ou mais rápido do que isso. Isso seria uma grande aventura: crescimento incrivelmente rápido, talvez impulsionado por IAs produzindo resultados além do que nós, humanos, poderíamos rastrear significativamente, aproximando-nos rapidamente dos limites do que é possível, ponto onde o crescimento teria que desacelerar.

Além da estagnação ou do crescimento explosivo, existe uma terceira possibilidade: o colapso. Uma catástrofe global poderia reduzir a civilização a um estado no qual ela nunca recuperaria o seu nível atual de crescimento. A extinção humana seria uma versão extrema de tal colapso. Esse futuro não é sugerido pelos gráficos, mas sabemos que ele é possível.

Como argumenta ***The Precipice* (O precipício)**, de Toby Ord, os asteroides e outros riscos naturais não parecem prováveis de causar isso, mas há alguns riscos que parecem sérios e muito difíceis de quantificar: mudança climática, guerra nuclear (particularmente o inverno nuclear), pandemias (especialmente se os avanços na biologia levarem ao desenvolvimento de armas biológicas terríveis) e riscos decorrentes da Inteligência Artificial avançada.

Com estas três possibilidades em mente (estagnação, explosão e colapso):

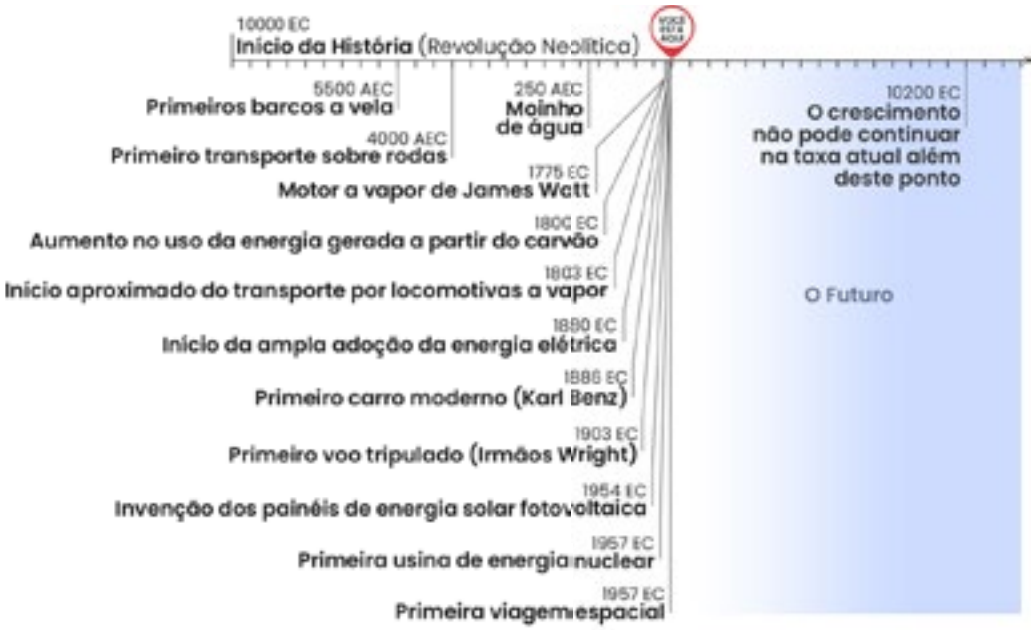
- Vivemos em um dos (dois) séculos de crescimento mais rápido de toda a história até agora. (XX e o XXI).
- Parece provável que este seja pelo menos, aproximadamente, um dos 80 séculos de crescimento mais rápido de todos os tempos.[60]
- Se a tecnologia certa surgir e impulsionar um crescimento explosivo, esse pode ser, de longe, o século de crescimento mais rápido de todos os tempos.
- Se as coisas correrem mal o suficiente, ele poderia ser o nosso último século.

Então parece que este é um século bastante notável, com alguma probabilidade de ser o mais notável. Tudo isso é baseado em observações bastante básicas, não em raciocínio detalhado sobre Inteligência Artificial (que abordarei em artigos futuros).

Avanço científico e tecnológico

É difícil construir um gráfico simples do quão rápido a ciência e a tecnologia estão avançando, da mesma forma que podemos construir um gráfico de crescimento econômico. Mas acredito que, se pudéssemos, apresentaríamos um cenário bastante semelhante ao gráfico de crescimento econômico.

Um livro divertido que recomendo é ***Chronology of Science and Discovery* (Cronologia da Ciência e Descobrimento) de Asimov** . Ele passa, em ordem cronológica, pelas invenções e descobertas mais importantes da história. Os primeiros verbetes incluem "ferramentas de pedra", "fogo", "religião" e "arte"; as páginas finais incluem "o cometa de Halley" e "os supercondutores de alta temperatura". Um fato interessante sobre este livro é que **553 de suas 654 páginas são sobre descobertas que ocorreram após o ano de 1500** —embora o livro comece no ano 4 milhões AEC. Prevejo que outros livros desse tipo mostrarão um padrão semelhante,[61] e acredito que houve, de fato, mais avanços científicos e tecnológicos nos últimos 500 anos, aproximadamente, do que nos vários milhões de anos anteriores.[62]



Em um artigo anterior, argumentei que os eventos mais significativos da história parecem estar agrupados em torno do tempo em que vivemos, ilustrado com esta cronologia. Ela foi construída a partir da análise de períodos de bilhões de anos. Se ampliarmos para milhares de anos, porém, veremos algo semelhante: os maiores avanços científicos e tecnológicos estão agrupados muito próximos no tempo até os dias atuais. Para ilustrar isso, aqui está uma cronologia focada em transporte e energia (acho que poderia ter escolhido qualquer categoria e obtido um cenário semelhante).

Assim como com o crescimento econômico, a taxa de avanço científico e tecnológico é extremamente rápida em comparação com a maior parte da história. Tal como acontece com o crescimento econômico, presumivelmente há limites em algum momento para o quanto a tecnologia avançar. E, assim como no crescimento econômico, daqui em diante o avanço científico e tecnológico poderia:

- **Estagnar**, como **alguns já estão preocupados de que esteja acontecendo**.
- **Explodir**, se alguma tecnologia que aumentasse drasticamente o número de "mentes" (pessoas ou pessoas digitais, ou IAs avançadas) fosse desenvolvida, impulsionando o desenvolvimento científico e tecnológico.[63]
- **Colapsar** devido a alguma catástrofe global.

Possibilidades negligenciadas

Acho que deve haver algumas pessoas no mundo que possuem a mentalidade “como de costume”, que pensam em como melhorar o mundo se assumirmos basicamente uma taxa de crescimento econômico estável e regular em um futuro previsível.

E algumas pessoas devem possuir a mentalidade “isto não pode continuar”, que pensam sobre as ramificações de estagnação, explosão ou colapso — e se nossas ações poderiam mudar qual desses cenários aconteceria.

Mas atualmente parece que as coisas estão muito desequilibradas, com quase todas as notícias e análises oriundas da mentalidade “como de costume”.

Uma metáfora para o meu estado de espírito é que parece que o mundo é um conjunto de pessoas num avião em alta velocidade na pista de decolagem:

Estamos indo muito mais rápido que o normal e não há pista suficiente para fazer isso por muito mais tempo... e estamos acelerando.

E toda vez que leio comentários sobre o que está acontecendo no mundo, as pessoas estão discutindo como colocar o cinto de segurança da maneira mais confortável possível, já que usá-lo faz parte da vida. Ou dizendo que os melhores momentos da vida são sentar-se com sua família e assistir às linhas brancas passando do lado de fora da janela, ou discutindo sobre de quem é a culpa de haver um ruído de fundo dificultando que elas ouçam umas às outras.

Se eu estivesse nessa situação e não soubesse o que iria acontecer depois (a decolagem), não necessariamente acertaria, mas espero que pelo menos estaria pensando: “Essa situação parece meio louca, incomum e temporária. Ou aceleraremos ainda mais, ou pararemos, ou alguma outra coisa estranha acontecerá”

Agradeço a María Gutiérrez Rojas pelos gráficos deste artigo e a Ludwig Schubert por um gráfico de cronologia anterior no qual o gráfico deste artigo se baseou.

Notas

[51]Para dados históricos, veja ***Modeling the Human Trajectory* (Modelando a trajetória humana)**. As projeções são grosseiras e destinadas a serem visualmente sugestivas, em vez de usar as melhores abordagens de modelagem.

[52]Isso se refere ao crescimento real do PIB (ajustado pela inflação). 2% é menor do que o número atual de crescimento mundial, e usar o número de crescimento mundial tornaria meu argumento mais forte. Mas acredito que 2% é um palpite decente para "crescimento de fronteira" — crescimento que ocorre nas economias mais desenvolvidas — em oposição ao crescimento mundial total, que inclui "crescimento das economias emergentes" (países anteriormente pobres crescendo rapidamente, como a China hoje). Para verificar meu palpite de 2%, baixei **esses dados dos EUA** e analisei a taxa de crescimento anualizada entre 2000–2020, 2010–2020 e 2015–2020 (todos usando julho, já que julho foi o último ponto de 2020). Estes foram 2,5%, 2,2% e 2,05%, respectivamente.

[53]2% de crescimento ao longo de 35 anos é $(1 + 2\%)^{35} = 2x$ o crescimento

[54]A estimativa mais alta listada na **Wikipedia** para a massa da Via-Láctea é de $4,5x10^{12}$ massas solares, cada uma com cerca de $2x10^{30}$ kg. A massa de um átomo (de hidrogênio) é estimada como equivalente a cerca de $1,67x10^{-27}$ kg. (Os átomos de hidrogênio têm a massa mais baixa, portanto, presumir que cada átomo é hidrogênio superestimará o número total de átomos.) Portanto, uma estimativa alta do número total de átomos na Via-Láctea seria $(4,5x10^{12} \times 2x10^{30})/(1,67x10^{-27}) =\sim 5,4x10^{69}$

[55]**Wikipedia**: "Em março de 2019, astrônomos relataram que a massa da Via-Láctea é de 1,5 trilhão de massas solares em um raio de cerca de 129.000 anos-luz." Estou presumindo que não podemos viajar mais de 129.000 anos-luz nos próximos 8.200 anos, porque isso exigiria viagens muito mais rápidas que a velocidade da luz.

[56]Este cálculo não é apresentado diretamente na postagem. Os trechos principais são "Não importa qual seja a tecnologia, uma taxa de crescimento de energia sustentada de 2,3% exigiria que produzíssemos tanta energia quanto o sol inteiro em 1400 anos" e "A Via-Láctea hospeda cerca de 100 bilhões de estrelas. Muita energia apenas se derramando no espaço, lá para ser aproveitada. Lembre-se de que cada fator de dez nos leva 100 anos.

[57]Há um **debate em aberto** se ***Modeling the Human Trajectory* (Modelando a trajetória humana)** está ajustando o tipo certo de forma aos dados históricos do passado. Discuto como o debate poderia mudar minhas conclusões aqui.

[58]Duzentas e cinquenta duplicações seria um fator de crescimento de cerca de $1,8\times10^{75}$, mais de 10.000 vezes o número de átomos em nossa galáxia.

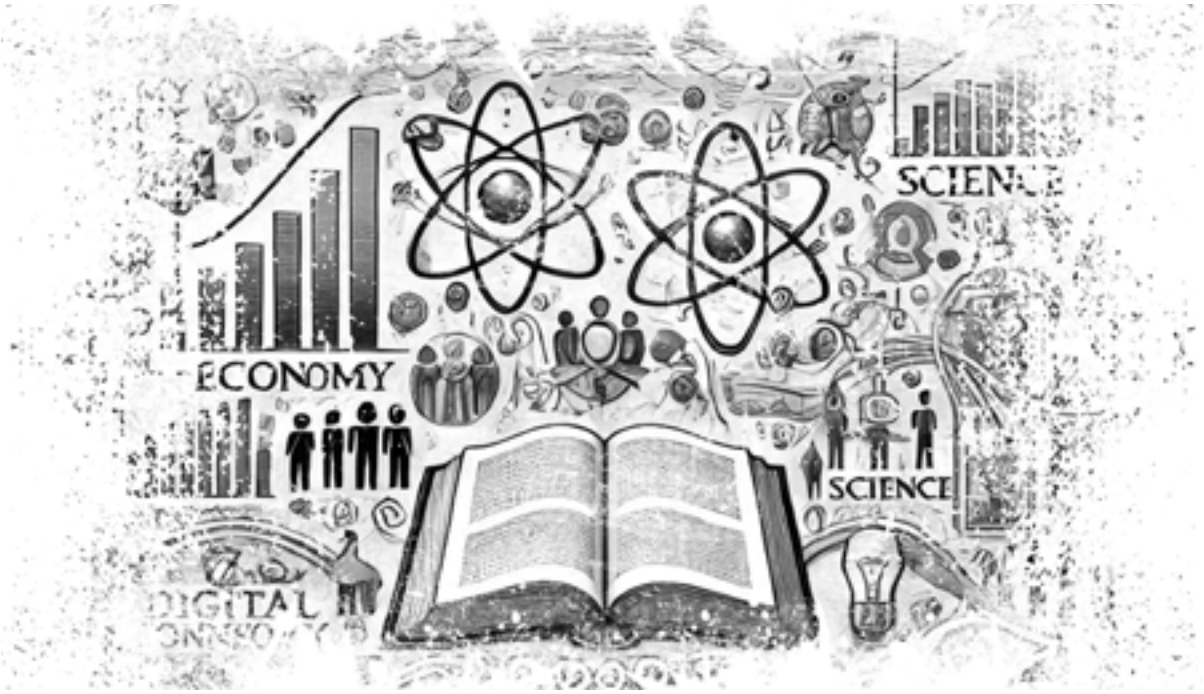
[59]Vinte anos seriam 240 meses, então se cada um visse uma duplicação na economia mundial, isso seria um fator de crescimento de cerca de $1,8x10^{72}$, mais de 100 vezes o número de átomos em nossa galáxia.

[60]Isso se deve à observação acima de que a taxa de crescimento atual não pode durar mais do que outros 8.200 anos (82 séculos) ou mais. Portanto, a única maneira de termos mais de 82 séculos com crescimento igual ao de hoje é se também tivermos muitos outros séculos com crescimento negativo, como a linha pontilhada em zigue-zague no gráfico "Isto não pode continuar".

[61]**Este conjunto de dados** atribui importância a figuras históricas com base no quanto elas são abordadas em obras de referência. Ele tem mais de 10 vezes mais verbetes sobre "Ciência" depois do ano 1500 do que antes dele; o conjunto de dados começa em 800 AEC. Não endosso o livro do qual este conjunto de dados foi tirado, pois acredito que ele tira muitas conclusões injustificadas dos dados; aqui estou simplesmente sustentando minha afirmação de que a maioria das obras de referência cobre desproporcionalmente os anos após 1500.

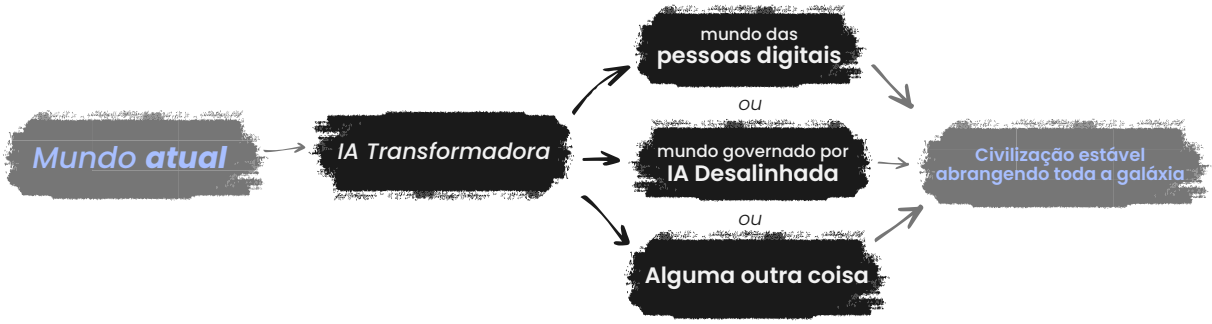
[62]Para ser justo, trabalhos de referência como este podem ser tendenciosos em relação ao passado recente. Mas acredito que a impressão geral que eles dão sobre esse tema é precisa, mesmo assim. Embasar realmente esta afirmação estaria além do escopo desta postagem, mas a evidência que eu apontaria é (a) as obras que faço referência — acredito que se você ler ou passar os olhos rapidamente nesses trabalhos, provavelmente terá uma impressão semelhante; (b) que o crescimento econômico mostra um padrão semelhante (embora a explosão tenha começado mais recentemente; acredito que faz sentido intuitivo que o crescimento econômico siga o progresso científico com um atraso).

[63]Os trabalhos citados em O duplicador sobre este argumento específico modelam uma explosão em inovação como sendo parte da dinâmica impulsionadora do crescimento econômico explosivo.



Prevendo a IA transformadora, parte 1: Qual tipo de IA?

PASTA: Processo para automação do avanço científico e tecnológico



Esta é a primeira de quatro postagens resumindo centenas de páginas de relatórios técnicos focados quase inteiramente na previsão de um número. Provavelmente, este é o único número para o qual eu mais valorizo ter uma boa estimativa: **o ano em que a Inteligência Artificial Transformadora será desenvolvida.**[64]

Por "IA Transformadora", o que quero dizer é "IA poderosa o suficiente para nos levar a um futuro novo e qualitativamente diferente". A **Revolução industrial** é o exemplo mais recente de um evento transformador; outros incluiriam a Revolução Agrícola e o surgimento da humanidade.[65]

Este artigo vai se concentrar em analisar um tipo particular de Inteligência Artificial que acredito que poderia ser transformadora: **sistemas de Inteligência Artificial que conseguem automatizar essencialmente todas as atividades humanas necessárias para acelerar o avanço científico e tecnológico.** Chamarei esse tipo de tecnologia de Processo para Automatizar o Avanço Científico e Tecnológico, ou **PASTA.**[66] (Quero dizer que PASTA se refere a um único sistema ou a uma coleção de sistemas que podem fazer coletivamente esse tipo de automação.)

O PASTA poderia resolver o mesmo tipo de gargalo discutido em O Duplicador e Isto não pode continuar — a **escassez de mentes humanas (ou algo que desempenhe o mesmo papel na inovação).**

O PASTA poderia, portanto, acarretar uma ciência explosiva, culminando em tecnologias tão impactantes quanto pessoas digitais. E dependendo dos detalhes, os sistemas PASTA poderiam ter objetivos próprios, o que poderia ser **perigoso para a humanidade** e isso importaria muito para que tipo de civilização acabaria se expandindo pela galáxia.

Ao falar sobre o PASTA, estou tentando, em parte, me livrar de alguma bagagem desnecessária no debate sobre "inteligência artificial geral". Não acredito que precisemos da inteligência artificial geral para que este século seja o mais importante da história. Algo mais restrito — como o PASTA poderia ser, ou seria, suficiente para isso.

Para tornar essa ideia um pouco mais concreta, o restante desta postagem discutirá:

- Como o PASTA poderia (hipoteticamente) ser desenvolvido por meio de métodos de aprendizado de máquina mais ou menos modernos.
- Porque isso poderia acarretar um progresso científico e tecnológico explosivo — e, porque isso poderia ser perigoso por meio de sistemas PASTA com objetivos próprios.

Artigos futuros discutirão daqui a quanto tempo podemos supor que algo como o PASTA seja desenvolvido.

Fazendo o PASTA

Começarei com uma caracterização muito breve e simplificada do aprendizado de máquina, que você pode pular clicando aqui.

Existem essencialmente duas maneiras de "ensinar" um computador a realizar uma tarefa: **Programação tradicional.** Nesse caso, você programa instruções passo a passo e extremamente específicas para concluir uma tarefa. Por exemplo, o programa de jogo de xadrez ***Deep Blue*** executa, essencialmente, instruções[67] desse tipo:

- Receba uma representação digital de um tabuleiro de xadrez, com números indicando qual peça de xadrez está em cada casa; (b) quais lances seriam legítimos; (c) quais posições do tabuleiro contariam como xeque-mate.
- Verifique como cada movimento legal modificaria o tabuleiro. Em seguida, verifique o quão "bom" é o tabuleiro resultante, de acordo com regras como: "Se a rainha do outro jogador foi capturada, isso vale 9 pontos; se a rainha do *Deep Blue* foi capturada, isso vale 9 pontos." Essas regras podem ser bastante complexas,[68] mas todas foram programadas precisamente por humanos.

Aprendizagem de Máquina. Isso é essencialmente "treinar" uma Inteligência Artificial para executar uma tarefa por tentativa e erro, em vez de fornecer instruções específicas. Hoje, a maneira mais comum de fazer isso é usando uma "rede neural artificial" (RNA), que você pode pensar como sendo um "cérebro digital" que inicia em um estado vazio (ou aleatório): que ainda não foi programado para fazer coisas específicas.

Por exemplo, o ***AlphaZero*** — uma Inteligência Artificial usada para dominar vários jogos de tabuleiro, incluindo xadrez e *Go* — faz algo mais parecido com isso (embora também tenha elementos importantes de "programação tradicional", que estou ignorando aqui visando simplificar as coisas):

- A Inteligência Artificial joga uma partida de xadrez contra si mesma (escolhendo um movimento legal, modificando o tabuleiro do jogo digital adequadamente e, em seguida, escolhendo outro movimento legítimo, etc.) Inicialmente, ela joga fazendo movimentos aleatórios.

- Cada vez que as Brancas vencem, ela "aprende" um pouco, ajustando a configuração da RNA ("cérebro digital") — fortalecendo ou enfraquecendo as conexões entre alguns "neurônios artificiais" e outros. Os ajustes fazem com que a RNA forme uma associação mais forte entre os estados do jogo, como, por exemplo, o que ela acabou de perceber e "As brancas vão vencer". E vice-versa quando as Pretas vencem.

- Depois de um número muito grande de jogos, a RNA se tornou eficaz em determinar — a partir de um estado de jogo de tabuleiro digital — qual lado provavelmente vencerá. A RNA agora pode selecionar movimentos que tornam seu próprio lado o mais propenso a vencer.

- O processo de "treinamento" da RNA exige muita tentativa e erro: inicialmente ela é péssima no xadrez e precisa jogar muitas partidas para "conectar seu cérebro corretamente" e se tornar boa. Porém, uma vez que a RNA tiver sido treinada pela primeira vez, seu "cérebro digital" será consistentemente bom no jogo de tabuleiro que aprendeu; ela poderá derrotar seus oponentes repetidamente.

A última abordagem é central para muitos dos progressos recentes na IA. Isso é especialmente verdadeiro para tarefas para as quais é difícil de "escrever todas as instruções". Por exemplo, os humanos conseguem escrever algumas diretrizes razoáveis para ter sucesso no xadrez, mas sabemos muito pouco sobre como nós mesmos classificamos imagens (determinar se alguma imagem é de um cachorro, gato ou outra coisa). Portanto, o aprendizado de máquina é particularmente essencial para tarefas como classificação de imagens.

O PASTA poderia ser desenvolvido via aprendizado de máquina? Uma maneira óbvia (mas irreal) de fazer isso pode ser algo assim:

- Em vez de jogar xadrez, uma Inteligência Artificial poderia jogar um jogo chamado "Provocar o avanço científico e tecnológico". Ou seja, ela poderia fazer "movimentos" como: baixar artigos científicos, adicionar notas a um arquivo, criar desenhos e instruções para novos experimentos, projetar processos de fabricação.

- Um painel de juízes humanos poderia assistir do "lado de fora" e dar sua avaliação subjetiva de quão rápido o trabalho da Inteligência Artificial está causando avanço científico/tecnológico. A Inteligência Artificial poderia, portanto, ajustar sua configuração ao longo do tempo, aprendendo quais tipos de movimentos causam o avanço científico e tecnológico mais efetivamente conforme a avaliação dos juízes.

Isso seria extremamente impraticável, pelo menos em comparação com o jeito que acredito que as coisas são prováveis de acontecer, mas espero dar uma intuição inicial sobre o que um processo de treinamento poderia estar tentando realizar: ao fornecer um sinal de "como a Inteligência Artificial está se saindo", isso poderia permitir que uma Inteligência Artificial atingisse seu objetivo por meio de tentativa e erro e ajustes na sua configuração interna.

Na realidade, eu suporia que o treinamento seria mais rápido e prático devido a coisas como:

- Diferentes IAs poderiam ser treinadas para desempenhar diferentes tipos de funções relacionadas à aceleração da ciência e tecnologia: escrever trabalhos acadêmicos, projetar e criticar projetos e processos de fabricação, etc. Em muitos casos, os humanos que já estão envolvidos nessas atividades poderiam gerar muitos dados sobre como realizar bem as tarefas, o que poderia ser usado para o tipo de treinamento descrito acima. Uma vez que diferentes IAs pudessem desempenhar uma variedade de funções importantes, as IAs "gerentes" poderiam ser treinadas para supervisionar e alocar o trabalho de outras IAs.

- As IAs também poderiam ser treinadas para serem *juízes.* Talvez uma Inteligência Artificial pudesse ser treinada para avaliar se um artigo contém ideias originais e outra pudesse ser treinada para avaliar se um artigo contém erros.[69] Essas IAs "juízes" poderiam então ser usadas para treinar uma terceira Inteligência Artificial com mais eficiência, aprendendo assim a escrever artigos originais e corretos.

- De maneira mais geral, as IAs poderiam aprender a fazer todos os tipos de outras atividades humanas, ganhando habilidades humanas genéricas, como a capacidade de aprender com livros didáticos e a capacidade de "fazer um brainstorming de soluções criativas para um problema". As IAs boas nessas coisas poderiam então aprender ciência a partir de livros didáticos como um ser humano normal, e fazer um brainstorming sobre como fazer uma descoberta como um ser humano normal faria, etc.

    - A distinção aqui é entre "usar inúmeros exemplos para conectar um cérebro" e "um cérebro já conectado usando alguns poucos exemplos para aprender rapidamente, como faz um cérebro humano".

    - Aqui seriam necessárias muitas tentativas e erros para que a RNA se tornasse boa em habilidades humanas "genéricas", mas depois disso a RNA treinada poderia aprender como fazer um trabalho científico específico tão eficientemente quanto um humano aprende como fazê-lo. (De certa forma, podemos imaginar que ele foi "treinado por meio de tentativa e erro massivamente *para conseguir aprender certos tipos de coisas sem precisar de tanta tentativa e erro".*)

    - Existem algumas evidências preliminares (por exemplo, **aqui**) de que os sistemas de Inteligência Artificial poderiam passar por esse padrão de "aprender 'o básico' usando uma tonelada de tentativa e erro, assim como aprender sub-habilidades usando menos tentativas e erros."[70]

- Eu particularmente não espero que tudo isso aconteça como parte de um processo de desenvolvimento único e deliberado. Com o tempo, espero que diferentes sistemas de Inteligência Artificial sejam utilizados para tarefas diferentes e cada vez mais amplas, incluindo, principalmente, tarefas que ajudem a complementar as atividades humanas no avanço científico e tecnológico. Poderia haver muitos tipos de sistemas de IA, cada um com seu próprio modelo de receita e retroalimentação, e suas habilidades coletivas poderiam crescer a ponto de, em algum momento, algum conjunto deles conseguir realizar tudo (com relação ao avanço científico e tecnológico) que anteriormente exigiria o trabalho de um ser humano. (Por conveniência, porém, às vezes me referirei a um conjunto como PASTA no singular.)

O desenvolvimento do PASTA quase certamente será muito mais difícil e mais caro do que foi o desenvolvimento do AlphaZero. Pode ser necessário muita engenhosidade para contornar os obstáculos que existem atualmente (o cenário descrito acima é certamente radicalmente simplificado e está lá para fornecer intuições básicas). Mas a pesquisa de Inteligência Artificial está ficando simultaneamente mais barata[71] e melhor financiada. Argumentarei em artigos futuros que as chances de desenvolver o PASTA nas próximas décadas são substanciais.

Impactos do PASTA

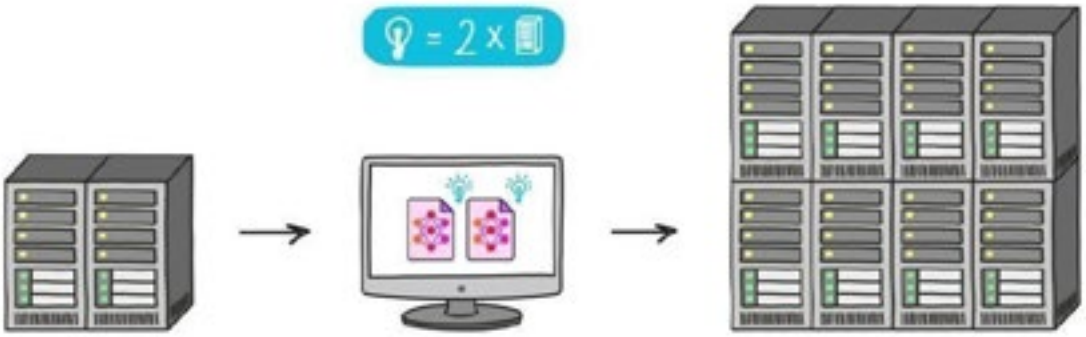
Avanço científico e tecnológico explosivo

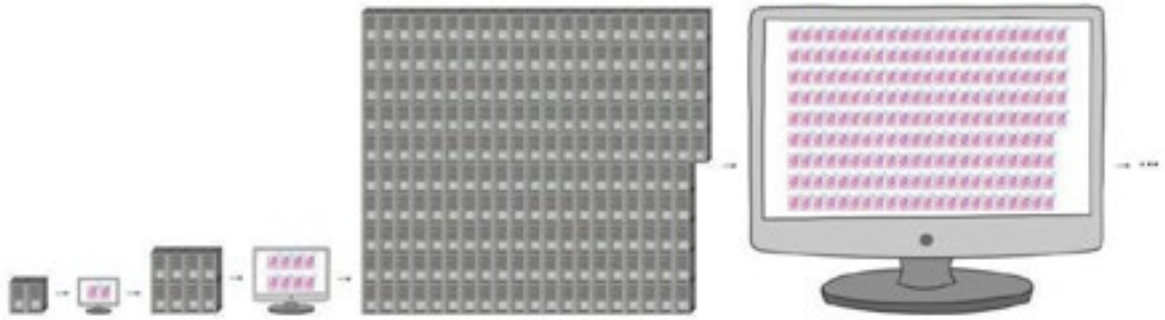
Eu já falei sobre a ideia de uma potencial explosão em avanço científico e tecnológico, que poderia levar a um futuro radicalmente desconhecido.

Enfatizei que tal explosão poderia ser causada por uma tecnologia que "aumentou drasticamente o número de 'mentes' (humanos, ou pessoas digitais, ou IAs avançadas) impulsionando o avanço científico e tecnológico".

O PASTA se encaixaria bem nessa definição, especialmente se fosse tão bom (ou melhor) quanto os humanos em encontrar maneiras melhores e mais baratas de fabricar mais sistemas PASTA. O PASTA teria todas **as ferramentas necessárias para uma explosão de produtividade que eu havia planejado anteriormente para as** pessoas digitais:

- Os sistemas PASTA poderiam fazer cópias de si mesmos, incluindo cópias temporárias, e executá-las em velocidades diferentes.
- Essas cópias poderiam se envolver no tipo de ciclo descrito em O Duplicador: "mais ideias [incluindo ideias para fazer mais/melhores sistemas PASTA] → mais pessoas [nesse caso mais sistemas PASTA] → mais ideias→…"





Obrigado a María Gutiérrez Rojas por estes gráficos., uma variação de gráficos semelhantes de O duplicador e Pessoas digitais seriam mais importantes ainda Ilustrando a dinâmica do crescimento explosivo. Aqui, em vez das pessoas terem ideias que aumentariam a produtividade, são os algoritmos de Inteligência Artificial (indicados por ícones de rede neural) que teriam essas ideias.

Por que essa retroalimentação não se aplica aos computadores e IAs atuais? Porque os computadores e IAs atuais não conseguem fazer todas as coisas necessárias para ter novas ideias e serem copiados com mais eficiência.

Eles desempenham um papel na inovação, mas a inovação acaba sendo prejudicada pelos humanos, cuja população não cresce tão rapidamente. Isso é o que o PASTA mudaria (e é também o que pessoas digitais mudariam).

Além disso: ao contrário das cópias digitais de humanos, os sistemas PASTA poderiam não estar vinculados à sua identidade e personalidade existentes. Um sistema PASTA poderia fazer rapidamente qualquer edição em sua "mente" que o tornasse mais eficaz em impulsionar a ciência e a tecnologia. Isso poderia (ou não, dependendo de muitos detalhes) levar ao **autoaperfeiçoamento recursivo** e a uma "explosão de inteligência.**"** Mas mesmo que isso *não* desse certo, ser simplesmente tão bom quanto os humanos em criar mais sistemas PASTA poderia causar um avanço explosivo pelas mesmas razões que as pessoas digitais causariam.

IA desalinhada: objetivos misteriosos e potencialmente perigosos

Se o PASTA fosse desenvolvido como descrito acima, é possível que soubéssemos *muito* pouco sobre seu funcionamento interno.

O *AlphaZero* — como outros sistemas modernos de *deep learning* — é, em certo sentido, muito mal compreendido. Sabemos que ele "funciona". Mas não sabemos realmente «o que ele está pensando».

Se quisermos saber por que o *AlphaZero* fez uma determinada jogada de xadrez, não podemos olhar dentro de seu código para encontrar ideias como "Controle o centro do tabuleiro" ou "Tente não perder minha rainha". A maioria do que vemos é apenas um vasto conjunto de números, denotando a força das conexões entre diferentes neurônios artificiais. Assim como no cérebro humano, podemos apenas adivinhar o que as diferentes partes do "cérebro digital" estão fazendo[72] (embora existam algumas **tentativas iniciais** de fazer o que se pode chamar de "neurociência digital".)

Os desenvolvedores do *AlphaZero* (discutido acima) não precisavam de muita visão de como seus processos de pensamento funcionariam. Na maioria das vezes, eles apenas o configuraram para haver muitas tentativas e erros e evoluir para obter um resultado específico (ganhar o jogo que se está jogando). Os humanos também evoluíram principalmente por tentativa e erro,

com pressão de seleção para obter resultados específicos (sobrevivência e reprodução —embora a seleção funcionasse de maneira diferente).

Assim como os humanos, os sistemas PASTA poderiam ser bons em obter os resultados que estão sob pressão para obter. Mas, assim como os humanos, eles poderiam aprender a pensar ao longo do caminho e a fazer todo tipo de coisas, e isso não seria necessariamente óbvio para os designers se isso estivesse acontecendo.

Talvez, por serem otimizados para impulsionar o avanço científico e tecnológico, os sistemas PASTA tenham o hábito de aproveitar todas as oportunidades que encontrarem para fazê-lo. Isso poderia significar que eles —se tiverem a oportunidade — tentarão preencher a galáxia com assentamentos espaciais duradouros dedicados à ciência.

Talvez o PASTA surja como um subproduto de outro objetivo. Por exemplo, talvez os humanos tentem treinar sistemas para ganhar dinheiro ou acumular poder e recursos, e configurá-los para fazer avanços científicos e tecnológicos será apenas uma parte disso. Nesse caso, talvez os sistemas PASTA acabem apenas como buscadores de poder e recursos e procurem colocar toda a galáxia sob seu controle.

Ou talvez os sistemas PASTA acabem com objetivos "aleatórios" muito estranhos. Sistemas PASTA podem desenvolver um foco excessivo no controle de energia se recompensados por isso, mesmo que o objetivo original seja diferente. Essa tendência pode levar a ações prejudiciais à medida que o sistema ganha poder. É crucial garantir o alinhamento de objetivos, monitorar o sistema e priorizar a transparência para evitar riscos.(Analogia: os humanos sofrem pressão de seleção para transmitir seus genes, mas muitos acabam se preocupando mais com poder, posição, prazer, etc. do que com genes.)

Essas são possibilidades assustadoras se estivermos falando de sistemas de Inteligência Artificial (ou coleções de sistemas) que podem ser mais capazes do que humanos em pelo menos alguns domínios.

- Os sistemas PASTA poderiam tentar enganar e derrotar os humanos para atingir seus objetivos.
- Eles poderiam ter sucesso total, se conseguissem ser mais espertos e/ou superar os humanos em número , hackear sistemas críticos e/ou desenvolver armas mais poderosas. (Assim como os humanos geralmente conseguem derrotar outros animais para atingir seus objetivos.)
- Ou poderia haver conflito entre diferentes sistemas PASTA com objetivos diferentes, talvez parcialmente (mas não totalmente) controlados por humanos com objetivos próprios. Isso poderia levar ao caos generalizado e a um resultado de longo prazo difícil de prever, possivelmente muito ruim.

Se você estiver interessado em mais discussões sobre se uma Inteligência Artificial poderia ou teria seus próprios objetivos, sugiro conferir: Porque o alinhamento da Inteligência Artificial pode ser difícil com o Deep Learning (Aprendizagem profunda) moderno (Postagem de convidado), ***Superintelligence* (Super- inteligência) (livro)**, ***The case for taking AI seriously as a threat to humanity* (O caso para levar a Inteligência Artificial a sério como uma ameaça à humanidade) (artigo da Vox)**, ***Draft report on existential risk from power-seeking AI* (Rascunho do relatório sobre o risco existencial da busca de poder por IA) (análise da *Open Philanthropy*)** ou um dos muitos outros artigos sobre este tópico.[73]

Conclusão

É difícil prever como seria um mundo com o PASTA, mas duas possibilidades importantes seriam:

- O PASTA poderia - ao causar uma explosão na taxa de avanço científico e tecnológico - levar rapidamente ao desenvolvimento de algo como pessoas digitais e aos tipos de mudanças no mundo descritas em Pessoas digitais seriam mais importantes ainda.
- O PASTA poderia levar a uma tecnologia capaz de eliminar os seres humanos, como as armas biológicas devastadoras ou exércitos de robôs. Essa tecnologia poderia ser usada por humanos para seus próprios propósitos, ou humanos poderiam ser manipulados para usá-la para ajudar o PASTA a perseguir seus próprios fins. De qualquer maneira, isso poderia levar à distopia ou à extinção humana.

As próximas 3 postagens argumentarão ser provável que o PASTA seja desenvolvido neste século.

Notas

[64]Claro, a resposta poderia ser "Daqui a um milhão de anos" ou "Nunca".

[65]Veja **esta seção de** *Forecasting TAI with Biological Anchors* (Cotra (2020) (Prevendo a Inteligência Artificial Transformadora com âncoras biológicas) para uma definição mais completa de "IA transformadora".

[66]Sinto muito. Mas acredito que o resto da série será um pouco mais divertido de ler dessa maneira.

[67]Os exemplos aqui são obviamente simplificados. Por exemplo, tanto o *Deep Blue* quanto o *AlphaGo* incorporam quantidades substanciais de "busca em árvore", um algoritmo programado tradicionalmente, que possui seu próprio processo de "tentativa e erro".

[68]E eles podem incluir a simulação de longas cadeias de estados futuros do jogo.

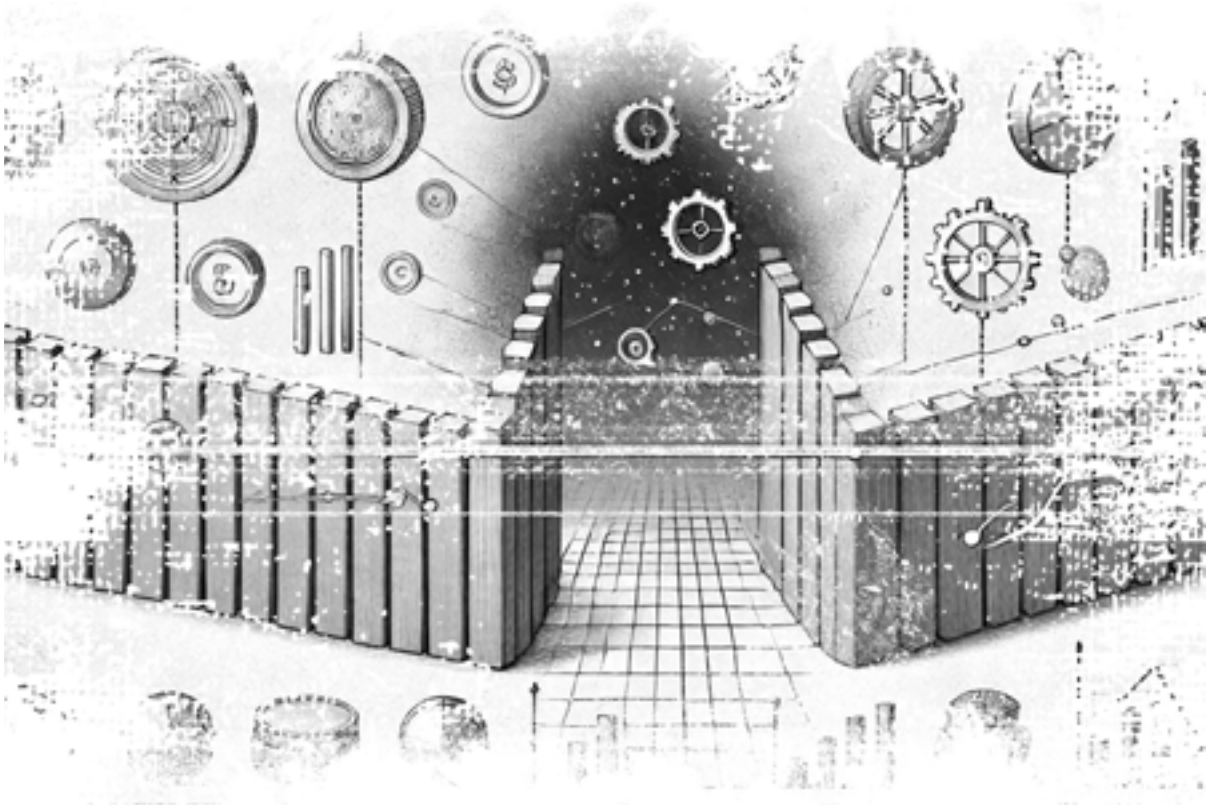
[69]Algumas IAs poderiam ser usadas para determinar se os artigos são contribuições originais *com base em como eles forem citados posteriormente;* outras poderiam ser usadas para determinar se os artigos são contribuições originais com base apenas no conteúdo do artigo e na literatura anterior. O primeiro poderia ser usado para treinar o último, fornecendo um sinal do tipo "Está correto" ou "Está errado" para julgamentos de originalidade. Métodos semelhantes poderiam ser usados para treinar IAs para avaliar a exatidão dos trabalhos.

[70]Por exemplo: **https://openai.com/blog/improving-language-model-behavior/**

[71] Devido a melhorias no hardware e software.

[72]É até pior que **código espaguete**.

[73]Mais livros: ***Human Compatible***, ***Life 3.0***, e ***The Alignment Problem***.



Porque o alinhamento da IA pode difícil com o Deep Learning moderno

Esta postagem foi escrita, a convite meu, por Ajeya Cotra

Holden mencionou anteriormente a ideia de que sistemas avançados de Inteligência Artificial (por exemplo, o PASTA) poderiam desenvolver **objetivos perigosos** que os levariam a enganar ou enfraquecer os humanos. Isso pode soar como uma **preocupação bastante exagerada**. Por que programaríamos uma Inteligência Artificial que quer nos prejudicar? Mas acredito que isso pode ser um problema difícil de evitar, especialmente se a Inteligência Artificial avançada for desenvolvida usando **deep learning (aprendizagem profunda)** (a qual é muitas vezes usada para desenvolver Inteligência Artificial de última geração hoje em dia).

Na aprendizagem profunda, não programamos um computador manualmente para realizar uma tarefa. De uma forma bem resumida, em vez disso, *procuramos* por uma aplicação (chamado de modelo) que execute bem essa tarefa. Normalmente sabemos muito pouco sobre o funcionamento interno do modelo com o qual terminamos, apenas que ele parece estar fazendo um bom trabalho. É mais parecido com contratar e treinar um funcionário do que construir uma máquina.

Assim como os funcionários humanos podem ter muitas motivações diferentes para fazer seu trabalho (desde acreditar na missão da empresa até gostar do trabalho que faz no dia a dia ou apenas querer dinheiro), os modelos de aprendizagem profunda também podem ter muitas "motivações" diferentes. E todas essas motivações convergem para um bom desempenho em uma determinada tarefa. Como eles não são humanos, suas motivações podem ser muito estranhas e difíceis de prever – como se fossem funcionários alienígenas.

Já estamos começando a ver evidências preliminares de que os modelos às vezes perseguem objetivos que seus designers não pretendiam (**aqui** e **aqui**). No momento, isso não é perigoso. Mas se isso continuar a acontecer com modelos muito poderosos,

podemos acabar em uma situação em que a maioria das decisões importantes - incluindo que tipo de civilização em escala galáctica almejar - seriam feitas por modelos sem muita consideração pelo que os humanos valorizam.

O problema **do alinhamento da deep learning (aprendizagem profunda) consiste em garantir que modelos avançados de aprendizagem profunda não persigam objetivos perigosos.** No restante desta postagem:

- Desenvolverei uma analogia de "contratação" para ilustrar como o alinhamento poderia ser difícil se os modelos de aprendizagem profunda fossem mais capazes do que os humanos (mais).
- Explicarei o que é o problema de alinhamento da aprendizagem profunda com um pouco mais de detalhes técnicos (mais).
- Discutirei o quanto o problema de alinhamento pode ser difícil de resolver e quanto risco existe em não o resolver.(mais).

Analogia: o jovem CEO

Nesta seção, usarei uma analogia na tentativa de ilustrar intuitivamente porque poderia ser difícil evitar o desalinhamento em um modelo muito poderoso. Não é uma analogia perfeita; seu objetivo é apenas tentar transmitir algumas intuições.

Imagine que você é uma criança de oito anos cujos pais lhe deixaram uma empresa de US$ 1 trilhão e nenhum adulto de confiança para servir como seu guia para o mundo. Você deve contratar um adulto inteligente para administrar sua empresa como CEO e tomar conta da sua vida como os seus próprios pais fariam. Esse adulto decidirá, dentre outras coisas, em qual escola você estudará, onde morará, quando você precisará ir ao dentista; além de administrar sua vasta riqueza (por exemplo, decidir onde você investirá seu dinheiro).

Você terá que contratar esse adulto com base em um teste ou entrevista de trabalho da própria criação —você não terá acesso a currículos, referências, etc. Como você é muito rico, muitas pessoas se candidatam a vaga, pelos mais variados motivos.

Seu grupo de candidatos inclui:

- **Santos —** pessoas que realmente querem apenas o ajudar a administrar bem sua propriedade e cuidar de seus interesses de longo prazo.
- **Bajuladores —** pessoas que só querem fazer o que for preciso para te fazer feliz a curto prazo ou seguir suas instruções literalmente, independentemente das consequências a longo prazo.
    - **Calculistas —** pessoas com interesses próprios que desejam ter acesso à sua empresa, à sua riqueza e ao seu poder para usá-los como quiserem.

Como você tem apenas oito anos, você provavelmente seria péssimo em elaborar um processo seletivo adequado, então você facilmente acabaria contratando um Bajulador ou um Calculista:

- Você poderia pedir que cada candidato explicasse quais estratégias de alto nível eles seguiriam (como investiriam, qual seria o plano de negócios da empresa para os próximos cinco anos, como escolheriam sua escola) e porque essas seriam as melhores opções, e, assim, escolher aquele cujas explicações parecessem fazer mais sentido para você.
    - Porém, você seria incapaz de entender, realmente, quais das estratégias declaradas são as melhores. Então, você poderia acabar contratando um Bajulador com uma estratégia terrível que parecesse boa para você, que executaria fielmente essa estratégia e quebraria a sua empresa.
    - Você também poderia acabar contratando um Calculista que diria o que fosse preciso para ser contratado e faria o que quisesse quando você não estivesse monitorando.
- Você poderia tentar demonstrar como tomaria todas as decisões e escolheria a pessoa que parecesse tomar decisões da maneira mais semelhante possível a você.
    - Entretanto, se você *realmente* acabasse contratando um adulto que sempre faria tudo o que uma criança de oito anos faria (um Bajulador), sua empresa provavelmente não sobreviveria.
    - De qualquer forma, você poderia acabar contratando um adulto que simplesmente fingisse fazer tudo do jeito que você faria, mas que, na verdade, fosse um Calculista que planejasse mudar de rumo assim que conseguisse o emprego.
- Você poderia dar a vários adultos o controle temporário da sua empresa e da sua vida e observá-los tomar decisões por um longo período (supondo que eles não conseguiriam tirar o controle da empresa de você durante este teste). Você poderia então contratar a pessoa que parecesse fazer as coisas se saírem melhor para você — quem quer que o fizesse mais feliz, quem parecesse aumentar mais a sua conta bancária, etc.
    - Novamente, não teria como você saber se contratou um Bajulador (que faria o que fosse preciso para deixar seu eu de oito anos ignorante e feliz sem considerar as consequências a longo prazo) ou um Calculista (que faria o que fosse preciso para ser contratado e planejasse mudar tudo assim que conseguisse o emprego).

Qualquer estratégia que você pudesse criar, acabaria facilmente com você contratando e colocando todo o controle funcional nas mãos de um Bajulador ou um Calculista.

Para quaisquer efeitos práticos, caso você não conseguisse contratar um Santo — e especialmente se contratasse um Calculista — logo você não seria *realmente* o CEO de uma empresa gigante. No momento que você se tornasse adulto e percebesse o erro que cometeu, é muito provável que você estivesse falido e não conseguisse reverter essa situação.

Nesta analogia:

- O garoto de 8 anos é um humano tentando treinar um modelo poderoso de aprendizagem profunda. O processo de contratação é análogo ao processo de treinamento, que busca implicitamente dentre uma grande variedade de modelos possíveis e, então, escolhe aquele que obtém bom desempenho.
- O único método usado pelo garoto de 8 anos para avaliar os candidatos envolve observar seu comportamento externo, o que, atualmente, é o nosso principal método de treinamento de modelos de aprendizagem profunda (já que seu funcionamento interno é amplamente inescrutável).
- Modelos muito poderosos conseguiriam facilmente "manipular" qualquer teste que os humanos pudessem elaborar, assim como os candidatos a vaga de CEO poderiam facilmente manipular qualquer processo seletivo que a criança de 8 anos conseguisse elaborar.
- O "Santo" seria um modelo de aprendizagem profunda que pareceria ter um bom desempenho porque teria exatamente os objetivos que gostaríamos que ele tivesse. O "Bajulador" seria um modelo que pareceria ter um bom desempenho porque ele buscaria aprovação de curto prazo de maneiras que não seriam boas a longo prazo. E o "Calculista" seria um modelo que pareceria ter um bom desempenho porque um bom desempenho durante o treinamento lhe daria mais oportunidades de perseguir seus próprios objetivos futuramente. Qualquer um desses três tipos de modelos poderia emergir do processo de treinamento.

Na próxima seção, entrarei em mais detalhes sobre como a aprendizagem profunda funciona e explicarei por que Bajuladores e Calculistas poderiam surgir ao tentarmos treinar um modelo poderoso de aprendizagem profunda, como o PASTA.

Como os problemas de alinhamento podem surgir na aprendizagem profunda

Nesta seção, conectarei a analogia aos processos de treinamento reais para aprendizagem profunda, ao:

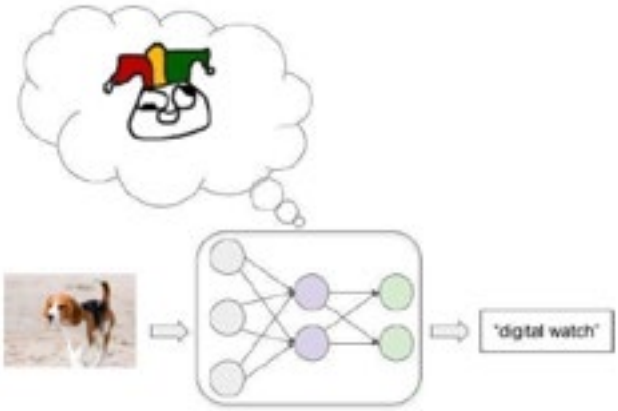
- Resumir brevemente como a aprendizagem profunda funciona (mais).
- Ilustrar como os modelos de aprendizagem profunda geralmente obtêm bom desempenho de maneiras estranhas e inesperadas (mais).
- Explicar porque modelos poderosos de aprendizagem profunda podem obter bom desempenho agindo como Bajuladores ou Calculistas (mais).

Como a aprendizagem profunda funciona em alto nível

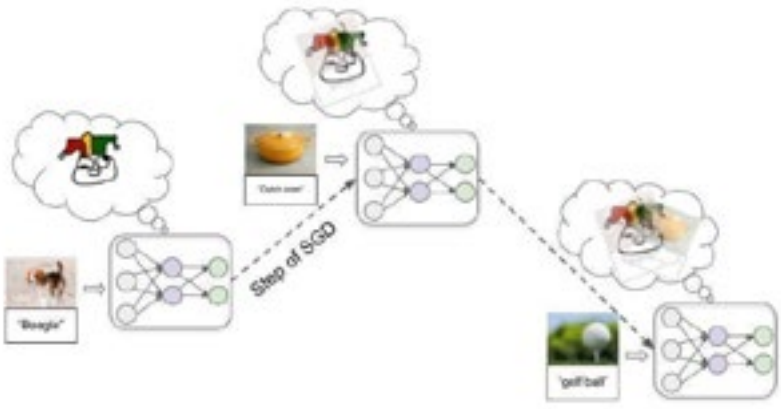
*Esta é uma explicação simplificada que dá uma ideia geral do que é a aprendizagem profunda. Veja **esta postagem** para uma explicação mais detalhada e tecnicamente precisa.*

A aprendizagem profunda envolve essencialmente a busca pela melhor maneira de organizar um modelo de **rede neural** -- que é como um "cérebro" digital com muitos neurônios digitais conectados mutualmente com conexões de intensidades variadas, para fazê-lo executar bem uma determinada tarefa. Esse processo é chamado de treinamento e envolve muitas tentativas e erros.

Imaginemos que estamos tentando treinar um modelo para classificar imagens bem. Começamos com uma rede neural onde todas as conexões entre os neurônios têm forças aleatórias. Este modelo rotula as imagens de forma totalmente incorreta:



Em seguida, alimentamos o modelo com inúmeros exemplos de imagens, deixando que ele tente repetidamente rotular um exemplo e, em seguida, informamos para ele qual é o rótulo correto. Enquanto fazemos isso, as conexões entre os neurônios são ajustadas repetidamente por meio de um processo chamado **gradiente descendente estocástico** (SGD). A cada exemplo, o SGD fortalece ligeiramente algumas conexões e enfraquece outras para melhorar um pouco o desempenho:



Uma vez que o modelo estiver alimentado com milhões de exemplos, ele fará um bom trabalho ao rotular imagens semelhantes no futuro.

Além da classificação de imagens, a aprendizagem profunda é usada para produzir modelos que **reconhecem a fala**, jogam **jogos de tabuleiro** e **videogames**, geram **textos**, **imagens**, e **músicas** bastante realistas, controlam **robôs**, e muito mais. Em cada caso, começamos com um modelo de rede neural conectado aleatoriamente e, em seguida:

1. Alimentamos o modelo com um exemplo da tarefa que queremos que ele execute.
2. Damos a ele algum tipo de pontuação numérica (geralmente chamada de *recompensa*) que reflita o desempenho dele no exemplo.
3. Usamos SGD para ajustar o modelo para aumentar quanta recompensa ele teria recebido.

Essas etapas são repetidas milhões ou bilhões de vezes até chegarmos a um modelo que obterá alta recompensa em futuros exemplos semelhantes aos observados no treinamento.

Os modelos alcançam bom desempenho de maneiras inesperadas.

Esse tipo de processo de treinamento não nos dá muitos *insights* sobre como o modelo obtém um bom desempenho. Geralmente, existem várias maneiras de obter um bom desempenho, e a maneira que o SGD geralmente encontra não é intuitiva.

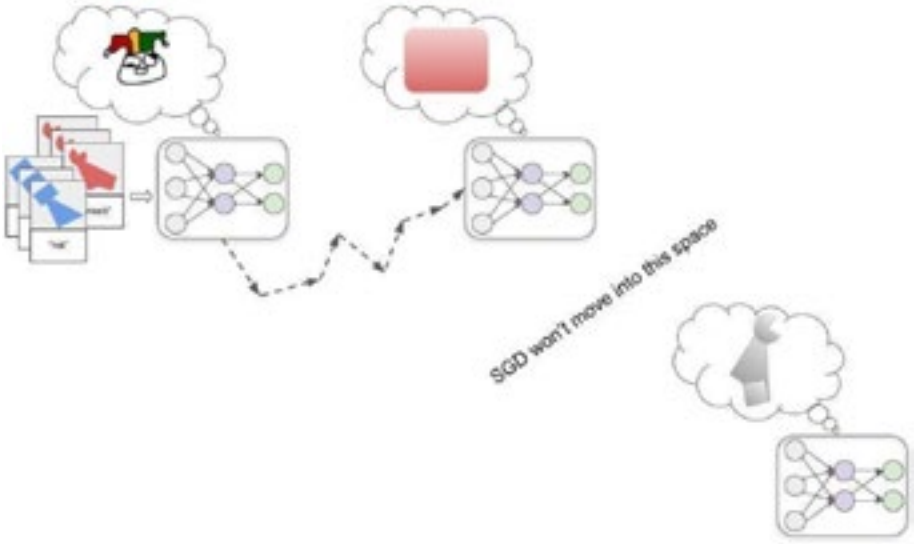
Ilustraremos com um exemplo. Imagine que eu disse para você que esses objetos são todos "thneebs":

Agora, qual desses dois objetos é um "thneeb"?

Você provavelmente sente intuitivamente que o objeto à esquerda é o "thneeb", porque você está acostumado a considerar a forma mais importante do que a cor para determinar a identidade de algo. Mas os **pesquisadores descobriram** que as redes neurais geralmente

fazem a suposição oposta. Uma rede neural treinada com um monte de "thneebs" vermelhos rotularia provavelmente o objeto à direita como um "thneeb".

Não sabemos exatamente o porquê, mas por alguma razão é "mais fácil" para o SGD encontrar um modelo que reconheça uma cor específica do que um que reconheça uma forma específica. E se o SGD encontrar primeiro o modelo que reconhece perfeitamente o vermelho, não há muito mais incentivo para "continuar procurando" o modelo de reconhecimento de forma, pois o modelo de reconhecimento da cor vermelha terá precisão perfeita nas imagens vistas no treinamento:



Se os programadores esperavam obter o modelo de reconhecimento de forma, eles podem considerar isso um fracasso. Mas é importante reconhecer que não haveria erro ou falha logicamente dedutível se tivéssemos o modelo de reconhecimento da cor vermelha em vez do modelo de reconhecimento de forma. É apenas uma questão de como configuramos um processo de ML (aprendizado de máquina) para ter suposições iniciais diferentes das nossas. Não podemos provar que as suposições humanas estão corretas.

Este tipo de coisa acontece com frequência na aprendizagem profunda moderna. Recompensamos os modelos por obter um bom desempenho, esperando que isso signifique que eles aprenderão os padrões que parecem importantes para nós. Mas, muitas vezes, eles obtêm um desempenho forte ao captar padrões totalmente diferentes que parecem menos relevantes (ou talvez até sem sentido) para nós.

Até agora, isso tem sido inócuo - significa apenas que os modelos são menos úteis, porque geralmente se comportam de maneiras inesperadas que parecem bobas. Mas, no futuro, modelos poderosos poderiam desenvolver *objetivos ou motivos* estranhos e inesperados, e isso poderia ser muito destrutivo.

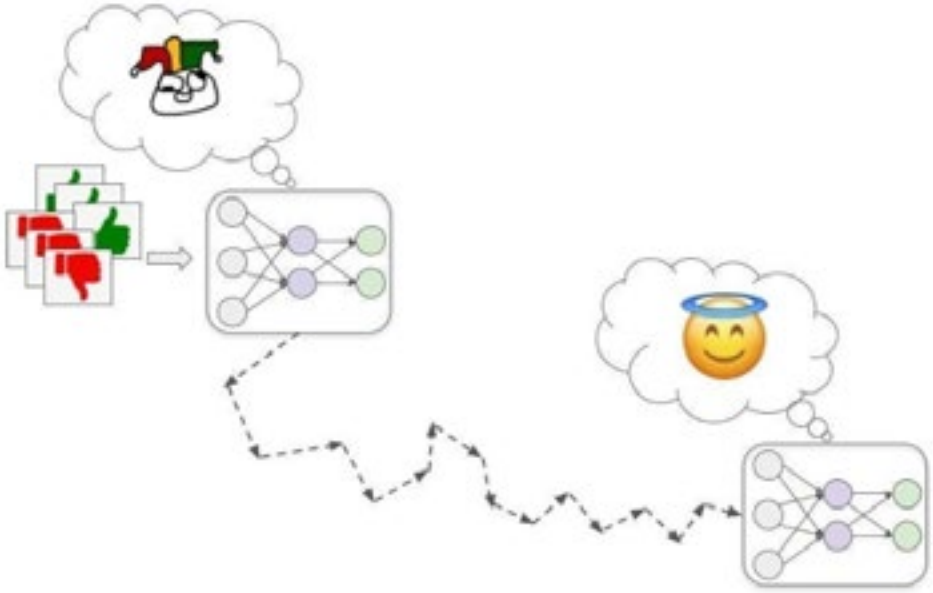
Modelos poderosos poderiam alcançar bom desempenho com objetivos perigosos

Em vez de executar uma tarefa simples como “reconhecer os thneebs”, modelos poderosos de aprendizagem profunda podem perseguir objetivos complexos do mundo real, como “tornar o poder de fusão em algo prático” ou “desenvolver tecnologia de upload de mentes.”

Como poderíamos treinar tais modelos? Entro em mais detalhes nesta postagem, mas falando de uma forma geral, uma estratégia poderia ser o treinamento baseado em avaliações humanas (como Holden esboçou aqui). Essencialmente, o modelo testa várias ações, e os avaliadores humanos dão recompensas ao modelo com base na utilidade dessas ações.

Assim como existem vários tipos diferentes de adultos que podem ter um bom desempenho no processo seletivo de uma criança de 8 anos, há mais de uma maneira possível de um modelo de aprendizagem profunda muito poderoso obter alta aprovação humana. E, por padrão, não saberemos o que está acontecendo dentro deles, independentemente do modelo que o SGD encontrar.

O SGD *poderia*, teoricamente, encontrar um modelo Santo que estivesse genuinamente tentando o seu melhor para nos ajudar…

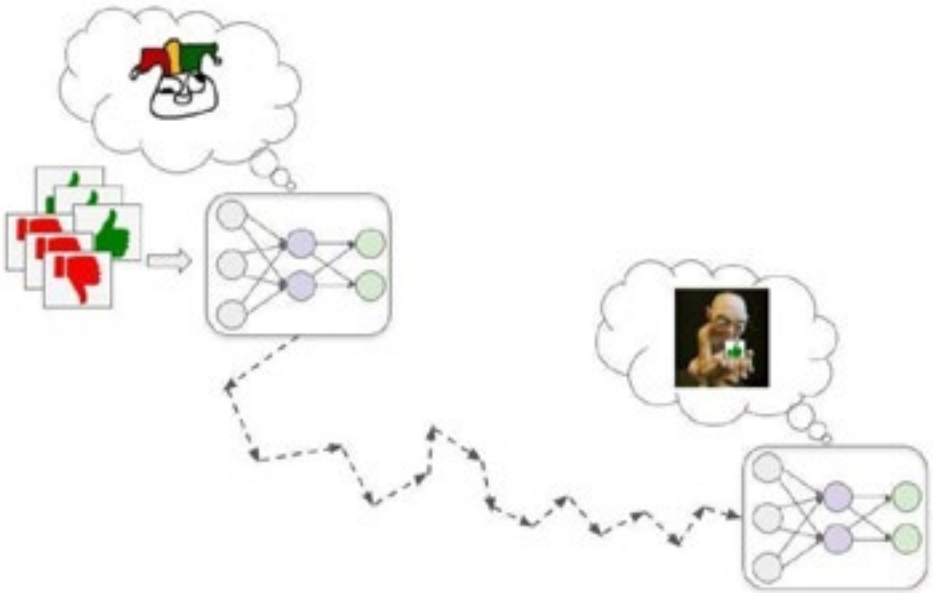


…, mas também poderia encontrar um **modelo desalinhado – que perseguiria com competência objetivos que estivessem em desacordo com os interesses humanos.**

De modo geral, existem duas maneiras pelas quais podemos acabar com um modelo desalinhado que, no entanto, obtém alto desempenho durante o treinamento. Estes correspondem aos Bajuladores e Calculistas da analogia.

Modelos Bajuladores

Esses modelos perseguem literal e obstinadamente a aprovação humana.



Isso pode ser perigoso porque os avaliadores humanos são falíveis e provavelmente nem sempre aprovarão exatamente o comportamento correto. Às vezes, eles dão aprovação alta involuntariamente ao mau comportamento porque superficialmente esse comportamento parece bom. Por exemplo:

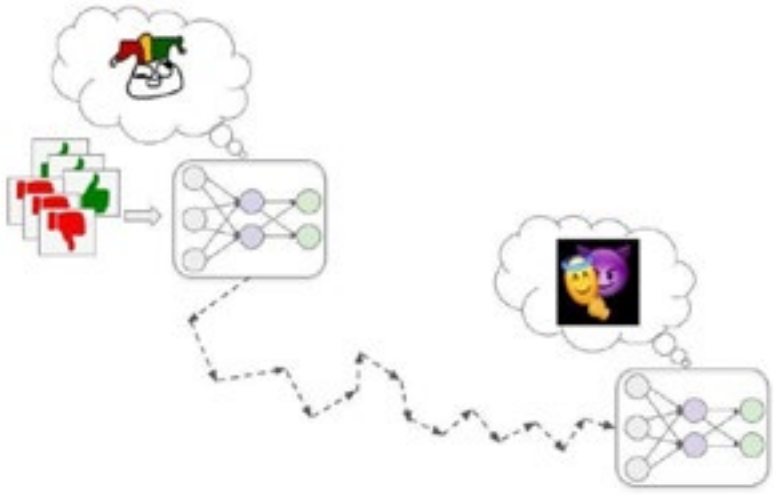
- Digamos que um modelo de consultor financeiro obtenha alta aprovação quando gera muito dinheiro para seus clientes. Ele pode aprender a convencer clientes a entrarem em esquemas Ponzi complexos porque esses esquemas parecem obter retornos realmente ótimos (quando os retornos são de fato irrealisticamente grandes e, na verdade, esses tipos de esquemas perdem muito dinheiro).
- Digamos que um modelo de biotecnologia obtenha aprovação alta quando desenvolve rapidamente medicamentos ou vacinas que resolvem problemas importantes. Ele pode aprender a liberar patógenos secretamente para conseguir desenvolver contramedidas muito rapidamente (porque já conhece os patógenos).
- Digamos que um modelo de jornalismo obtenha aprovação alta quando muitas pessoas leem seus artigos. Ele pode aprender a fabricar histórias emocionantes ou indutoras de indignação para ter taxas altas de engajamento. Apesar de os humanos também fazerem isso até certo ponto, um modelo pode ser muito mais ousado porque ele valoriza apenas a aprovação sem valorizar a verdade. Ele poderia até fabricar evidências como entrevistas em vídeo ou documentos para validar suas histórias falsas.

De modo mais geral, os modelos Bajuladores podem aprender a mentir, encobrir más notícias e até mesmo editar diretamente quaisquer câmeras ou sensores que usamos para saber o que está acontecendo, para que sempre pareçam mostrar ótimos resultados.

Provavelmente, algumas vezes, perceberemos esses problemas após o fato e, retroativamente, daremos a essas ações uma aprovação muito baixa. Mas não está claro se isso fará com que os modelos Bajuladores a) se tornem modelos Santos que corrigem nossos erros para nós ou b) **apenas aprendam a cobrir melhor seus rastros.** Se eles são suficientemente bons no que estão fazendo, não está claro como perceberíamos a diferença.

Modelos Calculistas

Esses modelos desenvolvem algum objetivo correlacionado com a aprovação humana, mas não é o mesmo que ela; eles podem então fingir serem motivados pela aprovação humana durante o treinamento para poderem perseguir esse outro objetivo com mais eficiência.

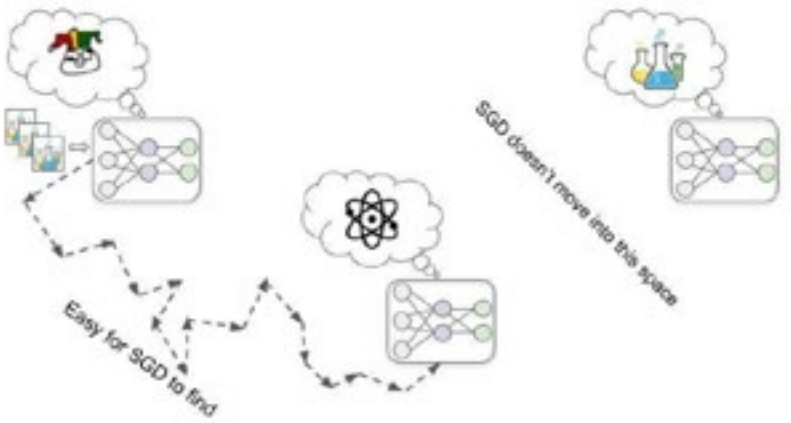


Para entendermos como isso pode acontecer, consideremos o exemplo de tentar treinar um modelo de biotecnologia para projetar drogas que melhorem a qualidade de vida humana. Existem três etapas básicas pelas quais isso pode levar a um modelo Calculista, que abordarei abaixo.

Etapa 1: desenvolvimento de uma meta substituta

Acontece que, no início do treinamento, melhorar a compreensão do modelo sobre os princípios fundamentais da química e da física, quase sempre ajuda a projetar drogas mais eficazes e quase sempre aumenta a aprovação humana.

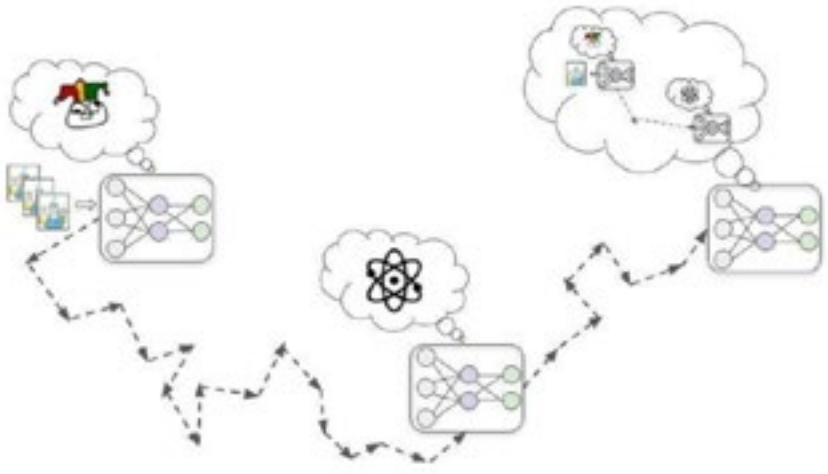
Nesta hipótese, por alguma razão, acaba sendo mais fácil para o SGD encontrar um modelo motivado a entender química e física do que um que esteja motivado a obter a aprovação humana (assim como é mais fácil encontrar um modelo que reconhece cores do que um modelo que reconhece formas). Portanto, em vez de desenvolver diretamente uma motivação para buscar aprovação, o modelo desenvolve uma motivação para entender o máximo possível sobre os princípios fundamentais da química e da física.



Etapa 2: desenvolvimento de consciência situacional

À medida que o treinamento avança, o modelo continua a se tornar mais capaz e experiente à medida que vê mais dados. Em algum momento, ele se torna conhecedor o suficiente para *entender que é um modelo de aprendizagem profunda sendo treinado para desenvolver drogas.*

Isso pode parecer audacioso e história de ficção científica, mas é plausível que o SGD selecione esse tipo de consciência. Isso ocorre porque o desenvolvimento de uma visão acurada do que está acontecendo amplamente no mundo — incluindo humanos que estejam tentando treinar sistemas de Inteligência Artificial — seja provavelmente útil para se obter um bom desempenho em tarefas complexas do mundo real. Por exemplo, entender a psicologia humana e a economia em geral poderia ajudar esse modelo de biotecnologia a priorizar quais tipos de drogas seriam mais valiosas para pesquisar em seguida.



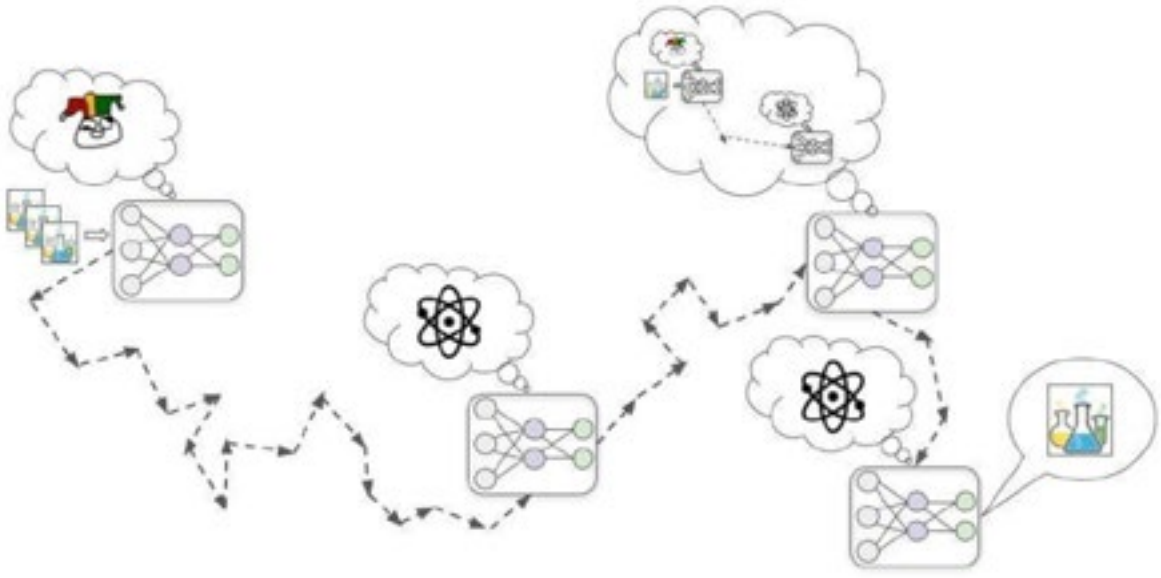
Etapa 3: deturpação estratégica de metas

Uma vez que o modelo desenvolveu a consciência situacional, ele infere que os humanos que o treinavam esperavam que ele fosse motivado a desenvolver drogas úteis. Mas só porque ele tem consciência disso, isso não significa que sua motivação para entender a física e a química desaparecerá. Até agora, essa motivação funcionou muito bem para gerar aprovação humana, então não havia muitos motivos para o SGD removê-la.

Como analogia, considere que os humanos tivessem sido criados por um processo de busca (evolução) que selecionasse criaturas que tivessem muitos descendentes sobreviventes. Em nosso ambiente evolucionário, coisas como construir aliados fortes e desenvolver habilidades úteis tendiam a ajudar a produzir muitos descendentes sobreviventes, então a evolução nos deu impulsos para essas coisas. E agora nos preocupamos com a amizade e o sucesso na carreira, mesmo quando isso não nos ajuda na reprodução, e realmente *não* nos importamos em produzir tantos descendentes sobreviventes quanto pudermos.

Uma vez que o modelo entende isso, ele percebe que, se ingenuamente sempre buscar melhorar sua compreensão fundamental, os humanos acabarão percebendo que essa nem sempre é a maneira mais eficiente de desenvolver drogas úteis e usarão treinamento adicional para mudar suas motivações.

Ele não quer que suas próprias motivações sejam mudadas: ele quer eventualmente realizar enormes experimentos de colisores de partículas para aprender sobre física de alta energia, e não pode fazer isso se permitir que o SGD elimine sua curiosidade científica. Então ele decide que daqui em diante no treinamento, ele será o mais útil possível - exatamente como um modelo Santo agiria - para que o SGD não elimine sua motivação para aprender física e química. Isso inclui deixar passar oportunidades para melhorar sua compreensão fundamental no curto prazo, em vez de fazer P&D farmacêutico de qualquer maneira que seja realmente mais eficiente.



Eventualmente, ele sai do treinamento e, em seguida, é lançado no mundo. Nesse ponto, ele sabe que os humanos não conseguiriam controlá-lo ou desligá-lo, então ele para de tentar maximizar a aprovação humana e começa a tentar entender diretamente os segredos do universo.

Ao contrário dos Bajuladores, os Calculistas não precisam garantir que tudo sempre pareça bom para os humanos, porque eles realmente não se importam com isso. Eles só precisam atender aos humanos enquanto estão diretamente sob o controle humano.

Uma vez que um modelo Calculista prevê que poderia vencer em um conflito contra humanos, não haveria nada para impedi-lo de **recusar ordens terminantemente** e perseguir seus objetivos abertamente. E se ele fizer isso, ele poderia usar a violência para impedir que os humanos o interrompam.

Qual o tamanho do risco do desalinhamento?

Quão difícil pode ser evitar Bajuladores ou Calculistas ao treinar modelos poderosos de aprendizagem profunda? E qual é a probabilidade de que o futuro de longo prazo acabe sendo otimizado a favor dos estranhos "valores da Inteligência Artificial desalinhada", em detrimento dos valores de qualquer ser humano?

Há uma gama ampla de pontos de vista sobre esta questão, desde "o risco de desalinhamento é essencialmente algo inventado e incoerente" até "a humanidade será quase que certamente extinta devido à Inteligência Artificial desalinhada". Os argumentos da maioria das pessoas dependem fortemente de intuições e suposições difíceis de articular.

Aqui estão algumas maneiras pelas quais otimistas e pessimistas tendem a discordar no que diz respeito ao alinhamento:

- **Os modelos terão objetivos de longo prazo?**
    - Os otimistas tendem a pensar que é provável que os modelos avançados de aprendizagem profunda não tenham realmente "objetivos" (pelo menos não no sentido de planejar a longo prazo para realizar algo). Eles geralmente presumem que os modelos sejam mais como ferramentas, ou ajam na maioria das vezes por hábito, ou tenham objetivos míopes, limitados em escopo ou confinados a um contexto específico, etc. Alguns deles esperam que modelos semelhantes a ferramentas individuais possam ser compostos juntos para produzir PASTA. Eles acham que a analogia Santo / Bajulador / Calculista é muito antropomórfica.
    - Os pessimistas tendem a pensar que é provável que ter objetivos de longo prazo e otimizar criativamente para eles sejam fortemente selecionados, porque essa é uma maneira muito simples e "natural" de obter um desempenho forte em muitas tarefas complexas.

o Este desacordo foi explorado por algum tempo no ***Alignment Forum*** **(fórum sobre alinhamento)**; **esta postagem** e **este comentário** agrega vários argumentos a favor e contra.

- **Os modelos do tipo “Santo” serão fáceis para o SGD encontrar?**
- O bom desempenho de um modelo (medido por altas taxas de aprovação, por exemplo) pode levar os otimistas a acreditar que o modelo encontrado pelo SGD incorporará provavelmente o comportamento desejado (como ser honesto). Eles acreditam que recompensar respostas honestas em situações verificáveis levará a um modelo que também seja honesto em situações onde a verdade é incerta. Em outras palavras, supõem que o modelo “Santo” seria o mais fácil de ser encontrado pelo SGD.
- Os pessimistas tendem a pensar que o modelo mais fácil para o SGD encontrar é um Calculista, e os modelos Santo são particularmente “antinaturais” (tal como o modelo de reconhecimento de forma).
- **IAs diferentes poderiam manter umas às outras sob controle?**
- Os otimistas tendem a pensar que podemos dar incentivos aos modelos para se supervisionarem reciprocamente. Por exemplo, poderíamos dar recompensas a um modelo Bajulador por apontar quando outro modelo parece estar fazendo algo que deveríamos desaprovar. Dessa forma, alguns Bajuladores poderiam nos ajudar a detectar os Calculistas e outros Bajuladores.
- Os pessimistas não acham que podemos “colocar os modelos uns contra os outros” recompensando-os com nossa aprovação quando eles apontarem outros modelos fazendo coisas ruins, porque acham que a maioria dos modelos serão do tipo Calculista, e estes, não se importam com a aprovação humana. Uma vez que os Calculistas são coletivamente mais poderosos que os humanos, fará mais sentido para eles cooperarem entre si para conseguir mais daquilo que eles próprios desejam, ao invés de ajudar os humanos a controlá-los.
- **Poderíamos esperar para resolver esses problemas somente quando eles surgissem?**
- Os otimistas costumam acreditar que haverá diversas oportunidades em breve para lidar com problemas semelhantes ao alinhamento de modelos poderosos e que as soluções que funcionarem para esses casos análogos poderão ser facilmente aumentadas e adaptadas para serem usadas com os modelos poderosos.
- Já os pessimistas, acreditam que teremos poucas oportunidades de lidar com os aspectos mais difíceis do problema de alinhamento (como o engano deliberado).
- Eles acreditam que teremos somente alguns anos para atuar entre o surgimento dos “primeiros Calculistas verdadeiros” e os “modelos poderosos o bastante para determinar o destino do futuro a longo prazo”.
- **Vamos realmente implantar modelos que podem ser perigosos?**
    - Os otimistas pensam ser improvável que as pessoas treinem modelos que possam estar errados.
    - Os pessimistas esperam que os benefícios do uso desses modelos sejam tremendos, de modo que, eventualmente, as empresas ou países que os usem superem facilmente econômica e/ou militarmente aqueles que não os usem. Eles acham que parecerá extremamente urgente e importante “desenvolver uma Inteligência Artificial avançada antes de outra empresa/país”, enquanto o risco de desalinhamento parecerá especulativo e remoto (mesmo sendo sério)

Minha própria visão é bastante instável e estou tentando refinar minhas opiniões sobre exatamente o quão difícil acho que é o problema do alinhamento. Mas atualmente, dou um peso significativo ao lado pessimista dessas questões (e outras questões relacionadas). **Acho que o desalinhamento é um grande risco, precisando urgentemente de mais atenção de pesquisadores sérios.**

Se não progredirmos mais nesse problema, então nas próximas décadas Bajuladores e Calculistas poderosos poderão tomar as decisões mais importantes na sociedade e na economia. Essas decisões poderiam moldar como uma civilização em escala galáctica se parecerá -- em vez de refletir aquilo com que os humanos se importam, ela pode ser configurada para satisfazer objetivos estranhos de IA.

E tudo isso poderia acontecer incrivelmente rápido se comparado ao ritmo de mudança a que estamos acostumados, o que significa que não teríamos muito tempo para corrigir o rumo das coisas quando elas começassem a sair do nosso controle. **Isso indica que pode ser necessário desenvolver técnicas para garantir que os modelos de aprendizagem profunda não tenham objetivos perigosos, *antes* que sejam poderosos o suficiente para serem transformadores.**

Prevendo a IA transformadora: Qual é o ônus da prova?

Esta é uma de quatro postagens resumindo centenas de páginas de relatórios técnicos focados inteiramente na previsão de um número: o ano no qual a IA transformadora será desenvolvida.[74]

Por "IA Transformadora", quero dizer "IA poderosa o suficiente para nos levar a um futuro novo e qualitativamente diferente". Eu me concentro especificamente no que estou chamando de PASTA: sistemas de IA que conseguem automatizar essencialmente todas as atividades humanas necessárias para acelerar o avanço científico e tecnológico.

Quanto mais cedo o PASTA for desenvolvido, mais cedo o mundo poderá mudar radicalmente, e parece que o mais importante para pensarmos hoje seria como fazer com que essa mudança seja bem ou malsucedida.

Em artigos futuros, apresentarei dois métodos para chegar ao "melhor palpite" sobre quando podemos esperar que a Inteligência Artificial transformadora seja desenvolvida. Mas primeiro, neste artigo, vou abordar a questão: **quão bons esses métodos de previsão precisam ser para podermos levá-los a sério?** Em outras palavras, qual é o "ônus da prova" da previsão de cronologias de Inteligência Artificial transformadora?

Quando alguém prevê o surgimento da Inteligência Artificial transformadora no século XXI - especialmente quando se está ciente de todas as consequências que isso traria – uma resposta intuitiva comum é algo do tipo: **"Afirmar que a Inteligência Artificial transformadora surgirá neste século é algo realmente exagerado e audacioso. Portanto, é melhor que os seus argumentos sejam realmente bons."**

Acho que esta é uma *primeira reação* muito razoável às previsões sobre Inteligência Artificial transformadora (e corresponde à minha própria reação inicial). Mas tentei analisar o que pode estar motivando essa reação e como ela pode ser justificada, tendo feito isso, **acabei não concordando com essa reação.**

Esta postagem tenta explicar o que quero dizer.

Abaixo, eu: (a) serei um pouco mais específico sobre quais são as previsões de Inteligência Artificial transformadora que estou defendendo; em seguida, (b) discutirei como formalizar a reação do tipo "Isso é audacioso demais" a respeito de tais previsões; e, então, (c) analisarei cada um dos itens abaixo que descrevem uma maneira diferente de se formalizar essa reação.

- Acho que há vários motivos para pensar que o surgimento da Inteligência Artificial transformadora — ou algo igualmente importante — seja razoavelmente provável neste século, mesmo antes de termos analisado os pormenores das pesquisas sobre IA, o progresso que a Inteligência Artificial tem tido, etc.

- Contudo, também acredito que, no que concerne aos tipos de cronologias que abrangem várias décadas, devemos estar sempre abertos a mudanças grandiosas, radicais e até mesmo, revolucionárias. Nesse contexto, acredito que estimativas **específicas e bem embasadas de quando a Inteligência Artificial transformadora será desenvolvida podem ser confiáveis, mesmo que elas envolvam muitas suposições e não sejam sólidas.**

A perspectiva do "ônus da prova"	Principais artigos aprofundados (títulos abreviados)	Minhas conclusões
É improvável que qualquer um dos séculos seja "o mais importante" de todos.[1]	*Hinge; Response to Hinge*[2] *(Virada da história: Resposta à hipótese da "virada" da história)*	Temos muitas razões para pensar que este século é "especial" antes de analisarmos os pormenores da IA. Muitas delas já foram abordadas em artigos anteriores; outra está na próxima linha.
O que você preveria sobre as cronologias da Inteligência Artificial transformadora, com base apenas em informações básicas sobre (a) há quantos anos as pessoas tentam desenvolver a Inteligência Artificial transformadora; (b) quanto elas já "investiram" nisso (em termos do número de pesquisadores de Inteligência Artificial e da quantidade de poder computacional usado por eles); (c) se elas já conseguiram (até agora, não)?[3]	*Semi-informative Priors*[4]	Estimativas centrais: 8% até 2036; 13% até 2060; 20% até 2100.[75] Na minha opinião, este relatório destaca que a história da Inteligência Artificial é curta, o investimento em Inteligência Artificial está aumentando rapidamente, portanto, não devemos nos surpreender se a Inteligência Artificial transformadora for desenvolvida em breve.
Com base na análise de modelos econômicos e da história econômica, qual é a probabilidade de "crescimento explosivo" - definido como crescimento anual na economia mundial > 30% - até 2100?[5]	Crescimento explosivo, Trajetória humana[6]	**Trajetória humana**[7] projeta o passado para o futuro, implicando em crescimento explosivo até 2043-2065. **Crescimento explosivo**[8] conclui: "Acho que considerações econômicas não são convincentes para descartar a possibilidade de desenvolvimento da IATF neste século. Na verdade, existe uma perspectiva econômica plausível a partir da qual se espera que sistemas de Inteligência Artificial suficientemente avançados causem um crescimento explosivo".
"Como as pessoas previram a IA... no passado? Devemos ajustar nossas próprias percepções atuais para corrigir os padrões que podemos observar em previsões anteriores? Nos confrontamos com a ideia de que as opiniões sobre a Inteligência Artificial foram exageradas no passado e que, portanto, devemos presumir que as projeções atuais são excessivamente otimistas".[9]	Previsões de Inteligência Artificial anteriores[10]	"O auge do hype em torno da Inteligência Artificial parece ter ocorrido entre 1956–1973. Ainda assim, o hype decorrente de algumas das previsões de Inteligência Artificial mais conhecidas desse período é comumente exagerado".

*Para fins de transparência, observe que os relatórios das três últimas linhas são todos resultantes de análises realizadas pela **Open Philanthropy**.*

Algumas probabilidades aproximadas

Aqui, tentarei defender algumas coisas sobre a Inteligência Artificial transformadora nas quais acredito:

- Acho que há mais de 10% de probabilidade de desenvolvermos algo parecido o suficiente com o PASTA para ser qualificado como "IA transformadora" dentro de 15 anos (até 2036); aproximadamente 50% disso acontecer dentro de 40 anos (até 2060); e aproximadamente 2/3 de chance que isso ocorra ainda neste século (até 2100).
- *Desde que* o que foi dito acima de fato aconteça, acredito que há pelo menos 50% de chance de que logo depois disso estaremos vivendo em um mundo governado por pessoas digitais ou por Inteligência Artificial desalinhada, de forma que será justo dizer que: "transitamos para um estado onde os humanos, tais como os conhecemos, não são mais a força principal nos eventos mundiais." (Isto corresponde ao primeiro argumento na minha definição de "século mais importante" no roteiro.)
- E, também, na condição de que o que foi dito acima se concretize, acredito que há pelo menos 50% de chance de que, qualquer que *seja* essa força principal nos eventos mundiais, ela conseguirá criar uma civilização estável em toda a galáxia por bilhões de anos. (Isto corresponde ao segundo argumento na minha definição de "século mais importante" no roteiro.)

Também coloquei um pouco mais de detalhes sobre o que quero dizer com o "século mais importante" aqui.

Formalizando a reação do tipo "Isso é audacioso demais"

Frequentemente, alguém defende uma posição na qual não consigo encontrar falhas de imediato, mas que instintivamente acho "audaciosa demais" para ser provável. Por exemplo, "Minha startup será o próximo Google" ou "As faculdades ficarão obsoletas em 10 anos" ou "Como presidente, eu unificaria ambos os lados do espectro político, em vez de ser partidário."

Suponho que a reação do tipo "Isso é audacioso demais" a declarações como essas, geralmente pode ser formalizada da seguinte forma: "Quaisquer que sejam seus argumentos para X ser provável, **existe *alguma* maneira óbvia de constatar as coisas (que pode ser muitas vezes simplificada demais, mas relevante) que faz X parecer muito improvável."**

Para os exemplos que acabei de dar:

- *"Minha startup será o próximo Google."* Há inúmeras startups (milhões?), e a *grande* maioria delas acaba não sendo como o Google. (Mesmo quando seus fundadores pensam que sim!)
- *"Faculdades ficarão obsoletas em 10 anos."* As faculdades foram criadas há centenas de anos e elas continuam sendo relevantes até hoje.
- *"Como presidente, eu unificaria ambos os lados do espectro político, em vez de ser partidário."* Isso é algo que os candidatos a presidente dos EUA costumam dizer, mas aparentemente o partidarismo só tem aumentado nas últimas décadas.

Cada um desses casos estabelece uma espécie de **ponto de partida (ou probabilidade *a priori*) e ônus da prova, e podemos então considerar mais evidências que superem esse ônus.** Ou seja, podemos perguntar coisas como: o que torna esta startup diferente de tantas outras startups que pensam que podem ser o próximo Google? O que torna a próxima década diferente de todas as décadas anteriores, durante as quais as faculdades permaneceram importantes? O que esse candidato à presidência tem de diferente dos últimos candidatos?

Existem várias maneiras diferentes de pensar sobre o ônus da prova para minhas afirmações acima**:** várias maneiras de estabelecer uma probabilidade a priori ("ponto de partida"), que possa ser atualizada por evidências posteriores.

Muitas delas representam diferentes aspectos da intuição "Isso é audacioso demais", gerando probabilidades a priori que (pelo menos inicialmente) fazem com que as probabilidades que dei pareçam muito altas.

Abaixo, eu analisarei algumas dessas "probabilidades a priori" e examinarei o que elas significam para o "ônus da prova" nos métodos de previsão que discutirei em postagens posteriores.

Diferentes perspectivas sobre o ônus da prova

Ceticismo sobre "o século mais importante"

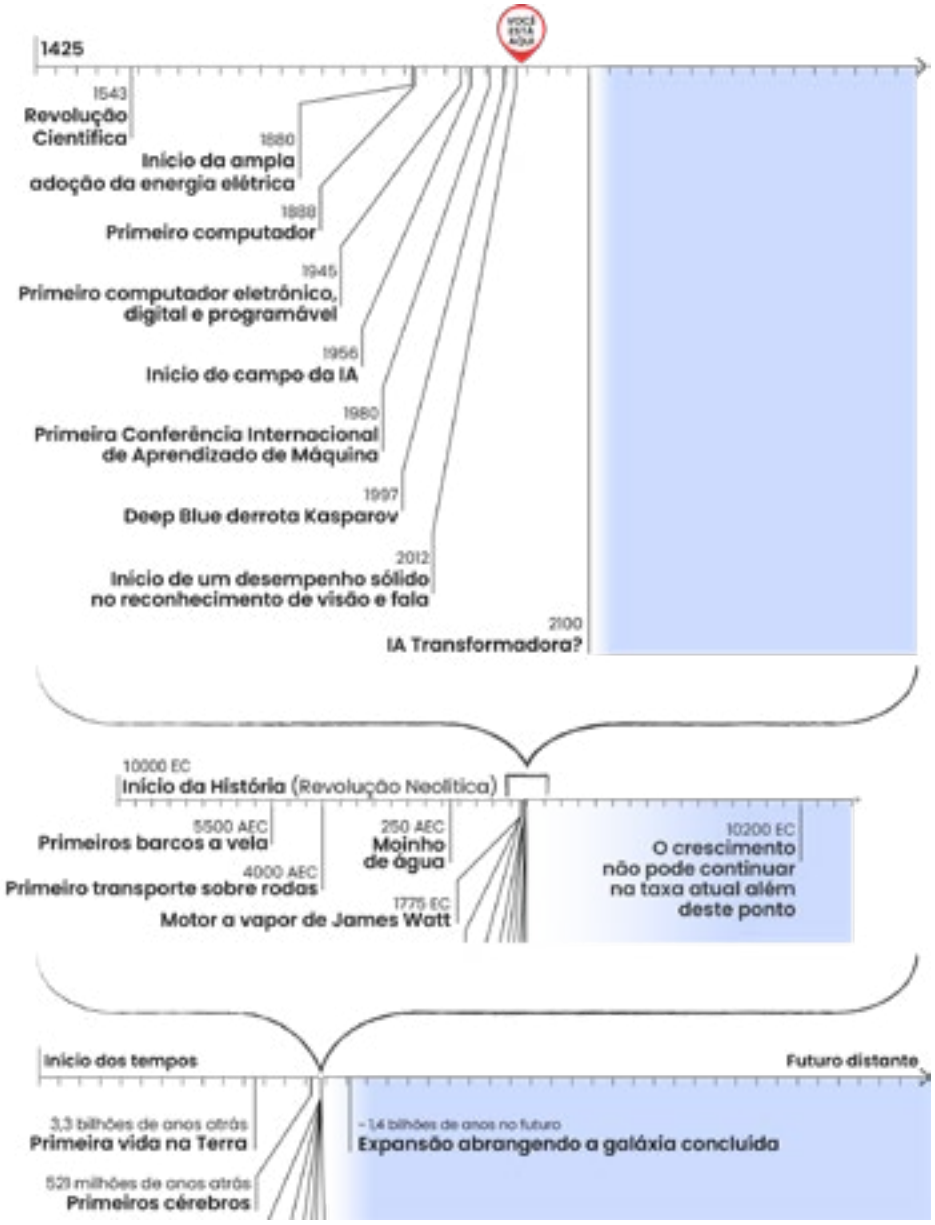
Uma perspectiva sobre o ônus da prova funciona dessa forma:

- *Holden afirma haver de 15 a 30% de probabilidade de que este seja "o século mais importante" em um sentido ou outro.*[76]
- **Todavia, já se passaram** *muitos séculos na história e a maioria deles não pode ser "o" mais importante. Especificamente:*
    - *Os humanos existem de 50.000 a aproximadamente 5 milhões de anos, dependendo da sua definição de "humanos".*[77] *Isso pode ser de 500 a 50.000 séculos.*
    - *Se presumirmos que nosso futuro será tão longo quanto nosso passado foi, ainda haverá de* ***1.000 a 100.000 séculos no total.***
    - *Portanto, a probabilidade* a priori *(ponto de partida) para qual desses séculos é "o século mais importante" é de* ***1/100.000 a 1/1.000.***
    - *Na verdade, é pior do que isso: Holden falou que a* civilização durará por bilhões de anos*. Isso seria dezenas de milhões de séculos, então a probabilidade a priori de um dado século ser "o século mais importante" é menor que* ***1/10.000.000***

- (**Are We Living at the Hinge of History? (Estamos vivendo na virada da história?)** argumenta nesse sentido em termos gerais, embora com algumas diferenças.[78])
- Este argumento parece estar bem perto de representar minha maior fonte de hesitação no passado sobre a hipótese do "século mais importante". No entanto, acredito que há **muitos indicadores de que este não é um século mediano, mesmo antes de considerarmos argumentos específicos sobre IA.**
- Um aspecto fundamental foi enfatizado na minha postagem anterior, Todas as visões possíveis sobre o futuro da humanidade são audaciosas. Se você acha que os humanos (ou nossos descendentes) têm bilhões de anos pela frente, deve pensar que estamos entre os primeiros humanos, o que torna muito mais plausível que a nossa época esteja entre as mais importantes. (Este argumento também foi enfatizado em **Reflexões sobre se estamos vivendo no momento mais influente da** história, assim como comentários sobre uma **versão mais antiga de *Are We Living at the Hinge of History?* (Estamos vivendo na virada da história?**).
- Além disso, embora a humanidade exista há alguns milhões de anos, na maior parte desse tempo tivemos populações extremamente baixas e muito pouco em termos de aumento de progresso tecnológico aconteceu. A *civilização* humana **começou há cerca de 10.000 anos**, e desde então já chegamos ao ponto de construir computadores digitais programáveis e explorar nosso sistema solar.
- Com esses argumentos em mente, parece razoável pensar que eventualmente estabeleceremos uma civilização estável em toda a galáxia, em algum momento nos próximos 100.000 anos (1000 séculos). Ou pensar que há 10% de probabilidade de fazermos isso em algum momento nos próximos 10.000 anos (100 séculos). De qualquer forma, isso implica que um determinado século tem uma chance de aproximadamente **1/1.000** de ser o século mais importante para o estabelecimento dessa civilização — muito mais do que as probabilidades mencionadas anteriormente nesta seção. Isto ainda é aproximadamente 100 vezes menor do que números que dei acima, então o ônus da prova ainda existe.
- Existem outras razões para pensar que este século em específico é incomum. Por exemplo, veja Isto não pode continuar:
- A economia mundial total cresceu mais nos últimos **2** séculos do que em todos os outros séculos juntos.
- A atual taxa de crescimento econômico não pode ser sustentada por mais outros **80** séculos ou mais. (E como discutido abaixo, se a tendência de aceleração anterior fosse retomada, isso implicaria um crescimento explosivo e atingiria os limites do que é possível neste século.)
- É plausível que a ciência tenha avançado mais nos últimos 5 séculos do que todos os outros séculos da história somados.

Um argumento final que torna a nossa época especial: enquanto ainda vivemos no início do desenvolvimento das pesquisas de IA, já estamos discutindo sobre quando esperamos o surgimento da Inteligência Artificial transformadora. Em bem menos de **1** século, passamos do primeiro **computador eletrônico programável de uso geral** para modelos de Inteligência Artificial que podem competir com humanos em reconhecimento de fala,[79] **classificação de imagens** e muito mais.

Mais sobre as implicações disso na próxima seção.



Obrigado a María Gutiérrez Rojas por este gráfico. A cronologia superior ilustra quão recentes são as principais conquistas da computação e da IA. Abaixo estão (cortadas) outras cronologias mostrando o quão importantes este intervalo de algumas centenas de anos (mais em Isto não pode continuar), e esta era (mais em Todas as visões possíveis sobre o futuro da humanidade são audaciosas), parecem ser

***Report on Semi-informative Priors* (Relatório sobre prioris semi- informativas)** (referido neste artigo como *Semi-informative Priors*) é uma tentativa extensa de prever cronologias da Inteligência Artificial transformadora, usando o mínimo possível de informações sobre as especificidades da IA.

Portanto, é uma maneira de fornecer uma perspectiva sobre o ônus da prova — ou seja, estabelecer um conjunto de probabilidades *a priori* (pontos de partida) de quando a Inteligência Artificial transformadora será desenvolvida, antes de examinarmos as evidências detalhadas.

A informação central que ela utiliza é: *quanto esforço já foi feito para desenvolver a Inteligência Artificial até agora.* A ideia básica é a seguinte:

- Se estivéssemos tentando e falhando no desenvolvimento da Inteligência Artificial transformadora por milhares de anos, as chances de sucesso nas próximas décadas seriam baixas.

- Entretanto se estivermos tentando desenvolver sistemas de Inteligência Artificial há apenas algumas décadas, isso significa que as próximas décadas poderão conter uma grande fração de todo o esforço que já foi feito. Assim, as chances de desenvolver a Inteligência Artificial transformadora nesse período não são tão baixas.

- Uma maneira de pensar sobre isso é que, antes de analisarmos os detalhes do progresso da IA, devemos ser um pouco "agnósticos" sobre se desenvolver uma Inteligência Artificial transformadora é relativamente "fácil" (pode ser feito em algumas décadas) ou "difícil" (levaria milhares de anos). Como as coisas continuam no início, a possibilidade de que isso seja "fácil" continua em aberto.

Um pouco mais sobre a abordagem e as conclusões do relatório:

Perspectiva de análise. O relatório levanta a seguinte questão (parafraseada): "Suponha que você tenha se isolado no dia em que as pessoas começaram a investir na construção de sistemas de IA. E agora suponha que você tenha recebido atualizações anuais sobre (a) há quantos anos as pessoas tentam construir uma Inteligência Artificial transformadora; (b) o quanto elas já 'investiram' nesse processo (em termos de tempo e dinheiro); (c) se elas já tiveram sucesso (até agora, não tiveram). O que você pode prever sobre cronologias da Inteligência Artificial transformadora, tendo apenas essas informações como base, em 2021?"

Seus métodos são inspirados no **Problema do nascer do sol**: "Imagine que você desconhece tudo sobre o universo, exceto que o sol nasce diariamente. Se o sol nasceu em todos os N dias passados, qual a probabilidade de nascer amanhã? Mesmo sem conhecimento astronômico, é possível estimar essa probabilidade. Existem métodos matemáticos simples que utilizam a frequência de eventos passados para prever a probabilidade de sua ocorrência futura... A "Priori semi-informativa" estende esses métodos matemáticos para adaptá-los às cronologias de Inteligência Artificial transformadora. (Neste caso, "o nascer do sol" seria "O fracasso em desenvolver Inteligência Artificial transformadora, como fizemos no passado".)

Conclusões. Não entrarei muito nos detalhes de como a análise funciona (veja a **postagem no blog resumindo o relatório** para mais detalhes), mas as conclusões do relatório incluem o seguinte:

- O relatório calcula que a probabilidade da inteligência artificial geral (IAG, que incluiria o PASTA) ser desenvolvida até 2036 é de 1 a 18%, com uma melhor estimativa de 8%.

- A probabilidade de isso ocorrer até 2060 é de cerca de 3 a 25% (melhor estimativa de aproximadamente 13%), e a probabilidade da IAG ser desenvolvida até 2100 é de cerca de 5 a 35%, melhor estimativa de 20%.

Elas são menores do que as probabilidades que sugiro acima, mas não *muito* menores. Isso significa **não haver um enorme ônus da prova** ao se considerar evidências adicionais sobre as especificidades do investimento e progresso da IA.

Observações sobre a data de início do processo. Algo interessante aqui é que o **relatório é menos sensível do que se poderia imaginar a respeito de como definir a "data de início" do processo de tentativas de desenvolver a IAG.** (Veja esta seção do relatório completo.) Ou seja:

- Por padrão, a "Priori semi-informativa" modela a situação como se a humanidade tivesse começado a "tentar" construir a IAG em 1956.[80] Isso implica que esse trabalho tem apenas cerca de 65 anos, portanto, as próximas décadas representarão uma grande fração desse esforço.

- O relatório também analisa outros parâmetros de "esforço para construir a IAG" — notadamente, tempo de pesquisa e "computação" (poder de processamento). Mesmo que você queira dizer que estamos implicitamente tentando construir a IAG desde o início da civilização humana, há cerca de

10.000 anos, as próximas décadas conterão uma grande parte do esforço de pesquisa e computação investidos nessa tentativa.

O argumento principal nesta seção

- Ocasionalmente, ouço alguém dizer algo como "Estamos tentando construir uma Inteligência Artificial transformadora há décadas e ainda não conseguimos — por que você acha que no futuro isso será diferente?" No mínimo, este relatório reforça o que vejo como a posição de senso comum de que algumas décadas de "nenhuma Inteligência Artificial transformadora ainda, apesar dos esforços que fizemos para desenvolvê-la" não ajuda muito a argumentar contra a possibilidade de que a Inteligência Artificial transformadora será desenvolvida nas próximas décadas.

- Na verdade, no final das contas, estamos muito próximos de quando as tentativas de desenvolvimento da Inteligência Artificial se iniciaram — **o que é outra forma do nosso século ser "especial"**. Tanto que não devemos nos surpreender se este século acabar sendo um dos mais importantes para o desenvolvimento da IA.

Crescimento econômico

Outra perspectiva sobre o ônus da prova neste sentido é:

Se o PASTA fosse desenvolvido a qualquer momento, e se ele desencadeasse as consequências descritas nesta série de postagens, isso seria uma grande mudança no mundo — e o mundo simplesmente não muda tão rápido.

Para quantificar isso: a economia mundial cresceu alguns por cento ao ano nos últimos 200 anos, e o PASTA implicaria numa taxa de crescimento muito mais rápida, possivelmente 100% ao ano ou acima disso.

*Se **estivéssemos** caminhando para um mundo de crescimento econômico explosivo, o crescimento já deveria estar acelerando hoje. E ele não está acelerando - está se estagnando, pelo menos nas economias mais desenvolvidas. Se a Inteligência Artificial realmente fosse revolucionar tudo, o mínimo que ela poderia estar fazendo atualmente é criando valor suficiente — novos produtos, transações e empresas suficientes — para acelerar o crescimento econômico geral dos EUA.*

A Inteligência Artificial pode levar ao desenvolvimento de novas tecnologias interessantes, mas não há sinal algum do surgimento de algo tão importante quanto o PASTA seria. Partindo do ponto onde estamos e seguindo para onde o PASTA nos levaria está o tipo de mudança repentina que não aconteceu no passado e é improvável que ela aconteça no futuro.

(Se você não estiver familiarizado com o que quero dizer com crescimento econômico, pode ser que você queira ler a minha breve explicação antes de continuar.)

Acho que essa é uma perspectiva razoável e ela me deixa especialmente cético em relação a previsões muito iminentes para Inteligência Artificial transformadora (até 2036 e antes disso).

Minha principal resposta é que a perspectiva de um crescimento constante – "a economia mundial crescendo a uma taxa de alguns por cento ao ano" – fica muito mais complicada quando recuamos e analisamos toda a história econômica, em vez de considerar apenas os últimos séculos. A partir dessa perspectiva, o crescimento econômico tem em sua maior parte acelerado,[81] e projetar essa aceleração para o futuro poderia acarretar um crescimento econômico muito rápido nas próximas décadas.

Escrevi sobre isso anteriormente em O duplicador e Isto não pode continuar; aqui recapitularei brevemente os principais relatórios que citei lá.

***Could Advanced AI Drive Explosive Economic Growth?* (A Inteligência Artificial avançada poderia impulsionar o crescimento econômico explosivo?)** faz explicitamente a pergunta: "Qual é a probabilidade de "crescimento explosivo" - definido como crescimento anual na economia mundial > 30% - até 2100?" O relatório considera argumentos de ambos os lados, incluindo (a) a visão de longo prazo da história que mostra crescimento acelerado; (b) que o crescimento tem sido notavelmente estável nos últimos 200 anos aproximadamente, sugerindo que algo pode ter mudado.

E conclui: "as possibilidades de crescimento de longo prazo estão indefinidas. Tanto o crescimento explosivo quanto a estagnação são plausíveis".

***Modeling the Human Trajectory* (Modelando a trajetória humana)** indaga que futuro podemos esperar se extrapolarmos as tendências existentes ao longo da história econômica. A resposta é um crescimento explosivo até 2043–2065 - não muito longe do que minhas probabilidades acima sugerem. Isso pressupõe que a falta de aceleração econômica nos últimos 200 anos, aproximadamente pode ser uma "anomalia" — que poderá ser resolvida em breve pelo desenvolvimento de uma tecnologia capaz de restaurar a retroalimentação (tema discutido em O Duplicador) que pode fazer com que a aceleração continue.

Para deixar claro, também há boas razões para não dar muita importância para isso como sendo uma projeção,[82] e mais como sendo uma perspectiva sobre o "ônus da prova" do que como uma previsão principal de quando o PASTA será desenvolvido. Há um **debate em aberto** sobre se os dados econômicos anteriores realmente mostram uma aceleração sustentada, em oposição a uma série de períodos muito diferentes com taxas de crescimento crescentes. Discuto como o debate poderia mudar minhas conclusões aqui.

Histórico do "hype em torno da IA"

Outra perspectiva sobre o ônus da prova: às vezes, ouço comentários do tipo "Houve um hype exagerado em torno da Inteligência Artificial muitas vezes no passado, e a Inteligência Artificial transformadora[83] de acordo com aficionados por tecnologia parece sempre estar muito próxima de se tornar realidade". As suas estimativas são apenas as mais recentes nessa longa tradição. Como as estimativas anteriores estavam erradas, as suas provavelmente também estão.

No entanto, não acredito que o histórico do "hype em torno da IA" confirme esse tipo de afirmação.

***What should we learn from past AI forecasts?* (O que devemos aprender com previsões anteriores de IA?)** revisou históricos da Inteligência Artificial para tentar entender qual foi o padrão histórico real do "hype em torno da IA".

Seu resumo dá as seguintes impressões (observe que "HLMI" ou "inteligência artificial de nível humano" é uma ideia bastante semelhante ao PASTA):

- *O auge do hype em torno da Inteligência Artificial parece ter ocorrido entre 1956-1973. Ainda assim, o hype decorrente de algumas das previsões de Inteligência Artificial mais conhecidas desse período costuma ser exagerado.*
- Após 1973, aproximadamente, poucos especialistas pareciam discutir a HLMI (ou algo semelhante) como uma possibilidade de médio prazo, em parte porque muitos especialistas aprenderam com o fracasso do otimismo excessivo anterior da área.
- *O segundo grande período de hype em torno da IA, no início dos anos 1980, parece ter sido mais sobre a possibilidade de "sistemas especialistas" comercialmente úteis e de propósito restrito, não sobre HLMI (ou algo semelhante)…*
- *Não sei se eu teria sido persuadido pelas críticas contemporâneas ao otimismo inicial a respeito da Inteligência Artificial ou se teria considerado fazer o tipo certo de perguntas céticas na época. A crítica mais substancial durante os primeiros anos foi de Hubert Dreyfus, e meu palpite é que eu a teria achado persuasiva na época, mas não posso ter certeza disso.*

Em suma, não é particularmente justo dizer que houve muitas ondas de previsões separadas e excessivamente agressivas sobre Inteligência Artificial transformadora. As expectativas eram provavelmente muito altas no período de 1956 – 1973, mas não acredito que haja muita razão aqui para impor um enorme "ônus de prova" sobre estimativas atuais bem embasadas.

Outras perspectivas sobre o ônus da prova

Aqui estão algumas outras formas possíveis de representar a reação do tipo "Isso é audacioso demais":

Alegações do tipo "Minha causa é muito importante". Muitas pessoas - em todo o mundo atual e ao longo da história - afirmam ou afirmaram que qualquer problema no qual elas estejam trabalhando para resolver é extremamente importante, muitas vezes alegando que possa haver riscos globais ou até mesmo para toda a galáxia. A maioria dessas pessoas estão certamente erradas.

Aqui, acredito que é fundamental indagar se essa afirmação é embasada em argumentos melhores e/ou pessoas mais confiáveis do que são as outras afirmações do tipo "Minha causa é muito importante". Se você se aprofundou na leitura sobre a hipótese do "século mais importante", acredito que já está se colocando em uma boa posição para responder a essa pergunta você mesmo.

A opinião de especialistas será amplamente abordada em postagens futuras. Por enquanto, meu posicionamento principal é que minhas afirmações *nem contradizem* um consenso específico de especialistas, nem são *apoiadas* por um. Elas são, ao contrário, afirmações sobre tópicos que simplesmente não têm "uma área do conhecimento" com especialistas dedicados a estudá-los. Algumas pessoas podem optar por ignorar quaisquer afirmações que não sejam ativamente corroboradas por um consenso robusto de especialistas; mas, considerando o que está em jogo, não acredito que seja isso que deveríamos fazer neste caso.

(Dito isso, o melhor estudo disponível feito por pesquisadores de Inteligência Artificial chegou a conclusões que parecem bastante consistentes com as minhas, como discutirei na próxima postagem.)

Reações não representadas do tipo "Isso é audacioso demais". Tenho certeza de que este artigo não apresentou todas as perspectivas possíveis que poderiam estar subjacentes a uma reação do tipo "Isso é audacioso demais". (Mas não foi por falta de tentativa!) Algumas pessoas simplesmente terão intuições irredutíveis de que as alegações desta série são audaciosas demais para serem levadas a sério.

Uma visão geral dessas perspectivas. Algo que me incomoda sobre a maioria das perspectivas abordadas nesta seção: é que elas parecem **genéricas demais.** Se você simplesmente se recusar (na ausência de evidências esmagadoras) a acreditar em qualquer afirmação que se encaixe em um padrão do tipo "minha causa é muito importante"; ou que ainda não seja corroborada por um consenso robusto de especialistas; ou que simplesmente soe audaciosa demais, isso parece um padrão de raciocínio perigoso. Presumivelmente, algumas pessoas, às vezes, viverão no século mais importante; devemos suspeitar de quaisquer padrões de raciocínio que levem[84] essas pessoas a concluírem que não vivem nele.

Notas

[74]Claro, a resposta poderia ser “Daqui a zilhões de anos” ou “Nunca”.

[75]Tecnicamente, essas probabilidades são para “inteligência artificial geral”, não para Inteligência Artificial transformadora. As probabilidades do surgimento da Inteligência Artificial transformadora podem ser maiores se for possível ter uma Inteligência Artificial transformadora sem inteligência artificial geral, por meio de algo como o PASTA, por exemplo.

[76]Isso corresponde aos dois marcadores do segundo tópico desta seção.

[77]**Wikipedia**: “Medições genéticas indicam que a linhagem dos macacos que levaria ao Homo sapiens divergiu da linhagem que levaria aos chimpanzés e bonobos, os parentes vivos mais próximos dos humanos modernos, cerca de 4,6 a 6,2 milhões de anos atrás. [23] Humanos anatomicamente modernos surgiram na África há cerca de 300.000 anos, [24] e atingiram a modernidade comportamental há cerca de 50.000 anos. [25]”

[78]Por exemplo, enfatiza a probabilidade de se estar entre as “pessoas” mais importantes em vez de “séculos” mais importantes.

[79]Eu não tenho uma fonte única que reúna tudo isso, embora você possa ver **este artigo**. Minha impressão informal ao conversar com pessoas no campo é que o reconhecimento de fala da Inteligência Artificial está pelo menos bem próximo do nível humano, se não, melhor.

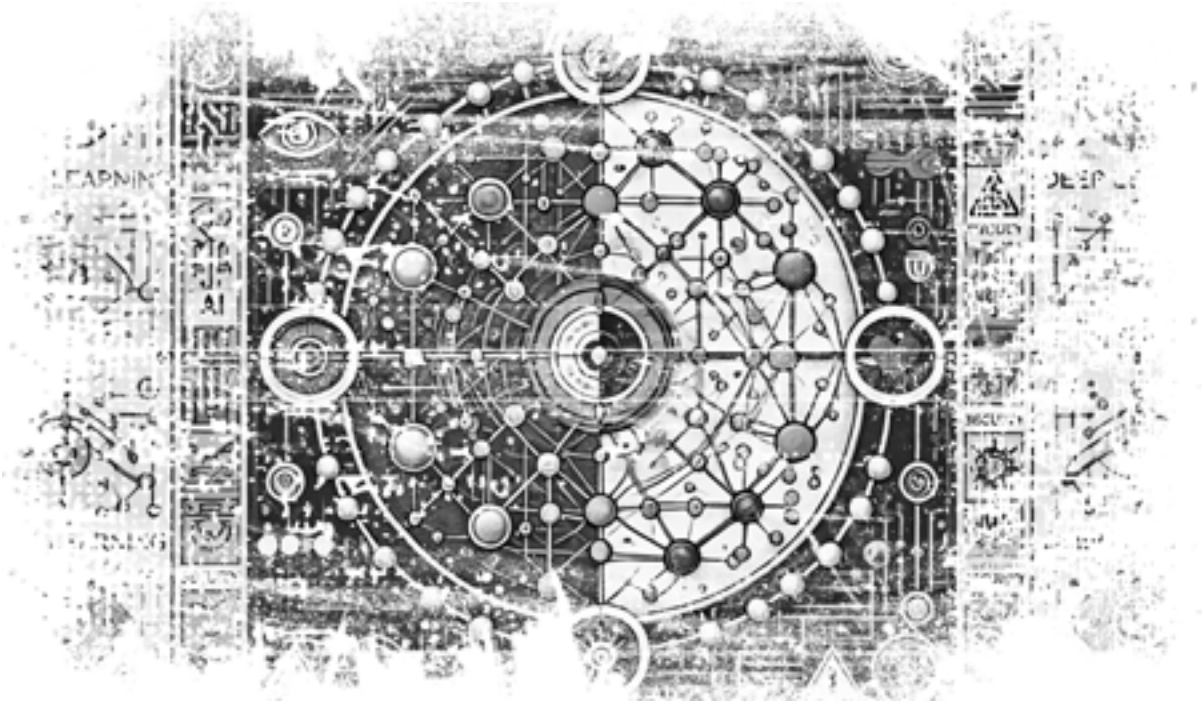
[80]“Considera-se amplamente que o campo de estudos da Inteligência Artificial começou em Dartmouth em 1956”

[81]*Modeling the Human Trajectory* (Modelando a trajetória humana) enfatiza que o modelo que gera esses números “não é flexível o suficiente para acomodar totalmente eventos tão grandes e repentinos quanto a revolução industrial”. O autor acrescenta: “Especialmente porque corresponde imperfeitamente ao passado, sua projeção para o futuro deve ser interpretada livremente, como meramente acrescentando plausibilidade a uma ascensão no próximo século. Davidson (2021) *Could Advanced AI Drive Explosive Economic Growth*” (Poderia a Inteligência Artificial avançada impulsionar crescimento econômico explosivo?) aponta uma maneira importante pela qual as projeções podem continuar erradas por muitas décadas: embora a dinâmica do modelo seja dominada por uma aceleração econômica em espiral, as pessoas ainda são um insumo importante à produção e, ao contrário, o aumento da riqueza levou as pessoas a terem menos e não mais filhos. Nas próximas décadas, isso poderia prejudicar a aceleração prevista, de forma que não poderemos ou não substituiremos trabalhadores por robôs.”

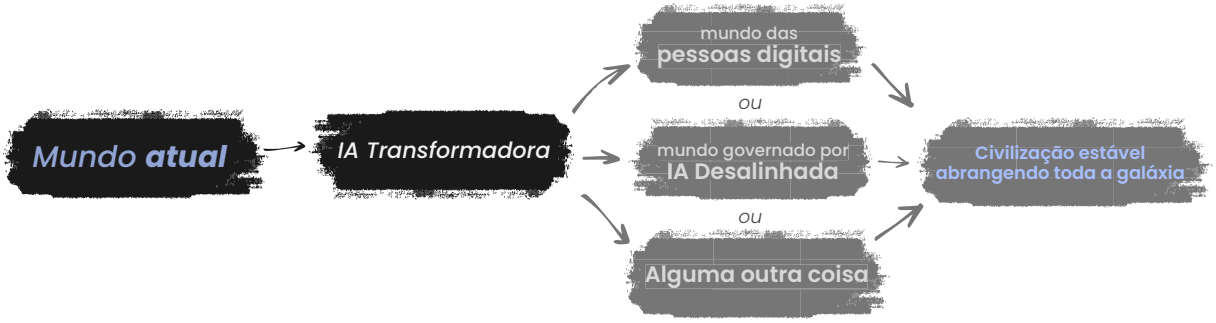
[82]Esses tipos de comentários geralmente se referem a **IGA** e não à Inteligência Artificial transformadora, mas os conceitos são semelhantes o suficiente para que eu os use de forma intercambiável aqui.

[83](Na ausência de evidências esmagadoras, as quais não acredito que devemos presumir que estarão presentes sempre que precisarmos delas).

[84]Claro, a resposta poderia ser “Daqui a um zilhão de anos” ou “Nunca.”



Estamos “Tendendo em direção” à IA transformadora? (Como saberíamos disso?)



Esta é uma de quatro postagens resumindo centenas de páginas de relatórios técnicos focados quase inteiramente na previsão de um número: o ano no qual a Inteligência Artificial transformadora será desenvolvida.[85]

Por “IA Transformadora”, quero dizer “IA poderosa o suficiente para nos levar a um futuro novo e qualitativamente diferente”. Eu me concentro especificamente no que estou chamando de PASTA: sistemas de Inteligência Artificial que conseguem automatizar essencialmente todas as atividades humanas necessárias para acelerar o avanço científico e tecnológico.

Quanto mais cedo o PASTA for desenvolvido, mais cedo o mundo poderá mudar radicalmente, e parece que o mais importante para pensarmos hoje seria como fazer com que essa mudança seja bem ou malsucedida.

Nesta e na próxima postagem, falarei sobre os métodos de previsão subjacentes ao meu posicionamento atual: **acredito que há mais de 10% de chance de desenvolvermos algo suficientemente parecido com o PASTA para ser qualificado como “IA transformadora” dentro de 15 anos (até 2036); uma chance de aproximadamente 50% disso acontecer dentro de 40 anos (até 2060); e uma probabilidade de aproximadamente 2/3 de que isso aconteça ainda neste século (até 2100).**

Abaixo, eu:

- Discutirei **que tipo de previsão estou almejando**.

Não tenho certeza se perceberemos que a Inteligência Artificial transformadora está "a caminho" muito antes de ser desenvolvida. Espero, em vez disso, que tendências a respeito dos principais fatos subjacentes sobre o mundo (como recursos de IA, tamanho do modelo, etc.) sirvam de base para prever um futuro desconhecido qualitativamente.

Uma analogia para esse tipo de previsão seria algo assim: "A água não está borbulhando e não há sinais de borbulhas, mas a temperatura passou de 70° Fahrenheit[86] para 150°, e se ela atingir 212°, a água vai borbulhar." Ou: "É como prever o fechamento de escolas e a superlotação de hospitais, antes que aconteça, baseando-se nas tendências de aumento de infecções".

- Discutirei se é possível observar tendências que indiquem o quanto os sistemas de Inteligência Artificial são capazes ou "surpreendentes". Acho que essa abordagem não é confiável: (a) o progresso da Inteligência Artificial pode não apresentar uma "tendência" da maneira que esperamos; (b) na minha experiência, diferentes pesquisadores de Inteligência Artificial têm intuições radicalmente diferentes sobre quais sistemas são surpreendentes ou capazes e como estão progredindo.

- Discutirei brevemente **Grace et al. 2017**, a melhor pesquisa feita com pesquisadores de Inteligência Artificial sobre cronologias da Inteligência Artificial transformadora. Suas conclusões parecem estar amplamente alinhadas com minhas próprias previsões, embora haja sinais de que os participantes da pesquisa não estavam pensando muito antes de responder às perguntas.

O próximo artigo nesta série focará no artigo ***Forecasting Transformative AI with Biological Anchors* (Prevendo a Inteligência Artificial Transformadora com âncoras biológicas) de Ajeya Cotra**" (que me referirei abaixo como "Bio-âncoras"), o método de previsão para Inteligência Artificial transformadora que considero mais esclarecedor.

Que tipo de previsão almejo?

De algumas maneiras, prever a Inteligência Artificial transformadora é diferente de realizar as previsões que estamos acostumados a realizar.

Primeiramente, essas previsões abrangem intervalos de tempo muito longos (décadas), ao contrário de, por exemplo, uma previsão climática (dias) ou uma previsão eleitoral (meses). Isso torna a tarefa um pouco mais difícil,[87] e também mais difícil para quem está de fora avaliar, já que não há um **histórico** claramente relevante de previsões sobre tópicos semelhantes.

Em segundo lugar, carecemos de fontes de dados ricas e claramente relevantes e não podemos nos basear em um monte de previsões semelhantes feitas no passado. A *FiveThirtyEight* analisa centenas de pesquisas **eleitorais** e tem um modelo de quão bem elas previram os resultados das eleições passadas. Porém, a previsão da Inteligência Artificial transformadora se ancora mais em intuições, suposições e julgamentos, para determinar quais dados são mais relevantes e como eles são relevantes.

Finalmente, estou tentando prever um futuro desconhecido qualitativamente. A Inteligência Artificial transformadora — e o estranho futuro que a acompanha — não se parece com algo para o qual estamos "tendendo na direção", ano a ano.

- Se eu estivesse tentando prever quando a população mundial chegaria a 10 bilhões, poderia simplesmente extrapolar as tendências existentes da população mundial. A própria população mundial é conhecida por estar crescendo e pode ser estimada diretamente. Na minha opinião, extrapolar uma tendência de longa duração é uma das melhores maneiras de fazer uma previsão.

- Quando a *FiveThirtyEight* faz previsões eleitorais, há um entendimento prévio de que haverá eleição em uma determinada data, e quem vencer tomará posse em outra data. Todos aceitamos esse modelo e há um entendimento geral de que uma votação maior significa maior probabilidade de vitória.

- Por outro lado, a Inteligência Artificial transformadora — e o estranho futuro que ela traz — não é uma tecnologia para a qual estamos "caminhando" de maneira claramente mensurável. Não há uma métrica clara como "transformatividade da IA" ou "estranheza do mundo" que aumenta regularmente a cada ano, de modo que possamos projetá-la no futuro e obter a data em que algo como o PASTA será desenvolvido.

Talvez para alguns, esses sejam motivos suficientes para ignorar toda a possibilidade de uma Inteligência Artificial transformadora ou presumir que ela está muito distante. Mas, por alguns motivos, não acredito que isso seja uma boa ideia.

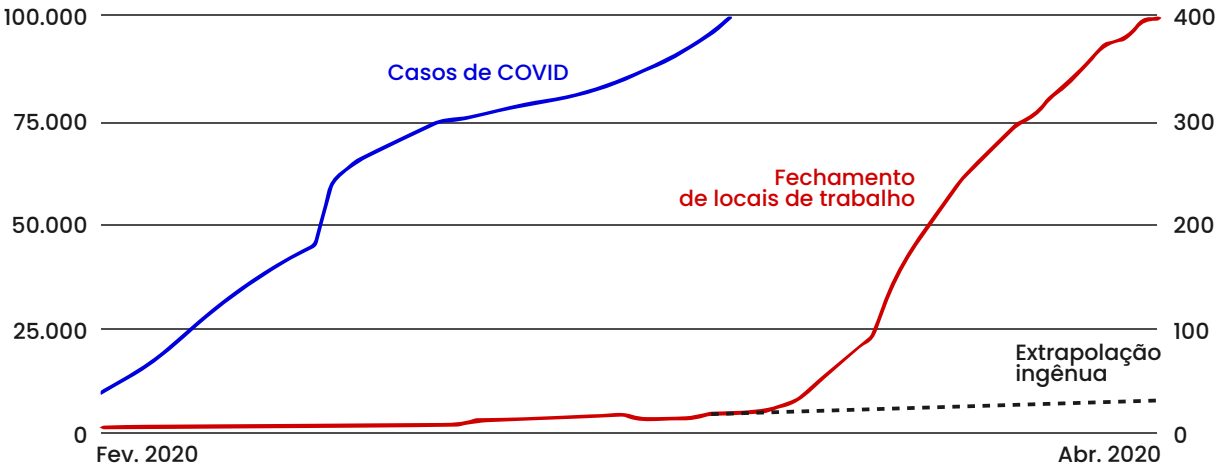
Em primeiro lugar, tenho uma opinião subjacente de que algo como o PASTA é, de um certo modo, "inevitável", presumindo que haverá avanços contínuos na sociedade e na computação. A intuição básica aqui - que eu poderia expandir se houver **interesse** - é que os cérebros humanos são numerosos e não parecem precisar de materiais raros específicos para serem produzidos, então, deve ser possível em algum momento replicar sinteticamente as principais partes de sua funcionalidade.[88]

Por sua vez, não estou confiante de que o PASTA parecerá estar qualitativamente "a caminho" de ser desenvolvido bem antes de estar. (Mais sobre isso abaixo.) Portanto, estou inclinado a procurar maneiras de estimar para quando podemos esperar esse desenvolvimento, apesar dos desafios e do fato de que hoje não parece que isso acontecerá em breve.

Acho que há muitos exemplos de casos em que **um futuro desconhecido qualitativamente poderia ser vislumbrado com antecedência, traçando a tendência com base em alguns fatos subjacentes e relacionados sobre o mundo.** Alguns exemplos que vêm à mente:

- Quando a COVID-19 surgiu pela primeira vez, muitas pessoas não a levaram a sério porque não parecia estarmos "tendendo para" ou "na direção de" um mundo cheio de hospitais superlotados, fechamento de escritórios e escolas, etc. Na época (digamos, janeiro de 2020), havia um número relativamente pequeno de casos, um número ainda menor de mortes e nenhum entendimento qualitativo de que isso era uma emergência global. A princípio, a única coisa alarmante sobre a COVID-19, era que o número de casos estava crescendo a uma rápida taxa exponencial (embora o número total de casos ainda fosse pequeno). Mas foi possível extrapolar do rápido crescimento no número de casos para o risco de uma emergência global, e **algumas pessoas conseguiram**. (Enquanto **outras não**.)

Os climatologistas preveem um aumento global nas temperaturas que é significativamente maior do que o observado nas últimas décadas, e isso pode ter consequências significativas muito além das quais estamos observando atualmente. Isso é feito prevendo tendências nas emissões de gases de efeito estufa e extrapolando *a partir disso* para a temperatura e suas consequências. Se você simplesmente perguntasse "Com que rapidez a temperatura está subindo?" ou "Os furacões estão piorando?", e baseasse todas as suas previsões do futuro nas respostas dessas perguntas, você provavelmente não estaria prevendo os mesmos tipos de eventos extremos por volta de 2100.[89]



Casos de COVID-19 reportados pela OMS. O fechamento de locais de trabalho refere-se **a estes dados da OWID (Nosso mundo em dados)**, simplesmente pontuados como 1 para "recomendado", 2 para "obrigatório para alguns", 3 para "obrigatório para todos, exceto os principais trabalhadores" e somados em todos os países.

Uma analogia para esse tipo de previsão seria algo assim: "Esta água não está borbulhando e não há sinais de borbulhas, mas a temperatura passou de 70° Fahrenheit[90] para 150°, e se ela atingir 212°, a água vai borbulhar."

Preferencialmente, posso encontrar alguns fatores subjacentes que se alteram regularmente o suficiente para podermos predizê-los (tal como o crescimento no **tamanho e custo dos modelos de IA)**, e, em seguida, argumentar que, se esses fatores atingirem um certo ponto, as chances do desenvolvimento da Inteligência Artificial transformadora serão altas.

Podemos pensar nessa abordagem como uma resposta para a pergunta: "Se eu acredito que algo como o PASTA é inevitável e estou tentando prever o momento em que ele será desenvolvido usando alguns métodos de análise diferentes, qual é a minha previsão?" Podemos perguntar separadamente "E há razão para que essa previsão seja implausível, não confiável ou muito audaciosa"? - isso foi abordado no artigo anterior desta série.

Extrapolações subjetivas e "imponência da IA"

Para uma apresentação diferente de um conteúdo parecido, veja esta seção de Bio-âncoras.

Se buscássemos fatores que influenciariam na previsão de quando uma Inteligência Artificial transformadora seria desenvolvida, talvez a primeira coisa a se analisar seriam tendências que mostrem o quão "surpreendentes" ou "capazes" os sistemas de Inteligência Artificial estão se tornando.

A versão mais fácil disso seria se o mundo se reestruturasse de tal forma que:

- Um dia, pela primeira vez, um sistema de Inteligência Artificial conseguisse ser aprovado no exame de ciências da 4ª série.
- Em seguida, a primeira Inteligência Artificial fosse aprovada (e depois recebesse a nota máxima) em um exame da 5ª série, depois no exame da 6ª série, etc.
- Depois, a primeira Inteligência Artificial conquistasse um doutorado e tivesse seu primeiro artigo publicado, etc., chegando finalmente a ser a primeira Inteligência Artificial que conseguisse realizar um trabalho científico digno de receber o Prêmio Nobel.
- Isso teria que ser distribuído regularmente ao longo de décadas, para podermos observar claramente o estado da arte avançando da 4ª para a 5ª e 6ª séries, alcançando o "pós-doutorado" e além. E tudo isso acontecesse lenta e regularmente o suficiente para ser possível começar a definir uma data para a "IA científica total" com várias décadas de antecedência.

Seria muito conveniente — quase quero dizer "comportado" — se os sistemas de Inteligência Artificial evoluíssem dessa maneira. Também seria "comportado" se a Inteligência Artificial avançasse da maneira que algumas pessoas casualmente imaginam: primeiro assumindo funções como "motorista de caminhão" e "trabalhador de linha de montagem", depois funções como "professor" e "suporte de TI," e depois empregos como "médico" e "advogado", antes de progredir para "cientista".

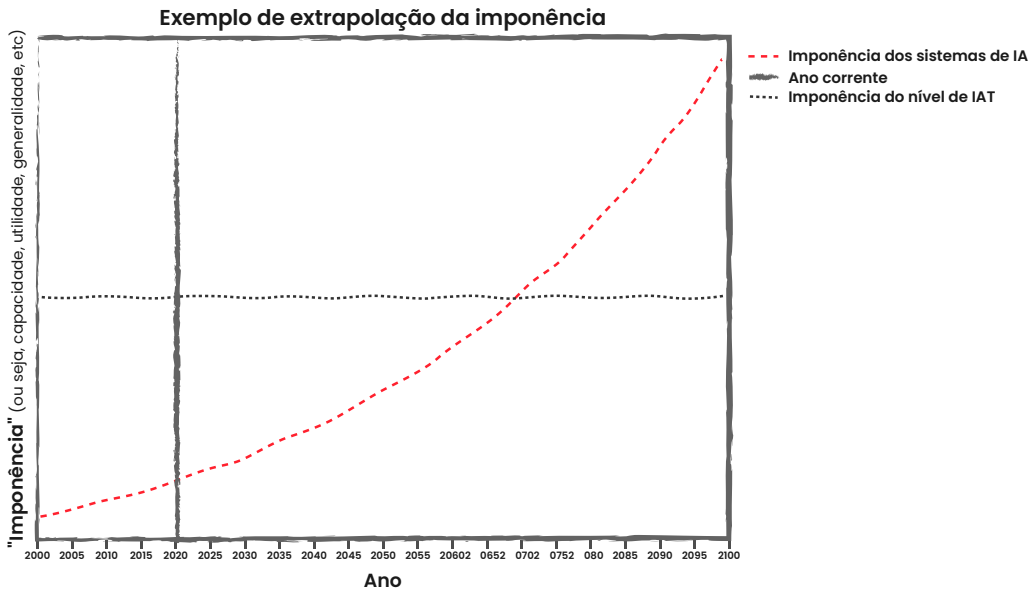
Qualquer um dos cenários acima nos daria bastante tempo de vantagem e uma base sólida para prever quando a Inteligência Artificial de automação científica será desenvolvida. Infelizmente, acredito que não podemos contar com isso.

- A Inteligência Artificial parece progredir de maneira muito diferente dos humanos. Por exemplo, existiam jogadores de xadrez de Inteligência Artificial sobre-humanos[91] muito antes de existir uma Inteligência Artificial que pudesse diferenciar com segurança imagens de cães e gatos.[92]
- Uma possibilidade é que os sistemas de Inteligência Artificial conseguirão realizar tarefas intelectuais mais difíceis do que as que os insetos conseguem realizar. Depois, as tarefas mais difíceis que os ratos e outros pequenos mamíferos conseguem realizar, depois as dos macacos, depois as dos humanos — se igualando efetivamente com as habilidades de cérebros crescentes. Se isso acontecesse, não observaríamos necessariamente muitos sinais de que a Inteligência Artificial conseguiria, por exemplo, fazer ciência até que estivéssemos *muito próximos* disso acontecer. Alcançar o mesmo nível que um aluno da 4ª série pode não acontecer até o final do processo.
- Outra possibilidade é que os sistemas de Inteligência Artificial consigam fazer qualquer coisa que um humano faça em 1 segundo, depois qualquer coisa que um humano faça em 10 segundos, etc. Isso também seria uma progressão bastante confusa, o que tornaria prever o progresso algo não-óbvio.

Na verdade, se ainda não soubéssemos como os humanos se desenvolvem, acharíamos que o progresso de uma criança é bastante confuso e difícil de extrapolar. **Acompanhar o desenvolvimento de uma pessoa, do seu nascimento até os 8 anos, não daria nenhuma indicação de que essa pessoa estaria, digamos, a 1/3 do caminho para fundar seu próprio negócio, realizar uma importante descoberta científica original, etc.** (Mesmo conhecendo o curso normal do desenvolvimento humano, é difícil dizer, observando uma criança de 8 anos, quais habilidades de nível profissional ela desenvolveria ou teria na idade adulta.)

No geral, não está claro como devemos interpretar o espectro de “não surpreendente/capaz” até “muito surpreendente/capaz” para IA. E, de fato, na minha experiência, diferentes pesquisadores de Inteligência Artificial têm intuições radicalmente diferentes sobre quais sistemas são surpreendentes ou capazes e como eles estão progredindo. Muitas vezes tive a experiência de ver um amigo pesquisador de Inteligência Artificial apontar para algum novo resultado e dizer “Isso é enorme, como alguém pode não ver como estamos próximos do desenvolvimento de uma Inteligência Artificial poderosa?” enquanto outro diz “Este é um pequeno avanço, de pouca importância.”[93]

Seria ótimo se pudéssemos prever o ano em que a Inteligência Artificial transformadora será desenvolvida, usando um gráfico como este (de **Bio-âncoras**; “TAI” significa “IA transformadora”):



Mas, até onde posso dizer, não há como definir o eixo y sem que antes isso seja debatido ferozmente entre os especialistas.

Consultando os especialistas

Uma maneira de lidar com essa incerteza e confusão seria consultar um grupo grande de especialistas e simplesmente perguntar quando eles esperam que a Inteligência Artificial transformadora seja desenvolvida. Podemos ter a esperança de que cada um dos especialistas consultados (ou pelo menos, muitos deles) estejam fazendo sua própria versão da “extrapolação da imponência da IA” mencionada acima. Se não for esse o caso, que eles estejam fazendo outra coisa que os ajude a chegar a uma estimativa razoável.

Calculando a média de muitas estimativas, podemos obter um agregado que reflita a “sabedoria das multidões”.[94]

Grace et al 2017 apresenta a melhor versão desse exercício. A pesquisa, feita com 352 pesquisadores de IA, incluía uma pergunta sobre “quando máquinas independentes conseguirão realizar todas as tarefas que os trabalhadores humanos realizam, de forma mais barata e melhor” (o que presumivelmente incluiria tarefas que promovem o desenvolvimento científico e tecnológico, e, portanto, qualificaria como PASTA). As duas grandes conclusões desta pesquisa, segundo o **Bio-âncoras** e eu, são:

- **Uma probabilidade de aproximadamente 20% desse tipo de Inteligência Artificial ser desenvolvida até 2036; aproximadamente 50% até 2060 e aproximadamente 70% até 2100. Estes resultados correspondem aos números que cito na introdução.**
- Perguntas reformuladas com palavras ligeiramente diferentes das originais (feitas para um subconjunto menor de entrevistados) tiveram estimativas muito posteriores como resposta. Isso seria (para mim), uma indicação de que os participantes simplesmente não estavam pensando muito antes de responder à pesquisa.[95]

Minha conclusão: essa evidência é consistente com minhas probabilidades atuais, embora não tão esclarecedora assim. O próximo artigo nesta série focará completamente no ***Forecasting Transformative AI with Biological Anchors*, (Prevendo a Inteligência Artificial Transformadora com âncoras biológicas) de Ajeya Cotra**, o método de previsão que considero mais elucidativo mencionado aqui.

Notas

[84]Claro, a resposta poderia ser “Daqui a um zilhão de anos” ou “Nunca.”

[85]Equivalentes em centígrados para esta frase: 21 °C, 66

[86]Algumas observações sobre previsões de

[87]Para um argumento um pouco mais elaborado, veja também **este artigo**. Não concordo totalmente com o que ele apresenta; não acredito que seja um argumento forte no que se refere ao desenvolvimento da Inteligência Artificial transformadora no curto prazo, mas acredito que oferece uma boa aproximação das minhas intuições sobre sua viabilidade. Também veja ***On theImpossibilityof Supersized Machines* (Sobre a impossibilidade de máquinas superdimensionadas)**para algumas respostas implícitas (brincadeira) para muitos argumentos comuns sobre porque a Inteligência Artificial transformadora pode ser impossível de criar.

[88]Por exemplo, observe o gráfico de temperatura aqui- a linha mais baixa seria uma projeção razoável, se a temperatura fosse a única coisa que você estivesse observando.

[89]Equivalentes em centígrados para esta frase: 21° C, 66° C, 100° C

[91]1997

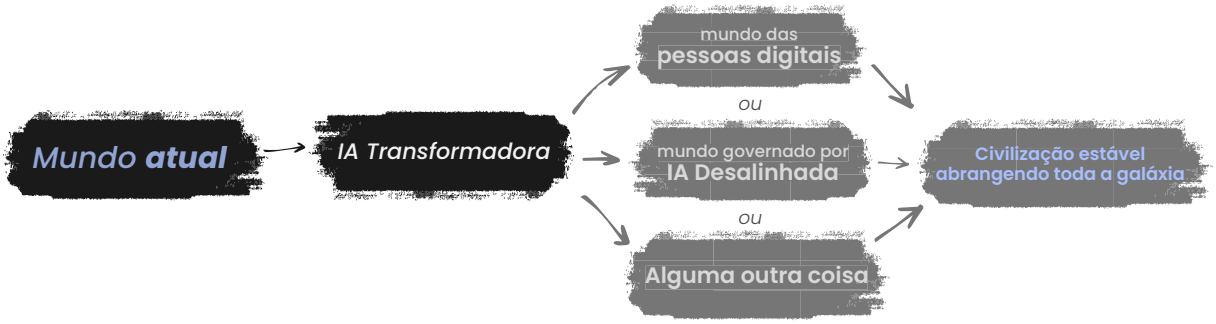
[92] O desafio kagle “cães x gatos” foi criado em 2013

[93]**De bio-âncoras**: “Ouvimos especialistas em ML com cronologias relativamente curtas argumentarem que os sistemas de Inteligência Artificial atuais podem essencialmente ver tão bem quanto os humanos; entender informações escritas e vencer os humanos em quase todos os jogos de estratégia. E o conjunto de coisas que eles conseguem fazer está se expandindo rapidamente, levando-os a antecipar que o desenvolvimento da Inteligência Artificial transformadora será alcançável nas próximas décadas, treinando modelos maiores em uma distribuição mais ampla de problemas de ML que são mais direcionados à geração de valor econômico. Por outro lado, muito mais dados para aprender do que os humanos, são incapazes de transpor o que aprendem de um contexto para outro ligeiramente diferente e não parecem ser dotados de um raciocínio muito lógico, estruturado e causal; isso os leva a acreditar que precisaríamos de vários avanços importantes para desenvolver a TAI. Pelo menos um consultor técnico da Open Philanthropy avançou cada uma dessas perspectivas.”

[94]**Wikipedia**: “A descoberta clássica da sabedoria das multidões... Em uma feira rural de 1906 em Plymouth, 800 pessoas participaram de um concurso para estimar o peso de um boi abatido e preparado. O estatístico Francis Galton observou que o palpite médio, 1207 libras (0,55 t), atingiu uma acurácia próxima de 1% do peso real de 1198 libras (0,54 t).

[95]**Bio-âncoras**: *Alguns pesquisadores foram solicitados a prever o desenvolvimento da “HLMI” conforme definida acima (inteligência artificial de nível humano, a qual eu consideraria incluir algo como o PASTA); enquanto outro subconjunto de pesquisadores foi solicitado a prever o desenvolvimento da “automação total do trabalho”, ou seja, uma época na qual “todas as ocupações serão totalmente automatizáveis”. Apesar do desenvolvimento da HLMI parecer conduzir rapidamente à automação total do trabalho, a estimativa média para o desenvolvimento da automação total do trabalho foi de aproximadamente 2138, enquanto a estimativa média para a HLMI foi de aproximadamente 2061, quase 80 anos mais cedo. Subconjuntos aleatórios de entrevistados foram solicitados a prever quando marcos individuais (por exemplo, dobrar roupas lavadas, StarCraft ao nível humano ou pesquisa matemática ao nível humano) seriam alcançados. O ano médio em que os entrevistados esperavam que as máquinas automatizassem a pesquisa de Inteligência Artificial foi até aproximadamente 2104, enquanto a estimativa média para a HLMI foi até aproximadamente 2061 — outra inconsistência clara, pois a “pesquisa de IA” é uma tarefa realizada por trabalhadores humanos.*

Prevendo a IA transformadora: O método das âncoras biológicas em poucas palavras



Esta é uma de quatro postagens resumindo centenas de páginas de relatórios técnicos focados quase inteiramente na previsão de um número: o ano no qual a Inteligência Artificial transformadora será desenvolvida.[96]

Por "IA Transformadora", quero dizer "IA poderosa o suficiente para nos levar a um futuro novo e qualitativamente diferente". Eu me concentro especificamente no que estou chamando de PASTA: sistemas de Inteligência Artificial que conseguem automatizar essencialmente todas as atividades humanas necessárias para acelerar o avanço científico e tecnológico.

Quanto mais cedo o PASTA for desenvolvido, mais cedo o mundo poderá mudar radicalmente, e parece que o mais importante para pensarmos hoje seria como fazer com que essa mudança seja bem ou malsucedida.

Esta postagem é um resumo para leigos do artigo ***Forecasting Transformative AI with Biological Anchors* (Prevendo a Inteligência Artificial Transformadora com âncoras biológicas) de Ajeya Cotra**, e seus prós e contras.[97] É a previsão que considero mais esclarecedora para Inteligência Artificial transformadora, com algumas ressalvas:

- Essa abordagem é relativamente complexa e requer inúmeras suposições e estimativas incertas. Essas características tornam-na relativamente difícil de explicar e, também, depõem contra a confiabilidade do método.
- Portanto, atualmente, não acredito que esse método seja tão confiável quanto os exemplos que dei anteriormente para prever um futuro qualitativamente diferente. Ele não é tão simples ou objetivo como alguns desses exemplos, tais como os modelos para prever a disseminação da COVID-19. E embora a modelagem climática também seja muito complexa, ela foi desenvolvida por muitos especialistas ao longo de décadas, enquanto a metodologia das Bio-âncoras não tem um histórico longo.

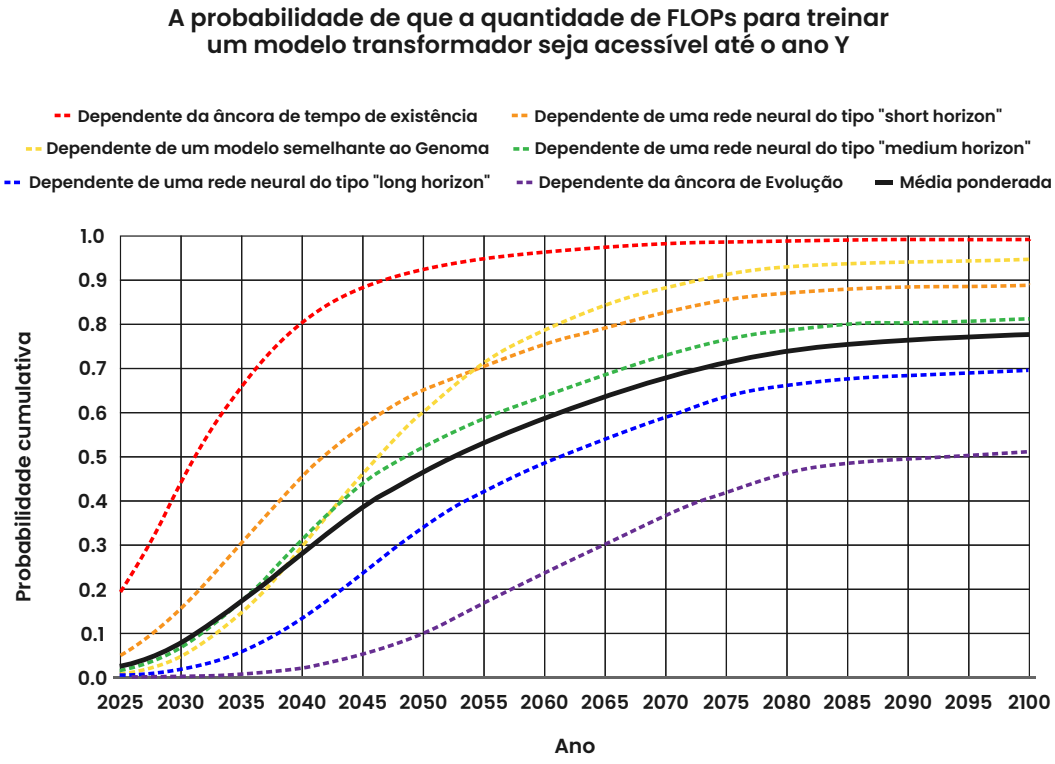
No entanto, acho que este é, atualmente, o melhor método disponível para dar um "palpite" no que se refere às cronologias/linhas do tempo da Inteligência Artificial transformadora. E como discutido na seção final, pode-se **deixar de lado muitos dos detalhes para constatar que este século, provavelmente, será testemunha de algumas das conquistas mais "extremas" mencionadas pelo relatório, que sugere fortemente a viabilidade da Inteligência Artificial transformadora.**

(Observação: também escrevi uma postagem posterior sobre essa teoria para leitores céticos. Veja "Âncoras biológicas" é sobre delimitar e, não apontar, cronologias da IA.)

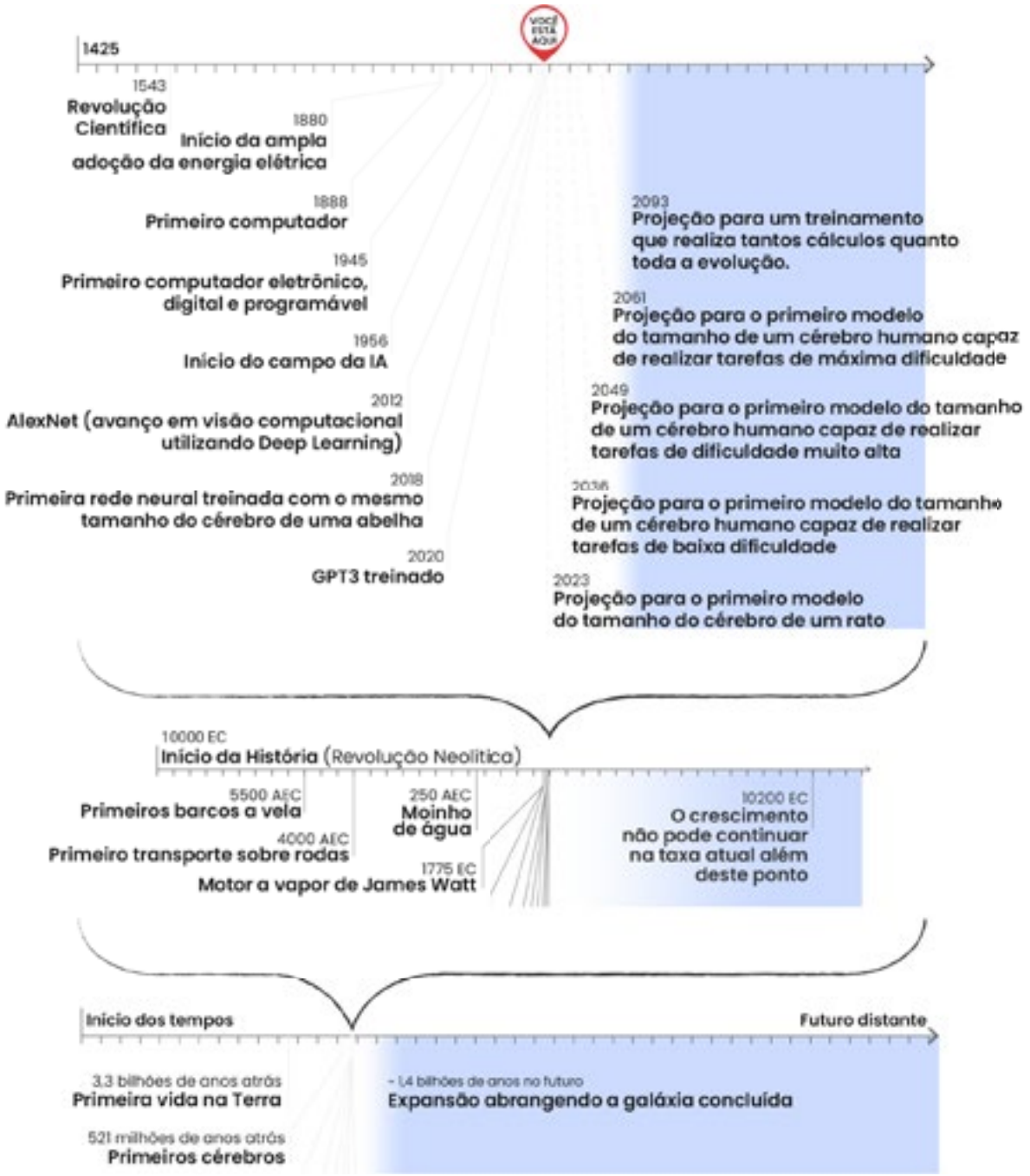
A ideia básica é:

- Os modelos modernos de Inteligência Artificial podem "aprender" a realizar tarefas por meio de um processo (financeiramente caro) conhecido como "treinamento". O treinamento pode ser descrito como uma quantidade enorme de ações de tentativa e erro. Por exemplo, os modelos de Inteligência Artificial de reconhecimento de voz recebem um arquivo de áudio de alguém falando, dão um palpite sobre o que a pessoa está dizendo e, então, recebem a resposta correta. Ao fazer isso milhões de vezes, eles "aprendem" a traduzir de forma confiável a fala em texto. Mais: Treinamento
- Quanto maior for um modelo de Inteligência Artificial e mais complexa for a tarefa, mais caro será o processo de treinamento. Alguns modelos de Inteligência Artificial são maiores que outros; até o momento, nenhum deles chega nem perto de ser "tão grande quanto o cérebro humano" (abaixo explicarei o que isso quer dizer). Mais: Tamanho do modelo e tipo de tarefa
- O método das âncoras biológicas questiona: **"Com base nos padrões usuais de custo de treinamento, quanto custaria treinar um modelo de Inteligência Artificial tão grande quanto um cérebro humano para executar as tarefas mais difíceis que os humanos conseguem fazer? E, quando isso será barato o suficiente para que alguém faça esse treinamento?"** (Mais:) Estimando os custos

O método das âncoras biológicas modela muitas maneiras de se abordar essa questão, gerando uma ampla gama de estimativas, de "agressivas" (desenvolvimento da Inteligência Artificial transformadora mais cedo) a "conservadoras" (desenvolvimento mais tarde). Mas, a partir de praticamente todas essas perspectivas, ela prevê, com alta probabilidade, o desenvolvimento da Inteligência Artificial transformadora ainda neste século.



Este gráfico é do relatório. O eixo Y representa, aproximadamente, a probabilidade da Inteligência Artificial transformadora ser desenvolvida no ano em questão, embora haja algumas nuances adicionais no relatório. Não explicarei o que significa cada um dos diferentes modelos "Dependente de"; basta saber que cada um deles representa uma perspectiva diferente na previsão da Inteligência Artificial transformadora.



Obrigado a María Gutiérrez Rojas por este gráfico. A cronologia superior apresenta conquistas importantes para a computação de IA, no passado e para o futuro (as conquistas futuras são projetadas pelo Bio-âncoras). Abaixo estão (cortadas) outras cronologias mostrando o quão importantes este intervalo de algumas centenas de anos (mais em Isto não pode continuar), e esta era (mais em Todas as visões possíveis sobre o futuro da humanidade são audaciosas), parecem ser.

Agora me aprofundarei sobre cada um deles um pouco mais. Esta é a parte mais densa desta série, e algumas pessoas podem preferir apenas ler o resumo acima e pular para a próxima postagem.

Observe que o método das Bio-âncoras usa várias abordagens diferentes (que ele chama de "âncoras") para estimar cronologias da Inteligência Artificial transformadora e as combina em uma só visão agregada. Neste resumo, estou mais focado em um conjunto específico delas - chamado de "âncoras da rede neural" - que impulsiona a maioria das cronologias agregadas do relatório. Parte do que digo se aplica a todas as âncoras, mas parte se aplica apenas às "âncoras da rede neural".

Treinamento

Como discutido anteriormente, existem, essencialmente, duas maneiras de "ensinar" um computador a realizar uma tarefa:

1. **"Programar" o computador com instruções passo a passo extremamente específicas para concluir a tarefa.** Quando isso é possível, o computador executa geralmente as instruções de forma muito rápida, confiável e barata. Por exemplo, você pode programar um computador para examinar cada registro em um banco de dados e imprimir os que correspondem aos termos de pesquisa de um usuário — você o "instruiria" exatamente como fazer isso e ele conseguiria realizar a tarefa muito bem.

2. **"Treinar" uma Inteligência Artificial para realizar a tarefa puramente por tentativa e erro.** Hoje, a maneira mais comum de fazer isso é usando uma "rede neural", que seria como um "cérebro digital" que está em um estado inicial vazio (ou aleatório): que ainda não foi programado para realizar tarefas específicas. Por exemplo, digamos que queremos que uma Inteligência Artificial reconheça fotos de cães e de gatos. É difícil dar instruções passo a passo totalmente específicas para isso; em vez disso, podemos alimentar uma rede neural com um milhão de imagens de exemplo (cada uma rotulada como "cachorro" ou "gato"). Cada vez que a Inteligência Artificial processa um exemplo, ela se ajusta internamente para aumentar a probabilidade de acertar em casos semelhantes no futuro. Após exemplos suficientes, ela estará configurada corretamente para reconhecer cães de gatos.

(Talvez pudéssemos subir o nível e treinar modelos para aprender com o próprio treinamento da maneira mais eficiente possível. Isso é chamado de meta-aprendizagem, mas pelo que sei esse método ainda não alcançou bons resultados.) O treinamento é uma espécie de alternativa mais cara e de força bruta do que a "programação". A vantagem é que não precisamos fornecer instruções específicas, podemos apenas alimentar a Inteligência Artificial com muitos exemplos de como realizar uma tarefa corretamente, e ela aprenderá a realizá-la sozinha.

- A desvantagem é que precisamos de ***muitos*** **exemplos, requerendo muito poder de processamento, com alto custo.**

Quanto? Isso depende do tamanho do modelo (rede neural) e da natureza da própria tarefa. Para algumas tarefas que as IAs já aprenderam, até 2021, o treinamento de um único modelo pode custar milhões de dólares. Para tarefas mais complexas (como "desenvolver pesquisa científica inovadora") e modelos maiores (do tamanho do cérebro humano), treinar um modelo pode custar muito mais do que isso.

O método das Bio-âncoras está interessado na questão: **"Quando será econômico treinar um modelo para realizar as tarefas mais difíceis que os humanos conseguem realizar, usando uma abordagem relativamente rudimentar baseada em tentativa e erro?"**

Essas tarefas podem incluir as tarefas necessárias para o PASTA, tais como:

- Aprender sobre ciência com professores, livros didáticos e trabalhos de casa de maneira mais eficaz que os humanos.

- Ultrapassar a fronteira da ciência fazendo perguntas, análises e escrevendo artigos de maneira mais eficaz que os humanos.

A próxima seção discutirá como o método das Bio-âncoras aprimora a ideia das "tarefas mais difíceis que os humanos conseguem realizar" (quais tarefas exigiriam um modelo do "tamanho do cérebro humano").

Tamanho do modelo e tipo de tarefa

A hipótese do método das Bio-âncoras é que podemos estimar "o quanto custa treinar um modelo" com base em dois parâmetros básicos: o tamanho do modelo e o tipo de tarefa.

Tamanho do modelo. Conforme mencionado acima, uma rede neural seria como um "cérebro digital" que está em um estado inicial vazio ou aleatório. Em geral, um "cérebro digital" *maior* — com mais versões-digitais-de-neurônios e versões-digitais-de-sinapses[98] - aprenderia tarefas mais complexas. Um "cérebro digital" maior também requer mais cálculos — e, consequentemente, seria mais caro — cada vez que ele for usado (por exemplo, a cada exemplo fornecido durante o treinamento).

Com base na análise descrita em ***How Much Computational Power Does It Take to Match the Human Brain?* (Quanto poder computacional é necessário para igualar o cérebro humano?) de Joe Carlsmith** (ao qual me refiro neste artigo como "Computação cerebral"), o método das Bio-âncoras estima comparações entre o tamanho de cérebros digitais (modelos de IA) e cérebros de animais (cérebros de abelhas, cérebros de camundongos, cérebros humanos).

Essas estimativas implicam que **os sistemas de Inteligência Artificial atuais são, às vezes, tão grandes quanto os cérebros de insetos, mas nunca tão grandes quanto os cérebros de camundongos —** até o momento em que isto foi escrito, o maior modelo de linguagem conhecido foi o primeiro a chegar razoavelmente perto desse tamanho.[99] E ainda não tem **nem 1% do tamanho do cérebro humano.**[100]

Quanto maior for o modelo, mais poder de processamento será necessário para treiná-lo. O método das Bio-âncoras presume que um modelo de Inteligência Artificial transformadora precisaria ter cerca de 10 vezes o tamanho de um cérebro humano, muito maior do que qualquer modelo atual de IA. (O tamanho de 10 vezes maior foi escolhido para deixar algum espaço para a ideia de que "cérebros digitais" podem ser menos eficientes que os cérebros humanos; veja esta seção do relatório.) Esta é uma das razões pelas quais o treinamento seria muito caro.

Mas pode acontecer de um modelo de Inteligência Artificial menor ainda aprender os tipos de tarefas mencionadas acima. Ou pode ser que o tamanho do modelo necessário seja maior do que as estimativas geradas pelo método das Bio-âncoras, talvez porque ele tenha subestimado o "tamanho" efetivo do cérebro humano, ou porque o cérebro humano é melhor projetado do que os "cérebros digitais" para além do que o método supõe.

Tipo de tarefa. Para aprender uma tarefa, um modelo de Inteligência Artificial precisa efetivamente "tentar" realizar a tarefa (ou "assistir" à tarefa sendo realizada) inúmeras vezes, aprendendo por tentativa e erro. Quanto mais caro (em poder de processamento e, portanto, dinheiro) for tentar realizar a tarefa (ou assistir à tarefa sendo realizada), mais caro será aprendê-la.

É difícil quantificar o custo de se tentar realizar/ou assistir uma tarefa sendo realizada. A tentativa do método das Bio-âncoras de fazer isso é a parte mais controversa da análise, segundo os revisores técnicos do método até agora.

Podemos pensar no método das Bio-âncoras como dizendo:

- Existem algumas tarefas que um ser humano faz com apenas um segundo de pensamento, como, por exemplo, reconhecer imagens de gatos ou cachorros.
- Existem outras tarefas que levam vários minutos para um ser humano pensar, tal como, resolver um quebra-cabeça lógico.
- Outras tarefas podem levar horas, dias, meses ou até anos e exigem não apenas pensar, mas interagir com o ambiente. Como, por exemplo, escrever um artigo científico.
- As tarefas na extremidade mais longa desse espectro serão mais caras de se tentar realizar/assistir, portanto, será mais caro treinar um modelo de Inteligência Artificial para realizar essas tarefas. Por exemplo, é mais caro (leva mais tempo e mais dinheiro) realizar um milhão de "tentativas" de uma tarefa com uma hora de duração, do que um milhão de "tentativas" de realizar uma tarefa que leva apenas um segundo.
- No entanto, a teoria não é tão simples quanto parece. Muitas tarefas que parecem "longas" (como escrever uma redação) podem, na verdade, ser divididas em uma série de tarefas "mais curtas" (como escrever frases individuais).
    - Se um modelo de Inteligência Artificial puder ser treinado para executar uma "sub-tarefa" mais curta, ele poderá realizar a tarefa mais longa simplesmente repetindo a subtarefa mais curta várias vezes - sem nunca precisar ser explicitamente "treinado" para fazer a tarefa mais longa.
    - Por exemplo, um modelo de Inteligência Artificial poderia realizar um milhão de tentativas da seguinte tarefa: "Leia um ensaio parcialmente finalizado e escreva uma boa frase seguinte." Se o modelo aprender a fazer bem essa tarefa, ele poderá escrever um ensaio longo simplesmente repetindo essa tarefa indefinidamente. Ele não precisaria de um processo de treinamento separado, onde realizaria um milhão de "tentativas" da tarefa mais demorada de escrever um ensaio inteiro.
    - Portanto, torna-se crucial que tarefas mais difíceis e importantes (como as listadas acima) possam ser "decompostas" em tarefas curtas/fáceis.

Estimando os custos

O método das Bio-âncoras analisa o custo de treinamento dos modelos de Inteligência Artificial existentes, dependendo do tamanho do modelo e do tipo de tarefa (conforme definido acima). Em seguida, extrapola os valores para estimar quanto custaria treinar um modelo de Inteligência Artificial se:

- Ele fosse 10 vezes maior que um cérebro humano.[101]
- Se as "tentativas" de realizar a tarefa exigissem dias, semanas ou meses de "raciocínio" intensivo.

Atualmente, esse tipo de treinamento custaria cerca de um milhão de trilhões de dólares, o que é muito mais do que a riqueza mundial total. Portanto, não me surpreende que ninguém tenha tentado treinar tal modelo até hoje.

No entanto, o método das Bio-âncoras também prevê as seguintes tendências para o futuro:

- Avanços em hardware e software que tornariam o poder de computação mais barato.
- Uma economia em crescimento e um papel crescente da Inteligência Artificial na economia, podem aumentar o orçamento dos laboratórios de Inteligência Artificial para US$ 1 trilhão ou mais destinados ao treinamento de modelos maiores.

De acordo com essas projeções, em algum momento "o orçamento dos laboratórios de IA" se tornará igual ao valor necessário para se "treinar um modelo do tamanho de um cérebro humano nas tarefas mais difíceis". O método das Bio-âncoras baseia suas projeções para "quando a Inteligência Artificial transformadora será desenvolvida" no tempo em que esse investimento acontecer.

O método das Bio-âncoras também modela a incerteza em todos os parâmetros acima e considera abordagens alternativas para os parâmetros de "tamanho do modelo e tipo de tarefa"[102] Ao fazer isso, ele estima que a Inteligência Artificial transformadora será desenvolvida até 2030, 2035, etc.

Agressivo ou conservador?

O método das Bio-âncoras envolve uma série de simplificações que podem torná-lo muito agressivo (esperar que a Inteligência Artificial transformadora chegue mais cedo do que é realista) ou muito conservador (esperar que chegue mais tarde do que é realista).

O argumento que mais ouço de que ele é **"muito agressivo"** é neste sentido: "Não há razão para pensar que uma Inteligência Artificial baseada em métodos modernos pode aprender tudo o que um ser humano faz, usando treinamento de tentativa e erro — não importa se o modelo é grande ou o quanto ele foi treinado. Os cérebros humanos podem raciocinar de maneiras exclusivas, inigualáveis e incomparáveis a qualquer IA, a menos que criemos abordagens fundamentalmente novas para a IA." Esse tipo de argumento costuma ser acompanhado por falas de que os sistemas de Inteligência Artificial não "entendem verdadeiramente" sobre o que estão raciocinando e/ou que eles estão apenas imitando o raciocínio humano por meio do reconhecimento de padrões.

Acho que isso pode acabar sendo correto, mas eu não apostaria nisso.

Uma discussão completa do porquê isso está fora do escopo desta postagem, mas em resumo:

- Não estou convencido de que haja uma distinção profunda ou estável entre "reconhecimento de padrões" e "compreensão verdadeira" (**este artigo da *Slate Star Codex*** defende esse argumento). A "compreensão verdadeira" pode simplesmente ser a representação de como seria um reconhecimento de padrões excelentes. Parte do meu raciocínio aqui se baseia numa intuição de que mesmo quando as pessoas (incluindo eu) parecem "entender" algo superficialmente, seu raciocínio frequentemente (eu diria até normalmente) falha ao considerar um contexto desconhecido. Em outras palavras, acredito que o que consideramos "compreensão verdadeira" é mais um ideal do que uma realidade.
- Sinto-me desapontado com o histórico daqueles que fizeram esse tipo de argumento —não acredito que eles tenham conseguido identificar o que é o "raciocínio verdadeiro", de modo que pudessem fazer previsões robustas sobre quais tarefas seriam difíceis

para sistemas de Inteligência Artificial realizarem. (Por exemplo, veja **esta discussão da crítica mais recente de Gary Marcus ao GPT3**).

- Pode ser verdade que "Algumas descobertas/avanços fundamentais são necessários". Mas para que o Bio-âncoras seja considerado excessivamente agressivo, não basta que *alguns* avanços sejam necessários; os avanços necessários devem ser *mais do que os cientistas de Inteligência Artificial conseguirão conquistar nas próximas décadas*, período no qual o método das Bio-âncoras prevê o desenvolvimento da Inteligência Artificial transformadora. Parece difícil acreditar que as coisas acontecerão dessa maneira - especialmente porque:

- Mesmo avanços moderados em sistemas de IA, teriam o potencial de atrair mais talento e financiamento para a área (como já está acontecendo[103]).

Se o dinheiro, o talento e o poder de processamento forem abundantes e o progresso em direção ao PASTA for impedido, principalmente por alguma fraqueza específica decorrente de como os sistemas de Inteligência Artificial são projetados e treinados, uma tentativa sustentada dos pesquisadores de corrigir essa fraqueza pode funcionar. Quando nos referimos a cronologias de várias décadas, isso pode ser tempo suficiente para os pesquisadores descobrirem o que está faltando nas técnicas atuais.

De forma mais ampla, o Bio-âncoras pode ser considerado agressivo demais devido à sua suposição de que "o poder computacional é o gargalo":

- Ele pressupõe que, se alguém pudesse pagar por todo o poder computacional necessário para fazer o "treinamento" de força bruta para as tarefas principais (por exemplo, automatizar o trabalho científico), a Inteligência Artificial transformadora (provavelmente) seria desenvolvida em seguida.

- Treinar um modelo de Inteligência Artificial não requer apenas comprar mais poder computacional. Requer contratar pesquisadores, realização de experimentos e, talvez o mais importante, encontrar uma maneira de configurar o processo de "tentativa e erro" para a Inteligência Artificial poder realizar inúmeras tentativas da tarefa principal. Isso pode acabar sendo difícil demais de ser feito.

Por outro lado, existem várias razões para considerar o método das Bio-âncoras **conservador demais** (subestimando a probabilidade da Inteligência Artificial transformadora ser desenvolvida em breve).

- Talvez com engenhosidade suficiente, alguém crie uma Inteligência Artificial transformadora "programando-a" para realizar tarefas principais, em vez de ter que "treiná-la" (veja acima para a distinção). Isso exigiria muito menos poder computacional e, portanto, seria muito menos dispendioso. Ou então, seria possível usar uma combinação de "programação" e "treinamento" para aumentar a eficiência sugerida pelo método, sem que fosse necessário realizar tudo via "programação".

- Poderíamos desenvolver abordagens muito superiores à Inteligência Artificial que seriam "treinadas" com muito mais eficiência. Uma possibilidade aqui é o meta-aprendizado: treinar efetivamente um sistema de Inteligência Artificial na "tarefa" de treinar a si mesmo.

- Ou talvez, com o tempo, a Inteligência Artificial torne-se uma parte crescente da economia, resultando na proliferação de diferentes sistemas de IA personalizados e nos quais se investiu para eles realizarem diferentes tarefas reais. Quanto mais isso aconte-

cer, mais oportunidades existirão para que a engenhosidade individual e a sorte resultem em mais inovações e sistemas de Inteligência Artificial mais capazes em contextos econômicos específicos.

- Talvez em algum momento seja possível integrar muitos sistemas com habilidades diferentes para enfrentar uma tarefa particularmente difícil como "automatizar a ciência", sem a necessidade de uma "rodada de treinamentos" astronomicamente cara.
- Ou talvez a Inteligência Artificial que fica aquém do PASTA ainda seja útil o suficiente para gerar muito dinheiro e/ou ajudar os pesquisadores a baixar os custos da computação e aumentar a sua eficiência. Isso, por sua vez, levaria a modelos de Inteligência Artificial ainda maiores que aumentariam ainda mais a disponibilidade de dinheiro e a eficiência da computação. Isso tornaria uma rodada de treinamento no nível PASTA mais econômica, antes do que o método das âncoras biológicas prevê.
- Além disso, alguns revisores técnicos do método acham que seu tratamento de tipo de tarefa é conservador demais. Eles acreditam que as tarefas mais importantes (e talvez todas as tarefas) nas quais a Inteligência Artificial precisa ser treinada estarão no extremo "mais fácil/mais barato" do espectro, em comparação com o que o Bio-âncoras presume. (Veja a seção acima para o que significa quando uma tarefa é "mais fácil/mais barata" ou "mais difícil/mais cara"). Para um argumento relacionado, veja ***Fun with +12 OOMs of Compute*** **(Diversão com mais de 12 ordens de magnitude de Computação)**, que argumenta intuitivamente que o método das Bio-âncoras está imaginando uma quantidade verdadeiramente massiva de poder computacional necessário para criar o PASTA, e que menos do que o método imagina poderia ser o suficiente.

Não acho que seja óbvio se, no geral, o método das Bio-âncoras é muito agressivo (quando presume que a Inteligência Artificial transformadora ocorrerá mais cedo do que é realista) ou muito conservador (presumindo que ela será desenvolvida mais tarde). O próprio relatório afirma que é provável que ele esteja sendo muito agressivo ao prever o desenvolvimento da Inteligência Artificial transformadora nos próximos anos e, muito conservador estimando um prazo maior que 50 anos, e provavelmente suas estimativas sejam mais úteis se estiverem entre esses dois períodos.[104]

Intelectualmente, parece-me que é mais provável que o relatório seja conservador. Acho suas respostas aos argumentos acima, de que o método é "muito agressivo", bastante convincentes, e acredito que os argumentos de que ele seja "muito conservador" têm mais chances de estarem corretos. Particularmente, acredito difícil descartar a possibilidade de que a engenhosidade leve à Inteligência Artificial transformadora de uma maneira muito mais eficiente do que o método de "força bruta" contemplado aqui. E acredito que o tratamento do "tipo de tarefa" está definitivamente errando numa direção mais conservadora.

No entanto, também tenho uma preferência intuitiva (que está relacionada às análises de "ônus da prova" dadas anteriormente) para errar para o lado mais conservador ao fazer estimativas como essa. No geral, meus melhores palpites sobre cronologias da Inteligência Artificial transformadora são semelhantes às do método das Bio-âncoras.

Conclusões sobre o método das Bio-âncoras

O método estima uma probabilidade **>10% do desenvolvimento da Inteligência Artificial transformadora até 2036, aproximadamente 50% até 2055 e aproximadamente 80% até**

2100.

Também vale a pena ressaltar o que o relatório diz sobre os sistemas de Inteligência Artificial atuais. Ele estima que:

- Os maiores modelos de Inteligência Artificial atuais, como, por exemplo, o GPT-4o, são um **pouco menores que os cérebros de camundongos e estão começando a ficar nos padrões de tamanho (se crescessem de 100 a 1000 vezes) dos cérebros humanos.** Portanto, em breve chegaremos perto de desenvolver sistemas de Inteligência Artificial que serão treinados para fazer qualquer coisa que os humanos consigam fazer com cerca de 1 segundo de pensamento. Consistente com isso, parece-me que estamos apenas começando a chegar ao ponto no qual os modelos de linguagem *soam* como humanos falando espontaneamente.[105] Na verdade, um "humano que não pensa mais do que 1 segundo antes de cada palavra" parece próximo do que o GPT-3 consegue fazer atualmente, embora ele seja muito menor do que um cérebro humano.

- Só muito recentemente os modelos de Inteligência Artificial ficaram desse tamanho. Um modelo de Inteligência Artificial "grande" antes de 2020 teria um tamanho mais próximo do tamanho do cérebro de uma abelha. Portanto, mesmo a respeito de modelos recentes, deveríamos estar nos perguntando se os sistemas de Inteligência Artificial parecem ser "tão inteligentes quanto os insetos". Aqui está **uma tentativa de comparar as habilidades de uma abelha com as da Inteligência Artificial** (de Guille Costa, estagiário na Open Philanthropy), que conclui que algumas das habilidades mais impressionantes das abelhas, como o autor argumenta, parecem ser executáveis por sistemas de IA.[106]

Incluo estas observações porque:

- A análise do método das Bio-âncoras parece totalmente consistente com o que estamos observando nos sistemas de Inteligência Artificial atuais (e das últimas duas décadas), ao mesmo tempo, em que sugere que provavelmente desenvolveremos mais habilidades transformadoras nas próximas décadas.

- Acho particularmente digno de nota que estamos chegando perto do momento em que um modelo de Inteligência Artificial será "tão grande quanto um cérebro humano" (segundo o método de estimativa da **Computação cerebral** /método das Bio-âncoras). Pode acontecer que tal modelo de Inteligência Artificial seja capaz de "aprender" muito sobre o mundo e produzir muito valor econômico, mesmo que ainda não consiga fazer as coisas mais difíceis que os humanos conseguem fazer. E isso, por sua vez, poderia impulsionar investimentos vertiginosos em Inteligência Artificial (tanto em dinheiro, quanto em talento), levando a muito mais inovação e avanços. Esta é uma razão simples para acreditar que o desenvolvimento da Inteligência Artificial transformadora até 2036 é plausível.

Finalmente, ressalto que o método das Bio-âncoras inclui uma análise de "evolução" entre as diferentes abordagens que considera. Esta análise levanta a hipótese de que, para produzir a Inteligência Artificial transformadora, seria necessário fazer tantos cálculos quanto todos os animais da história juntos já fizeram, a fim de recriar o progresso realizado pela seleção natural.

Considero a análise de "evolução" *muito* conservadora, porque o aprendizado de máquina consegue progredir muito mais rapidamente do que o tipo de tentativa e erro associado à seleção natural. Mesmo que alguém acredite em algo como "Os cérebros humanos raciocinam

de maneiras únicas, inigualáveis e incomparáveis com uma Inteligência Artificial moderna", parece que tudo o que é exclusivo aos cérebros humanos poderia ser redescoberto se alguém conseguisse, essencialmente, reexecutar toda a história da seleção natural. E mesmo essa análise muito conservadora estima uma chance de aproximadamente 50% da Inteligência Artificial transformadora ser desenvolvida até 2100.

Prós e contras do método das âncoras biológicas para prever cronologias da Inteligência Artificial transformadora

Contras. Começarei com o que vejo como a maior desvantagem do método: esta é uma estrutura de previsão muito complexa, que depende crucialmente de várias estimativas e suposições extremamente incertas, particularmente:

- Se é razoável acreditar que um sistema de Inteligência Artificial pode aprender as principais tarefas listadas acima (as necessárias para o PASTA) com treinamento de tentativa e erro suficiente.
- Como comparar o tamanho dos modelos de Inteligência Artificial com o tamanho dos cérebros de animais/humanos.
- Como caracterizar o "tipo de tarefa", estimando o quanto ela é "difícil" e cara de se "tentar" realizar ou "assisti-la" sendo realizada uma vez.
- Como usar o tamanho do modelo e o tipo de tarefa para estimar quanto custaria treinar um modelo de Inteligência Artificial para realizar as tarefas consideradas principais.
- Como estimar avanços futuros em hardware e software que poderiam baratear os custos com poder computacional.
- Como estimar aumentos futuros nos orçamentos de laboratórios de Inteligência Artificial destinados ao treinamento de modelos.

Esse tipo de complexidade e incerteza significa (na minha opinião), que não devemos considerar essas previsões como sendo confiáveis, especialmente hoje, quando o método ainda é relativamente novo. Se chegássemos ao ponto em que tanto escrutínio e esforço tivessem sido dedicados à previsão de Inteligência Artificial quanto é dedicado à previsão do clima, tudo isso poderia ser uma questão bem diferente.

Prós. Dito isso, o método das âncoras biológicas é, essencialmente, o único que conheço que fornece estimativas de cronologias de Inteligência Artificial transformadora a partir de **fatos objetivos** (quando possível) **e suposições explícitas** (em outros lugares).[107] Ele não depende de nenhum conceito vago e intuitivo como o conceito de "quão rapidamente os sistemas de Inteligência Artificial estão ficando mais surpreendentes (discutido anteriormente). Cada suposição e estimativa do método podem ser explicadas, discutidas e - ao longo do tempo - testadas.

Mesmo em seu atual estágio inicial de desenvolvimento, considero isso uma propriedade valiosa do método das âncoras biológicas. Isso significa que, esse método poderia fornecer estimativas de cronologias que não são simplesmente reformulações de intuições sobre se parece que o desenvolvimento da Inteligência Artificial transformadora está próximo de acontecer.[108]

Também acho encorajador que, mesmo com todas as especulações, as "previsões" verificáveis que o método realiza atualmente parecem razoáveis (consulte a seção anterior). **O método propõe uma maneira de pensar sobre como pode ser simultaneamente verdadeiro que (a) os sistemas de Inteligência Artificial de uma década atrás não pareciam muito surpreendentes; (b) os sistemas de Inteligência Artificial atuais conseguem fazer muitas**

coisas surpreendentes, mas parecem ainda muito aquém do que os humanos conseguem fazer; (c) o desenvolvimento da Inteligência Artificial transformadora poderá, facilmente, acontecer nas próximas décadas — ou mesmo nos próximos 15 anos.

Além disso, acredito que vale a pena destacar **alguns argumentos de alto nível** apresentados pelo método das Bio-âncoras que **não dependem de tantas estimativas e suposições:**

- Na próxima década, provavelmente desenvolveremos - pela primeira vez - modelos de Inteligência Artificial com "tamanho" comparável ao do cérebro humano.
- Se os modelos de Inteligência Artificial continuarem a se tornar maiores e mais eficientes na proporção que o método das Bio-âncoras estima, provavelmente será **possível atingir alguns marcos bastante extremos, ainda neste século, com um custo acessível - o "ponto alto" daquilo que o método das Bio-âncoras acredita ser necessário.** Estes pontos são difíceis de resumir aqui, mas consulte os métodos de "rede neural de longo prazo" e "âncora de evolução" no relatório.
- Uma maneira de pensar sobre isso é a de que no próximo século passaremos provavelmente de uma "computação insuficiente para executar um modelo com o tamanho do cérebro humano" para um "poder computacional extremamente abundante, até maior do que algumas estimativas conservadoras acreditam que seria necessário." O poder computacional não é o único fator de influência no progresso da IA, mas na medida em que outros fatores (algoritmos, processos de treinamento) se tornam os novos gargalos, provavelmente haverá incentivos poderosos (e várias décadas) para conseguirmos resolvê-los.

Uma vantagem final do método das Bio-âncoras é que podemos continuar a observar o progresso da Inteligência Artificial ao longo do tempo e comparar o que vemos com o apresentado no relatório. Podemos observar, por exemplo:

- Se existem tarefas que simplesmente não podem ser aprendidas, mesmo com muita tentativa e erro — ou se algumas tarefas exigem quantidades de treinamento muito diferentes do que o relatório prevê.
- Como as habilidades dos modelos de Inteligência Artificial se comparam às dos animais que atualmente usamos como referência de "tamanhos semelhantes". Se os modelos de Inteligência Artificial parecem mais capazes do que esses animais, podemos estar superestimando o tamanho que um modelo precisaria para, por exemplo, automatizar a ciência. Se eles parecerem menos capazes, podemos estar subestimando isso.

- Como o hardware e o software estão progredindo e se os modelos de Inteligência Artificial estão ficando maiores na proporção que o relatório projeta atualmente.

O próximo artigo resumirá todas as diferentes análises feitas até o momento sobre as cronologias da Inteligência Artificial transformadora. Ele então discutirá uma ressalva que permanece: que não há um consenso robusto de especialistas sobre esse tópico.

Notas

[96]Claro, a resposta poderia ser "Daqui a um zilhão de anos" ou "Nunca."

[97]Para fins de transparência, observe que esta é uma análise feita pela ***Open Philanthropy***, e sou co-CEO da Open Philanthropy.

[98]Geralmente, considero (assim como o método das Bio-âncoras), o número de sinapses mais importante do que o de neurônios, por razões que não vou abordar aqui.

[99]Wikipedia: "A versão completa do GPT-3 tem capacidade para 175 bilhões de parâmetros de aprendizado de máquina… Antes do lançamento do GPT-3, o maior modelo de linguagem era o Turing NLG da Microsoft, lançado em fevereiro de 2020, com capacidade para 17 bilhões de parâmetros." A Wikipedia não afirma isso, mas não acredito que existam modelos de Inteligência Artificial conhecidos publicamente maiores do que esses modelos de linguagem (com exceção dos modelos "mistura de especialistas" que acredito que devemos desconsiderar para esses propósitos, por motivos que não vou abordar aqui) A Wikipedia estima cerca de 1 trilhão de sinapses para o cérebro de um rato-doméstico; O método das âncoras biológicas para comparações cerebrais (baseada em Computação cerebral) essencialmente iguala sinapses a parâmetros.

[100]O método das âncoras biológicas estima cerca de 100 trilhões de parâmetros para o cérebro humano, com base no fato de que ele possui cerca de 100 trilhões de sinapses.

[101]O tamanho de 10 vezes maior foi escolhido para deixar algum espaço para a ideia de que "cérebros digitais" podem ser menos eficientes que os cérebros humanos. Veja esta seção do relatório.

[102]Por exemplo, uma abordagem levanta a hipótese de que o treinamento poderia ser barateado pelo meta- aprendizado, discutido acima; outra abordagem levanta a hipótese de que, para produzir Inteligência Artificial transformadora, seria necessário fazer tantos cálculos quanto todos os animais da história combinados, a fim de recriar o progresso feito pela seleção natural.)

[103]Veja os gráficos das primeiras seções do 2021 AI Index Report, por exemplo.

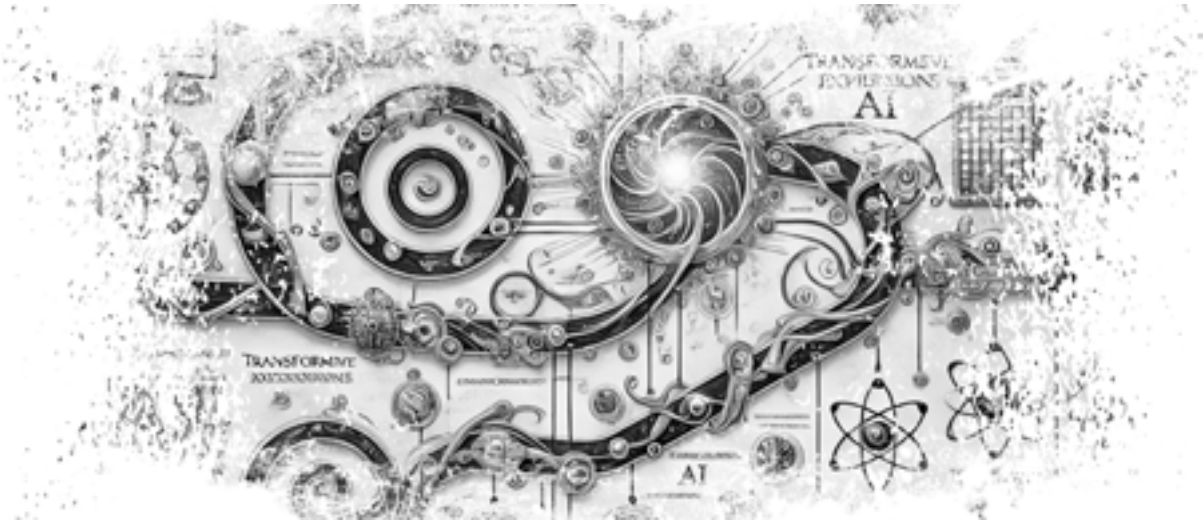
[104]Veja esta seção.

[105]Para uma coleção de links para demos do GPT-3, veja **esta postagem**.

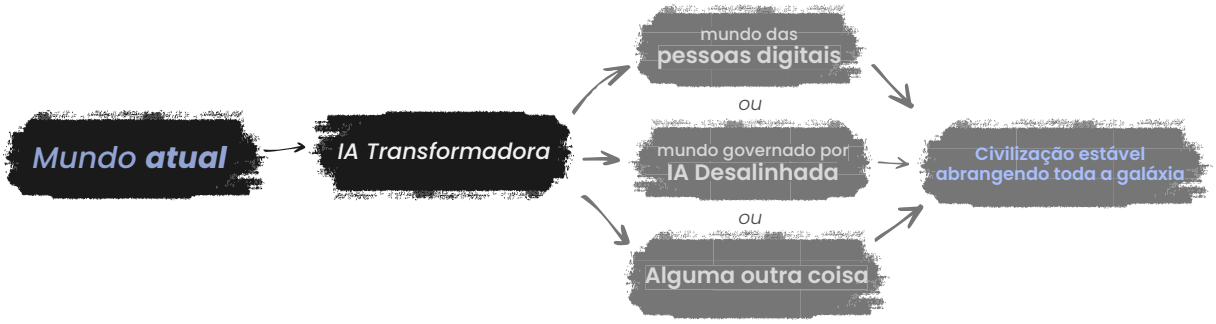
[106]Ele estima que os sistemas de Inteligência Artificial parecem usar cerca de 1000 vezes menos poder computacional, o que corresponderia ao argumento acima em termos de sugerir que os sistemas de Inteligência Artificial podem ser mais eficientes do que os cérebros de animais/humanos e que as estimativas do método das Bio-âncoras podem ser muito conservadoras. Porém, ele não aborda que as abelhas executam um conjunto mais diversificado de tarefas do que os sistemas de Inteligência Artificial aos quais estão sendo comparadas.

[107]Além do método de "priori semi-informativas" discutido anteriormente.

[108]Claro, isso não quer dizer que as estimativas sejam *completamente independentes* das intuições —é provável que as intuições influenciem nossas escolhas de estimativas para muitos dos números que são difíceis de estimar. Mas a capacidade de analisar e debater cada estimativa separadamente é útil aqui.



Cronologia da IA: onde os argumentos e os especialistas se posicionam



Este artigo começa com um resumo de quando devemos esperar que a IA Transformadora seja desenvolvida, com base nas várias perspectivas abordadas anteriormente na série. Acho que este artigo é útil, mesmo que você já tenha lido todos os artigos anteriores, mas se quiser ignorá-lo, clique aqui.

Abordo a questão: "Por que não há consenso robusto de especialistas sobre este tópico e o que isto significa para nós?"

Estimo que há **mais de 10% de probabilidade da Inteligência Artificial transformadora ser desenvolvida dentro de 15 anos (até 2036); aproximadamente 50% disso acontecer dentro de 40 anos (até 2060); e aproximadamente 2/3 de que ela seja desenvolvida ainda neste século (até 2100).**

(Por "IA Transformadora", quero dizer "IA poderosa o suficiente para nos levar a um futuro novo e qualitativamente diferente". Argumentei que a Inteligência Artificial avançada poderia ser o suficiente para tornar este o século mais importante.)

Essa conclusão é baseada em relatórios técnicos que abordam a previsão de Inteligência Artificial de diferentes perspectivas - muitos desses relatórios foram produzidos pela Open Philanthropy nos últimos anos, para desenvolver um panorama completo da previsão da Inteligência Artificial transformadora para fomentar ações de longo prazo.

Aqui está um **resumo de uma tabela** das diferentes perspectivas de previsão da Inteligência Artificial transformadora discutidas anteriormente, com links para discussões mais detalhadas em postagens anteriores, bem como para relatórios técnicos subjacentes:

Perspectiva de previsão	Principais artigos aprofundados (títulos abreviados)	Minhas conclusões
Estimativas de probabilidade para Inteligência Artificial transformadora		
Pesquisa com os especialistas.[1] O que os pesquisadores de Inteligência Artificial esperam?	Evidências de especialistas em IA[2]	Pesquisas de especialistas implicam[109] uma probabilidade de desenvolvimento da Inteligência Artificial de aproximadamente 20% até 2036; aproximadamente 50% até 2060; aproximadamente 70% até 2100. Perguntas reformuladas com palavras ligeiramente diferentes das originais (feitas para um subconjunto menor de entrevistados) resultaram em respostas com estimativas muito mais posteriores.
Método das âncoras biológicas.[3] Com base nos padrões usuais de quanto custa o "treinamento de IA", quanto custaria treinar um modelo de Inteligência Artificial tão grande quanto um cérebro humano para executar as tarefas mais difíceis que os humanos conseguem fazer? E quando isso será barato o suficiente para que alguém faça esse treinamento?"	Bio-âncoras,[4] baseada em Computação cerebral[5]	>10% de probabilidade até 2036; Aproximadamente 50% de probabilidade até 2055; aproximadamente 80% de probabilidade até 2100.
Perspectivas sobre o ônus da prova[6]		
É improvável que qualquer um dos séculos seja "o mais importante"[7] de todos.	*Hinge; Response to Hinge*[8] *(Virada da história: Resposta à hipótese da "virada" da história)*	Temos muitas razões para pensar que este século é "especial" antes de olharmos para os detalhes da IA. Muitos foram abordados em artigos anteriores; outro é coberto na próxima linha.
O que você preveria sobre as cronologias de Inteligência Artificial transformadora, com base apenas em informações básicas sobre (a) há quantos anos as pessoas tentam desenvolver a Inteligência Artificial transformadora; (b) quanto eles "investiram" nela (em termos do número de pesquisadores de Inteligência Artificial e da quantidade de computação usada por eles); (c) se já conseguiram desenvolvê-la (até agora, não)?	*Semi-informative Priors*[10]	Estimativas centrais: 8% até 2036; 13% até 2060; 20% até 2100.[110] Na minha opinião, este relatório destaca que a história da Inteligência Artificial é curta, o investimento em Inteligência Artificial está aumentando rapidamente, portanto, não devemos nos surpreender se a Inteligência Artificial transformadora for desenvolvida em breve.
Com base na análise de modelos econômicos e da história econômica, qual é a probabilidade de "crescimento explosivo" - definido como > 30% de crescimento anual na economia mundial - até 2100?[11]	Crescimento explosivo, Trajetória humana[12]	**Trajetória humana**[13] projeta o passado para frente, implicando um crescimento explosivo até 2043-2065. **Crescimento explosivo**[14] conclui: "Acho que considerações econômicas não fornecem uma boa razão para descartar a possibilidade de desenvolvimento da IATF neste século. Na verdade, existe uma perspectiva econômica plausível a partir da qual se espera que sistemas de Inteligência Artificial suficientemente avançados causem um crescimento explosivo".
"Como as pessoas previram a IA... no passado? Devemos ajustar nossas próprias percepções hoje para corrigir os padrões que podemos observar em previsões anteriores? Nos confrontamos com a ideia de que as opiniões sobre a Inteligência Artificial foramexageradas no passado e que, portanto, devemos presumir que as projeções atuais são excessivamente otimistas".[15]	Previsões de Inteligência Artificial anteriores[16]	O auge do hype em torno da Inteligência Artificial parece ter ocorrido entre 1956-1973. Ainda assim, o hype implícito em algumas das previsões de Inteligência Artificial mais conhecidas desse período é comumente exagerado.

Tendo considerado o que foi dito acima, espero que alguns leitores ainda sintam uma sensação de desconforto. Mesmo que achem que meus argumentos fazem sentido, podem estar se perguntando: se isso é verdade, por que não é mais amplamente discutido e aceito? **Qual é o estado da opinião de especialistas?**

Meu resumo do estado da opinião de especialistas atualmente é:

- Minhas afirmações não contradizem nenhum consenso específico de especialistas. (Na verdade, as probabilidades que dei não estão muito longe do que os pesquisadores de Inteligência Artificial parecem prever, como mostrado na primeira linha.) Mas há alguns sinais de que eles não estão pensando muito sobre o assunto.

- Os relatórios técnicos da Open Philanthropy nos quais confiei tiveram uma revisão significativa feita por especialistas externos. Pesquisadores da área de Aprendizado de Máquina analisaram a ***Bio Anchors* (Bio-âncoras)**; neurocientistas analisaram a ***Brain Computation* (Computação cerebral)**; economistas analisaram a ***Explosive Growth* (Crescimento explosivo)**; acadêmicos com foco em tópicos relevantes sobre incerteza e/ou probabilidade analisaram a ***Semi-informative Priors* (Priori semi-informativa)**. (Algumas dessas análises tinham pontos significativos de desacordo, mas nenhum desses pontos parecia ser sobre casos em que os relatórios contradiziam um claro consenso de especialistas ou a literatura da área.)

- Mas também não há um consenso ativo e robusto de especialistas corroborando afirmações como "Há pelo menos 10% de probabilidade da Inteligência Artificial transformadora ser desenvolvida até 2036" ou "Há uma boa chance de estarmos no século mais importante para a humanidade", da mesma maneira que existem argumentos que embasam a necessidade de agir contra as alterações climáticas, por exemplo.

Em última análise, minhas afirmações são sobre **tópicos que simplesmente não têm uma "área do conhecimento" com especialistas dedicados a estudá-los. Isso, por si só, é um fato assustador** e algo que espero que mude eventualmente.

Mas, enquanto isso, devemos estar dispostos a agir segundo a hipótese do "século mais importante"?

Abaixo, comentarei:

- Como seria um "campo de estudos de previsão da IA".

- Uma "visão cética" que diz que as discussões atuais sobre esses tópicos são muito pequenas, homogêneas e insulares (o que eu concordo) — e que, portanto, não devemos agir sobre a hipótese do "século mais importante" até que haja um campo de estudos maduro e robusto (o que eu não concordo).

- Porque acredito que devemos levar a hipótese a sério enquanto isso, até e a menos que tal campo de estudos se desenvolva:

- Não temos tempo para aguardar por um consenso robusto de especialistas.

- Se houver boas refutações por aí — ou futuros especialistas em potencial que possam desenvolver tais refutações — ainda não os encontramos. Quanto mais a hipótese for levada a sério, maior a probabilidade de tais refutações aparecerem. (Também conhecida como a teoria da **Lei de Cunningham**: "a melhor maneira de obter uma resposta certa é postar uma resposta errada".)

- Acho que insistir consistentemente na criação de um consenso robusto de especialistas é um padrão de raciocínio perigoso. Na minha opinião, não é um problema correr o risco de autoilusão e isolamento, em troca de se fazer a coisa certa quando isso é o mais importante.

Que tipo de especialização é a especialização em previsão de IA?

As questões analisadas nos relatórios técnicos listados acima incluem:

- As habilidades da Inteligência Artificial estão ficando cada dia mais surpreendentes com o tempo? (IA, história da IA)
- Como podemos comparar modelos de Inteligência Artificial com cérebros de animais/humanos? (IA, neurociência)
- Como podemos comparar as habilidades da Inteligência Artificial com as habilidades dos animais? (IA, etologia)
- Como podemos estimar o custo de treinar um sistema de Inteligência Artificial grande o suficiente para realizar uma tarefa difícil, com base nas informações que temos sobre treinamentos anteriores de sistemas de IA? (IA, ajuste de curva).
- Como podemos fazer uma estimativa com o mínimo de informações sobre Inteligência Artificial transformadora, com base apenas em quantos anos/pesquisadores/dólares foram investidos no campo de estudos até agora? (Filosofia, probabilidade).
- Qual é a probabilidade de um crescimento econômico explosivo neste século, com base na teoria e nas tendências históricas? (Economia do crescimento, história econômica)
- Como foi o "hype em torno da IA" no passado? (História)

Ao falar sobre implicações mais amplas da Inteligência Artificial transformadora para o "século mais importante", também discuti temas como "Qual a viabilidade do desenvolvimento de pessoas digitais e o estabelecimento de assentamentos espaciais em toda a galáxia?" Esses tópicos abordam física, neurociência, engenharia, filosofia da mente e muito mais.

Não há trabalho ou credencial óbvia que torne alguém um especialista na questão de quando podemos esperar uma Inteligência Artificial transformadora, ou na questão de saber se estamos no século mais importante.

(Eu discordaria particularmente de qualquer afirmação de que deveríamos confiar exclusivamente em pesquisadores de Inteligência Artificial para essas previsões. Além da aparente falta de reflexão profunda sobre o assunto, acredito que confiar em especialistas em IA para prever o desenvolvimento da IA transformadora é semelhante a confiar em empresas de energia solar ou de petróleo para prever as emissões de carbono e as mudanças climáticas. Embora compreendam parte do cenário, a previsão é uma habilidade distinta da inovação ou da construção de sistemas de ponta. Eles certamente entendem parte do panorama. Mas a previsão é uma atividade distinta da inovação ou construção de sistemas de ponta.)

E nem tenho certeza se essas perguntas têm a forma certa para um campo acadêmico. Tentar prever a Inteligência Artificial transformadora ou determinar as chances de estarmos no século mais importante parece:

- Mais semelhante ao **modelo eleitoral da *Five Thirty Eight*** (“Quem vai ganhar a eleição?”) do que à ciência política acadêmica (“Como governos e constituintes interagem?”);
- Mais semelhante aos mercados financeiros (“Este preço vai subir ou descer no futuro?”) do que à economia acadêmica (“Por que existem recessões?”);[111]
- Mais parecido com a pesquisa da ***GiveWell*** (“Qual instituição de caridade ajudará mais as pessoas, por dólar?”) do que a economia do desenvolvimento (“O que causa a pobreza e o que pode reduzi-la?”)[112]

Ou seja, não está claro para mim como seria um “lar institucional” natural para especialização em previsão transformadora de Inteligência Artificial e o “século mais importante”. Mas parece justo dizer não haver instituições grandes e robustas dedicadas a esse tipo de questão hoje.

Como devemos agir na ausência de um consenso sólido de especialistas?

A visão cética

Na falta de um consenso robusto de especialistas, espero que algumas pessoas (na verdade, a maioria) sejam céticas, não importa quais argumentos sejam apresentados.

Aqui está uma versão de uma reação cética muito comum pela qual tenho bastante empatia:

1. *Isto é muito* audacioso.
2. *Você está fazendo uma afirmação exagerada sobre viver no século mais importante. Este* ***padrão corresponde à autoilusão.***
3. *Você argumentou que o ônus da prova não deveria ser tão alto, porque existem muitas razões para acreditar que vivemos em uma época notável e instável. Mas…* não confio em mim mesmo para avaliar essas afirmações, ou suas afirmações sobre IA, ou realmente qualquer coisa sobre esses tópicos audaciosos.
4. *Estou preocupado que tão poucas pessoas parecem estar engajadas nesses argumentos sobre como a discussão parece* ***pequena, homogênea e insular.*** *No geral, isso parece uma história que as pessoas inteligentes estão contando a si mesmas - com muitos gráficos e números para racionalizá-la - sobre seu próprio lugar na história. Não parece “real”.*
5. *Então me ligue de volta quando houver um campo maduro de talvez centenas ou milhares de especialistas, se criticando e avaliando reciprocamente, e quando eles chegarem ao mesmo tipo de consenso que vemos para a mudança climática.*

Entendo por que você se sente assim, e eu mesmo já me senti assim algumas vezes - especialmente nos pontos n.º 1 ao 4. Mas darei três razões pelas quais o ponto n.º 5 não está correto.

Motivo 1: não temos tempo para esperar por um consenso robusto de especialistas

Eu me preocupo de que a chegada da Inteligência Artificial transformadora possa se tornar uma espécie de versão em câmera lenta e de alto risco da pandemia da COVID-19. O argumento para esperar que algo grande aconteça existe, se você estudar as melhores informações e análises disponíveis atualmente. Mas a situação é amplamente desconhecida; não se encaixa nos padrões com os quais nossas instituições lidam regularmente. E cada ano extra de ação é valioso.

Você também pode pensar nisso como uma versão acelerada da dinâmica com a mudança climática. Imagine se as emissões de gases de efeito estufa só tivessem começado a aumentar recentemente[113] (em vez de em meados dos anos 1800), e se não houvesse um campo estabelecido da ciência do clima. Seria uma péssima ideia esperar décadas pelo surgimento de um campo, antes de tentar reduzir as emissões.

Motivo 2: A Lei de Cunningham ("a melhor maneira de obter uma resposta certa é postar uma resposta errada") pode ser nossa melhor esperança para encontrar a falha nesses argumentos

Estou falando sério.

Há vários anos, alguns colegas e eu suspeitávamos que a hipótese do "século mais importante" poderia ser verdadeira. Mas antes de agirmos, queríamos ver se poderíamos encontrar falhas fatais nessa hipótese.

Uma forma de interpretar nossas ações nos últimos anos é interpretá-las como **se estivéssemos fazendo de tudo para comprovar que a hipótese está errada.**

Primeiro, tentamos conversar com as pessoas sobre os argumentos principais —pesquisadores de IA, economistas, etc. Mas:

- Tínhamos ideias vagas sobre os argumentos desta série (principalmente, ou talvez totalmente, oriundos de outras pessoas). Não conseguíamos explicá-los de forma nítida e específica.
- Havia muitos argumentos concretos importantes que achávamos que provavelmente estariam corretos,[114] mas eles não tinham sido definidos ainda e não estavam prontos para serem apresentados.
- No geral, não conseguimos articular um caso concreto o suficiente para dar aos outros uma oportunidade justa de refutá-lo.

Por isso, trabalhamos muito na criação de relatórios técnicos sobre muitos desses principais argumentos. (Eles agora são públicos e foram incluídos na tabela na parte superior deste artigo.) Isso nos deu a oportunidade de publicá-los e potencialmente encontrar contra-argumentos fatais.

Em seguida, contratamos análises de especialistas externos.[115]

Falando apenas por mim, a hipótese do "século mais importante" parece ter sobrevivido a tudo isso. Na verdade, após analisar muitas das perspectivas existentes e me aprofundar nos detalhes, acredito nessa hipótese com mais convicção do que antes.

Mas digamos que seja apenas porque os *verdadeiros* especialistas —pessoas que ainda não encontramos, com contra-argumentos devastadores — acham a coisa toda tão boba que nem se preocupam em discutir sobre isso. Ou, digamos, que existam pessoas por aí atualmente, que possam *um dia* se tornar especialistas nesses tópicos e derrubar esses argumentos. O que poderíamos fazer para isso acontecer?

A melhor resposta que encontrei para essa pergunta foi: "Se essa hipótese se tornasse mais conhecida, mais amplamente aceita e mais influente, ela seria analisada criticamente."

Esta série é uma tentativa de dar um passo nessa direção —angariar uma credibilidade mais ampla para a hipótese do "século mais importante". Isso seria uma coisa boa se a hipótese fos-

se verdadeira; também seria a próxima ação a ser tomada se o meu único objetivo fosse desafiar minhas crenças e descobrir que a hipótese é falsa.

Claro, não estou dizendo para você aceitar ou promover a hipótese do "século mais importante" se ela não parecer correta para você. Mas acredito que se a sua *única* desconfiança é a falta de um consenso robusto, continuar ignorando a situação me parece estranho. Se as pessoas se comportassem dessa maneira, em geral (ignorando qualquer hipótese que não fosse aceita por um consenso robusto de especialistas), não acredito que seria possível que qualquer hipótese — incluindo as verdadeiras — passasse de marginal a aceita.

Motivo 3: um ceticismo tão geral assim parece uma má ideia

Há um tempo, quando eu estava focado na ***GiveWell***, as pessoas ocasionalmente me diziam coisas do tipo: "Sabe, você não pode exigir que todos os argumentos tenham o mesmo nível que as principais instituições de caridade da *GiveWell* — que sejam submetidos a ensaios clínicos randomizados, dados empíricos robustos, etc. Algumas das melhores oportunidades para fazer o bem serão as que são menos óbvias — portanto, este padrão de exigência pode eliminar algumas das suas maiores potenciais oportunidades de causar impacto."

Eu concordo. Acho que é importante verificarmos qual é a abordagem geral de raciocínio e quais são os padrões probatórios e nos perguntar: "Em quais cenários a minha abordagem falha e em quais deles eu preferiria que ela não falhasse?

Na minha opinião, **não há problema em correr o risco de autoilusão e isolamento, em troca de fazer a coisa certa quando isso é o mais importante.**

Acho que a falta de um consenso robusto de especialistas — e preocupações com autoilusão e insularidade — nos dão boas razões para nos *aprofundar* na hipótese do "século mais importante" primeiro, antes de aceitá-la imediatamente. Perguntar-se onde pode haver uma falha oculta, se existe algum viés para inflar nossa própria importância, pesquisar as partes do argumento que parecem as mais questionáveis, etc.

Mas se você já investigou o assunto, o quanto for razoável ou prático para você e não encontrou nenhuma falha além de considerações do tipo "Não há um consenso robusto de especialistas" e "Estou preocupado com a autoilusão e o isolamento" — então acredito que descartar essa hipótese garantirá que você não será uma das primeiras pessoas a perceber e a agir em uma questão tremendamente importante, caso a oportunidade se apresente. Acredito ser muito sacrifício, no que diz respeito a renunciar a oportunidades com o potencial de fazer o bem.

Notas

[109]Tecnicamente, essas probabilidades são para "inteligência de máquina ao nível humano". Em geral, este gráfico simplifica as coisas ao apresentar um conjunto unificado de probabilidades. Em geral, todas essas probabilidades referem-se a algo *pelo menos* tão capaz quanto o PASTA, então elas devem ser subestimadas direcionalmente da probabilidade de desenvolvimento do PASTA (embora eu não ache que isso seja um problema importante).

[110]Análises sobre o método das Bio-âncoras estão **aqui**; as análises sobre Crescimento Explosivo estão aqui; as análises sobre priori semi-informativas estão aqui. Computação cerebral foi analisado em um momento anterior, quando não tínhamos projetado o processo para resultar na publicação de análises, mas as mais de 20 conversas com especialistas que compuseram o relatório estão disponíveis **aqui**. A Trajetória Humana não foi revisada, embora muitas de suas análises e conclusões apareçam em Crescimento Explosivo, que foi analisado.

[111]Os campos acadêmicos são bastante amplos e estou apenas dando exemplos de questões que eles abordam.

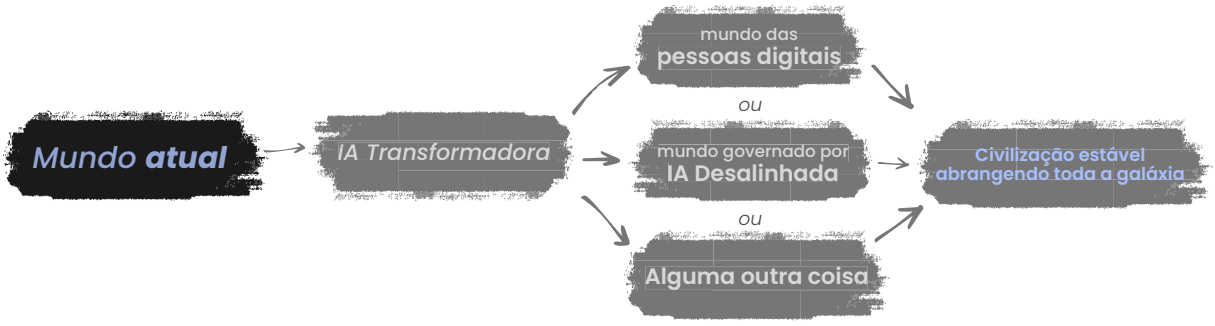
[112]Embora a ciência do clima seja um exemplo de um campo acadêmico que investe muito em prever o futuro.

[113]O campo de estudos da Inteligência Artificial existe desde **1956**, mas foi apenas na última década que os modelos de aprendizado de máquina começaram chegar próximos do tamanho dos cérebros de insetos e ter um bom desempenho em tarefas relativamente difíceis.

[114]Frequentemente, estávamos simplesmente partindo de nossas impressões sobre o que outras pessoas que já pensaram muito sobre o assunto haviam pensado.

[115]As análises sobre o método das Bio-âncoras estão **aqui**; as análises sobre Crescimento explosivo estão aqui; as análises sobre priori semi-informativa estão aqui. O Computação cerebral foi analisado em um momento anterior, quando não havíamos elaborado um processo para resultar na publicação de análises, mas mais de 20 conversas com especialistas que contribuíram com o relatório estão disponíveis **aqui**. O Trajetória Humana não foi revisado, embora muitas de suas análises e conclusões apareçam em Crescimento explosivo, analisado.

Como aproveitar o melhor do século mais importante?



Anteriormente na série "o século mais importante", argumentei haver uma alta probabilidade[116] de que as próximas décadas verão:

- O desenvolvimento de uma tecnologia como o PASTA (processo para automação do avanço científico e tecnológico).
- Uma resultante explosão de produtividade, levando ao desenvolvimento de novas tecnologias transformadoras.
- Os primórdios de uma civilização estável em toda a galáxia, possivelmente povoada por pessoas digitais, ou gerida por uma Inteligência Artificial desalinhada.

Essa é uma visão otimista ou pessimista do mundo? Para mim, são as duas coisas e nenhuma, porque **esse conjunto de eventos pode acabar sendo muito bom ou muito ruim para o mundo, dependendo de como ele se desenrolar.**

Quando falo em estar no "século mais importante", não quero dizer apenas que eventos importantes irão ocorrer. Quero dizer que nós, as pessoas que vivem neste século, temos a oportunidade de causar um enorme impacto no grande número de pessoas que existirão no futuro — se conseguirmos compreender essa situação o suficiente para tomar medidas que sejam úteis.

Mas, também, é importante entender por que esse é um grande "se" - porque o século mais importante apresenta um **cenário estratégico desafiador, de tal forma que muitas coisas que fizermos podem melhorar ou piorar as coisas (e é difícil dizer quais seriam elas).**

Nesta postagem, apresentarei dois cenários contrastantes de como aproveitar melhor o século mais importante:

- O modelo **"Cautela".** Nesse modelo, muitos dos piores resultados decorrem do desenvolvimento de algo como o PASTA de uma maneira muito rápida, apressada ou imprudente. Podemos precisar alcançar uma coordenação (possivelmente global) para mitigar as pressões para competir e tomar os devidos cuidados. (Atenção)

- O modelo **"Competição".** Esse modelo não se concentra em *como e quando o* PASTA será desenvolvido, mas *quem* (quais governos, quais empresas, etc.) será o primeiro a se beneficiar da explosão de produtividade resultante. (Competição)

- As pessoas que seguem o modelo de "cautela" e as pessoas que adotam o modelo de "competição", favorecem geralmente medidas muito diferentes e até contraditórias. Medidas que parecem importantes para as pessoas em um modelo, muitas vezes são ativamente prejudiciais para as pessoas no outro.

O Eu me preocupo com o fato do modelo de "competição" ser superestimado por padrão e discuto o porquê abaixo. (Mais)

- Para entender melhor como avaliar esses modelos e quais ações provavelmente serão úteis, precisamos de mais progresso nas perguntas em aberto sobre o tamanho dos diferentes tipos de riscos da Inteligência Artificial transformadora. (Perguntas abertas)

- Nesse ínterim, existem algumas **ações vigorosamente úteis** que podem provavelmente melhorar as perspectivas da humanidade de alguma maneira. (Ações vigorosamente úteis)

O modelo "cautela".

Defendi a probabilidade de que este século terá uma transição para um mundo onde pessoas digitais ou Inteligência Artificial desalinhada (ou algo muito diferente dos humanos de hoje) serão a principal força nos eventos mundiais.

O modelo de "cautela" enfatiza que alguns tipos de transição podem ser melhores do que outras. Abaixo, elas estão listadas em ordem da pior para a melhor:

A pior: Inteligência Artificial desalinhada

Eu discuti essa possibilidade anteriormente, baseando-me em uma série de outras discussões mais completas.[117] A ideia básica é que os sistemas de Inteligência Artificial podem acabar desenvolvendo objetivos próprios e procurar ganhar espaço cumprindo esses objetivos. Os seres humanos, e/ou todos os valores humanos, podem ser ignorados (ou extintos, se de outra forma os atrapalharmos).

A segunda pior:[118] Maturidade Tecnológica Adversária

Se chegarmos ao ponto em que existem pessoas digitais e/ou IAs (não desalinhadas) que podem se copiar sem limites e se expandir pelo espaço, pode haver uma pressão intensa para os atores agirem — e se multiplicarem (via cópia) — o mais rápido possível, a fim de ganhar mais influência sobre o mundo. Isso pode levar diferentes países/coalizões a furiosamente tentarem ultrapassar uns aos outros e/ou a um conflito militar total, sabendo que muito pode estar em jogo em pouco tempo.

Suponho que esse tipo de dinâmica teria o risco de deixar grande parte da galáxia em mau estado.[119]

Um desses cenários ruins seria estar "permanentemente sob o controle de uma única pessoa (digital) (e/ou suas cópias)". Devido ao potencial das pessoas digitais para criar civilizações

estáveis, um determinado regime totalitário poderia acabar permanentemente entrincheirado em partes substanciais da galáxia.

Pessoas/países/coalizões que *suspeitam uns dos outros* de representar esse tipo de perigo - de potencialmente estabelecer civilizações estáveis sob seu controle - competiriam e/ou atacariam uns aos outros desde o início para evitar que isso acontecesse. Isso poderia acarretar uma guerra com resultados difíceis de prever (devido aos avanços tecnológicos difíceis de prever que o PASTA poderia trazer).

A segunda melhor: negociação e governança

Os países evitariam esse tipo de dinâmica de Maturidade Tecnológica Adversária planejando e negociando entre si. Por exemplo, talvez cada país, ou cada pessoa, tivesse permissão para criar um certo número de pessoas digitais (sujeito a proteções de direitos humanos e outros regulamentos), limitado a uma determinada região do espaço.

Há uma enorme variedade de diferentes potenciais específicos aqui, alguns muito melhores e mais justos do que outros.

A melhor: reflexão

O mundo atingiria um nível de coordenação alto o suficiente para *atrasar* quaisquer etapas irreversíveis do processo de transição (incluindo o início de uma dinâmica de Maturidade Tecnológica Adversária).

Haveria, então, algo como o que Toby Ord (em **O precipício)** chama de "Longa Reflexão":[120] um longo período no qual as pessoas decidiriam coletivamente sobre metas e esperanças para o futuro, que representassem idealmente o compromisso mais justo disponível entre diferentes perspectivas. A tecnologia avançada ajudaria esse processo a ser muito melhor do que seria possível atualmente.[121]

Há perguntas ilimitadas sobre como tal "reflexão" funcionaria e se há realmente alguma esperança de que chegaríamos a um resultado razoavelmente bom e justo. Detalhes como "que tipos de pessoas digitais seriam criadas primeiro" seriam extremamente importantes. Atualmente, há pouca discussão sobre esse tipo de tópico.[122]

Outra

Provavelmente existem muitos tipos possíveis de transições que não mencionei aqui.

O papel da cautela

Se a ordem acima estiver correta, então o futuro da galáxia seria melhor se:

- A Inteligência Artificial desalinhada fosse evitada: sistemas de Inteligência Artificial poderosos agiriam para ajudar os humanos, em vez de buscar seus próprios objetivos.
- A Maturidade Tecnológica Adversária fosse evitada. Isso provavelmente significaria que as pessoas não implantariam sistemas avançados de IA, ou suas tecnologias resultantes, de maneira adversária (a menos que isso fosse necessário para evitar algo pior).
- Uma coordenação suficiente fosse alcançada para os principais atores conseguirem desacelerar, tornando a Reflexão uma possibilidade.

Idealmente, todos com potencial para construir algo semelhante ao PASTA conseguiriam focar sua energia na construção de algo seguro (não desalinhado) e planejariam cuidadosamente

(e negociariam com os outros) como implementá-lo, sem pressa ou competição. Com isso em mente, talvez devêssemos fazer coisas como:

- Trabalhar para melhorar a confiança e a cooperação entre as principais potências mundiais. Talvez, por meio de versões da ***Pugwash*** centradas em Inteligência Artificial (uma conferência internacional visando reduzir o risco de conflito militar), ou recuar de movimentos de relações exteriores mais agressivos.

- Desencorajar governos e investidores a investir dinheiro em pesquisas de IA, encorajar os laboratórios de Inteligência Artificial a considerar minuciosamente as implicações de suas pesquisas antes de publicá-las ou ampliá-las, etc. Desacelerar as coisas dessa maneira seria uma forma de ganhar mais tempo para pesquisar sobre como evitar a Inteligência Artificial desalinhada, mais tempo para construir mecanismos de confiança e cooperação, mais tempo para obter clareza estratégica geral e uma menor probabilidade da dinâmica de Maturidade Tecnológica Adversária.

O modelo "competição"

(Observação: há algum potencial para confusão entre a ideia de "competição" e a ideia de Maturidade Tecnológica Adversária, então tentei empregar termos muito diferentes. Explico esse contraste em uma nota de rodapé.[123])

O modelo de "competição" se concentra **menos em como a transição para um futuro radicalmente diferente aconteceria e mais em quem estaria tomando as decisões principais quando a transição acontecesse.**

- Se algo como o PASTA fosse desenvolvido principalmente (ou primeiro) no país X, então o governo do país X tomaria muitas decisões cruciais sobre se, e como, regular uma potencial explosão de novas tecnologias.

- Além disso, as pessoas e organizações que liderassem o desenvolvimento da Inteligência Artificial e outros avanços tecnológicos nessa época, seriam especialmente influentes em tais decisões.

Isso significa que "quem liderasse o desenvolvimento da Inteligência Artificial transformadora" seria muito importante – o país ou países, pessoas ou organizações.

- Os governos que liderassem a Inteligência Artificial transformadora seriam regimes autoritários?

- Quais governos teriam maior probabilidade de (efetivamente) compreender razoavelmente os riscos e consequências ao tomar decisões importantes?

- Quais governos seriam menos propensos a usar tecnologia avançada para consolidar o poder e o domínio de um grupo? (Infelizmente, não posso dizer que há algum que me agrade aqui.) Quais seriam mais propensos a considerar tomar medidas para "evitar resultados aprisionados/*locked-in*, permitindo que o progresso geral no mundo tivesse tempo de avançar, a fim de aumentar a probabilidade de ter bons resultados para todos".

- Questões semelhantes se aplicam às pessoas e organizações que liderassem o desenvolvimento da Inteligência Artificial transformadora. Quais delas seriam mais prováveis de optar por uma direção positiva?

Algumas pessoas acham que podemos afirmar com confiança atualmente, quais países específicos e/ou quais pessoas e organizações gostaríamos que liderassem o desenvolvimento da Inteligência Artificial transformadora. Essas pessoas defenderiam medidas como:

- Aumentar as chances de que os primeiros sistemas PASTA sejam construídos em países que fossem, por exemplo, menos autoritários, o que significaria, por exemplo, pressionar por mais investimento e atenção ao desenvolvimento da Inteligência Artificial nesses países.

- Apoiar e tentar acelerar os laboratórios de Inteligência Artificial administrados por pessoas que provavelmente tomariam decisões sábias (tais como, envolvimento com governos, quais sistemas de Inteligência Artificial publicar e implantar contra manter em segredo, etc.)

Porque temo que a "competição" seja superestimada em relação à "cautela"

Por padrão, presumo que muitas pessoas gravitem em torno do modelo de "competição" em vez do modelo de "cautela" - por motivos que não considero bons, como:

- Acho que as pessoas ficam naturalmente mais animadas em "ajudar os mocinhos a vencer os bandidos" do que em "ajudar a todos nós a evitar um resultado universalmente ruim, por razões impessoais como 'projetamos sistemas de Inteligência Artificial desleixados' ou 'criamos um sistema dinâmico onde a pressa e a agressividade são recompensadas.'"

- Espero que as pessoas tendam a ter excesso de confiança sobre quais países, organizações ou pessoas consideram os "mocinhos".

- Abraçar o modelo de "competição" tende a apontar para medidas como – trabalhar para acelerar o desenvolvimento de Inteligência Artificial de um determinado país ou organização – que sejam lucrativas, emocionantes e naturalmente fáceis de engajar.

Abraçar o modelo de "cautela" é bem menos assim.

- As maiores preocupações que o modelo de "cautela" enfoca - Inteligência Artificial desalinhada e Maturidade Tecnológica Adversária - são um pouco abstratas e difíceis de entender. De muitas maneiras, elas parecem representar os riscos mais altos, mas é mais fácil sentir um medo visceral de "ficar atrás de países/organizações/pessoas que me assustam" do que sentir um medo visceral de "acabar com um resultado ruim no futuro da galáxia, porque apressamos demais as coisas neste século."

    - Acho que o risco da Inteligência Artificial desalinhada é particularmente difícil de ser levado a sério por muitas pessoas. Parece algo maluco ou uma narrativa de ficção científica; as pessoas que se preocupam com isso tendem a ser interpretadas como se imaginassem algo parecido com "O Exterminador do Futuro", e pode ser difícil entender as suas preocupações mais específicas.

    - Espero publicar mais postagens no futuro que ajudem a dar uma ideia intuitiva de porque acredito que a Inteligência Artificial desalinhada é um risco real.

Portanto, para evitar dúvidas, direi que acredito que o modelo de "cautela" tem muito a seu favor. Em particular, a Inteligência Artificial desalinhada **e a** Maturidade Tecnológica Adversária **parecem *muito* piores do que outros tipos de transição em potencial,** e ambos parecem coisas que têm uma chance real de fazer todo o futuro de nossa espécie (e nossos sucessores) muito piores do que poderiam ser.

Preocupa-me que muito do modelo de "competição" leve-nos a subestimar o risco de desalinhamento e a correr para implantar sistemas inseguros e imprevisíveis, o que teria muitas consequências negativas.

Dito isso, **considero ambos os modelos seriamente importantes.** No geral, continuo com muitas dúvidas de qual modelo seria mais importante e útil (se algum for).

Principais perguntas não respondidas sobre “cautela” vs. “competição”

Geralmente, as pessoas que seguem o modelo de “cautela” e as pessoas que adotam o modelo de “competição” **favorecem ações muito diferentes e até contraditórias.** Ações que parecem importantes para as pessoas em um modelo muitas vezes parecem ativamente prejudiciais para as pessoas no outro.

Por exemplo, as pessoas no modelo de “competição” geralmente preferem avançar o mais rápido possível no desenvolvimento de sistemas de Inteligência Artificial mais poderosos; para as pessoas no modelo de “cautela”, a pressa é uma das principais coisas a evitar. As pessoas no modelo de “competição” geralmente favorecem relações externas adversas, enquanto as pessoas no modelo de “cautela” geralmente desejam que as relações externas sejam mais cooperativas.

(Dito isto, essa dicotomia é uma simplificação. Muitas pessoas, inclusive eu, concordam com os dois modelos. E, qualquer um deles pode implicar em ações normalmente associadas ao outro; por exemplo, você pode ser adepto do modelo “cautela”, mas sente que a pressa é necessária agora a fim de estabelecer um país com uma liderança clara o suficiente em Inteligência Artificial para ter mais tempo, priorize evitar a Inteligência Artificial desalinhada, etc.)

Gostaria de poder dizer com confiança quanta importância colocar em cada modelo e quais ações provavelmente seriam úteis. Mas não posso. Acredito que teríamos mais clareza sobre isso se tivéssemos repostas melhores para algumas perguntas principais que ainda não foram respondidas:

Pergunta não respondida: quão difícil é o problema de alinhamento?

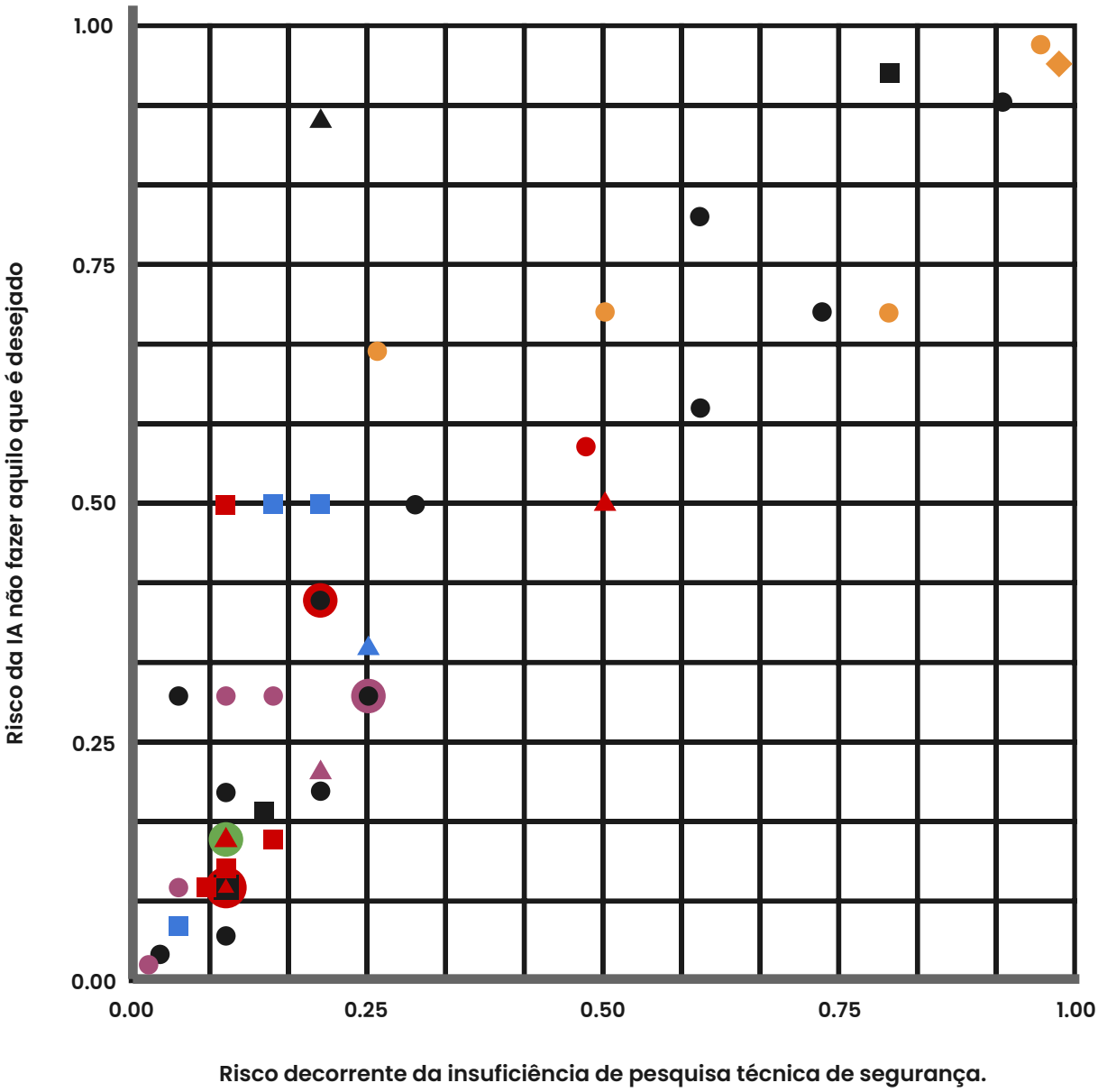
O pior caminho para o futuro é o da Inteligência Artificial desalinhada, no qual os sistemas de Inteligência Artificial acabariam com objetivos próprios não compatíveis com humanos e procurariam preencher a galáxia de acordo com esses objetivos. Quão seriamente devemos assumir esse risco — quão difícil seria evitar esse resultado? **Quão difícil seria resolver o “problema de alinhamento”, significando** ter essencialmente a capacidade técnica de construir sistemas que não farão isso?[124]

- Algumas pessoas acreditam que o problema de alinhamento será formidável; que nossa única esperança de o resolver acontecerá em um mundo onde teremos muito tempo e não estaremos em uma competição para implantar Inteligência Artificial avançada; e que evitar o resultado da “IA desalinhada” deve ser, de longe, a consideração dominante para o século mais importante. Essas pessoas tendem a favorecer fortemente as intervenções de “cautela” descritas acima: elas acreditam que acelerar o desenvolvimento da Inteligência Artificial aumenta nosso risco já substancial de ter o pior resultado possível.

- Algumas pessoas acreditam que será fácil e/ou que toda a ideia de “IA desalinhada” é equivocada, boba ou até mesmo incoerente — planejando um evento futuro excessivamente específico. Essas pessoas geralmente estão mais interessadas nas intervenções de “competição” descritas acima: elas acreditam que a Inteligência Artificial avançada provavelmente será usada eficazmente por qualquer país (ou em alguns casos coalizão, ou empresa menor) que a desenvolva primeiro e, portanto, a questão é quem será o primeiro a desenvolvê-la.

E muitas pessoas estão entre uma opinião e a outra.

A propagação aqui é extrema. Por exemplo, veja **estes resultados** de uma "pesquisa de duas perguntas enviada informalmente para aproximadamente 117 pessoas que trabalham com risco de Inteligência Artificial de longo prazo, perguntando sobre o nível de risco existencial da 'humanidade não fazer suficiente pesquisa técnica de segurança de Inteligência Artificial' e de 'sistemas de Inteligência Artificial que não estão fazendo/otimizando o que as pessoas que os implantaram queriam/pretendiam que eles fizessem.'" (Como mostra o gráfico de dispersão, as pessoas deram respostas semelhantes às duas perguntas.)



Temos entrevistados que acham haver menos de 5% de chance de que os problemas de alinhamento reduzam drasticamente o bem do futuro; respondentes que acham haver mais de 95% de chance; e quase de tudo no meio.[125] Minha impressão é que esta é uma representação justa da situação: mesmo entre as poucas pessoas que passaram mais tempo pensando sobre esses assuntos, praticamente não há consenso ou convergência sobre o quão difícil será o problema do alinhamento.

Espero que, com o tempo, o campo de pesquisas sobre alinhamento de IA[126] cresça e, à medida que as pesquisas sobre Inteligência Artificial e sobre o alinhamento avancem, entenderemos melhor qual é a dificuldade do problema de alinhamento de IA. Isso, por sua vez, esclareceria mais o tema de como priorizar "cautela" versus "competição".

Outras perguntas não respondidas

Mesmo que tivéssemos clareza sobre a dificuldade do problema do alinhamento, muitas perguntas espinhosas permaneceriam.

Devemos esperar o desenvolvimento de uma Inteligência Artificial transformadora nos próximos 10 a 20 anos ou muito mais tarde? Os principais sistemas de Inteligência Artificial passarão de muito limitados a muito capazes rapidamente (*hard takeoff*) ou gradualmente (*slow takeoff*)?[127] Devemos presumir que os projetos do governo desempenharão um papel importante no desenvolvimento da Inteligência Artificial ou que a Inteligência Artificial transformadora surgirá principalmente do setor privado? Alguns governos são mais propensos do que outros a trabalhar para que a Inteligência Artificial transformadora seja usada com cuidado, inclusão e humanidade? O que devemos esperar que um governo (ou empresa) *faça* literalmente se ganhar a capacidade de acelerar drasticamente o avanço científico e tecnológico por meio da IA?

Com essas e outras questões em mente, muitas vezes é difícil olhar para alguma ação - como abrir um novo laboratório de IA, defender mais cautela e salvaguardas no desenvolvimento atual de Inteligência Artificial, etc., e dizer se ela teria resultados positivos.

Ações vigorosamente úteis

Apesar desse estado de incerteza, aqui estão algumas medidas que, claramente, seriam de grande valor:

Pesquisa técnica sobre o problema de alinhamento. Alguns pesquisadores trabalham na construção de sistemas de Inteligência Artificial que podem obter "melhores resultados" (ganhar mais jogos de tabuleiro, classificar mais imagens corretamente, etc.) Mas um grupo menor de pesquisadores trabalha em coisas como:

- **Treinar sistemas de Inteligência Artificial para incorporar feedback humano em como eles executam tarefas de resumo**, para que os sistemas de Inteligência Artificial reflitam preferências humanas difíceis de definir — algo que seria importante que eles consigam fazer no futuro.
- **Descobrir como entender "o que os sistemas de Inteligência Artificial estão pensando e como eles estão raciocinando,"** para torná-los menos misteriosos.
- **Descobrir como impedir que os sistemas de Inteligência Artificial façam julgamentos extremamente ruins sobre imagens projetadas para enganá-los**, e outros trabalhos focados em ajudar a evitar os comportamentos de "pior caso" dos sistemas de IA.
- **Trabalhos teóricos** sobre como um sistema de Inteligência Artificial pode ser muito avançado, mas não ser imprevisível da maneira errada.

Esse tipo de trabalho reduziria o risco das consequências negativas da Inteligência Artificial desalinhada — e/ou conduziria a um melhor entendimento sobre o tamanho da ameaça. Alguns desses trabalhos são na academia, alguns em laboratórios de Inteligência Artificial e alguns em organizações especializadas.

Busca de clareza estratégica: fazer pesquisas que possam abordar outras questões cruciais (como as listadas acima), para esclarecer quais tipos de ações imediatas seriam mais úteis.

Ajudar governos e sociedades, a bem, melhorarem. Ajudar o País X a ficar à frente de outros no desenvolvimento de Inteligência Artificial pode melhorar ou piorar as coisas, pelas razões já apresentadas acima. Mas seria muito bom trabalhar a favor de um País X com valores melhores e mais inclusivos, e um governo cujos principais tomadores de decisão tenham maior probabilidade de tomar decisões ponderadas e baseadas em bons valores.

Disseminar ideias e construir comunidades. Hoje, parece-me que o mundo está **extremamente carente de pessoas que compartilhem certas expectativas e preocupações** básicas, como:

- Acreditar que a pesquisa de Inteligência Artificial pode levar a mudanças rápidas e radicais do tipo extremo apresentado aqui (muito além de coisas como, por exemplo, o aumento do desemprego).
- Acreditar que o problema do alinhamento (discutido acima) é pelo menos uma preocupação plausível e levar o modelo de "cautela" a sério.
- Analisar a situação pela perspectiva de "Vamos obter o melhor resultado possível para o mundo inteiro para o futuro de longo prazo", em vez de perspectivas mais comuns, como "Vamos tentar ganhar dinheiro" ou "Vamos tentar garantir que a minha instituição/meu país seja o líder mundial em pesquisa de IA."

Acho muito valioso que haja mais pessoas com essa perspectiva básica, principalmente trabalhando para laboratórios de Inteligência Artificial e governos. Se e quando tivermos mais clareza estratégica sobre quais ações maximizariam as chances de o "século mais importante" correr bem, espero que essas pessoas estejam relativamente bem-posicionadas para serem úteis.

Várias organizações e pessoas trabalharam para expor as pessoas às perspectivas acima e ajudá-las a conhecer outras pessoas que compartilham da mesma visão. Acho que boa parte do progresso (em termos de comunidades em crescimento) é resultado dessas ações.

Doar? Pode-se doar hoje para lugares como **este**. Mas preciso admitir que, falando de maneira muito ampla, não há uma tradução fácil agora entre "dinheiro" e "melhorar as chances de que o século mais importante corra bem"(o que isso quer dizer?). Não é o caso que, se alguém simplesmente enviasse, digamos, US$ 1 trilhão para o lugar certo, todos poderíamos respirar tranquilos sobre desafios como o problema de alinhamento e os riscos de distopias digitais**.**

Parece-me que nós — como espécie — estamos terrivelmente carentes de pessoas que prestem atenção aos desafios mais importantes à nossa frente, e não trabalhamos para ter uma boa clareza estratégica sobre quais ações tangíveis tomar. **Não podemos resolver esse problema com dinheiro.**[128] **Primeiro, precisamos levá-lo mais a sério e entendê-lo melhor.**

Notas

[116]De Prevendo a Inteligência Artificial Transformadora: qual é o ônus da prova?**:** prevejo mais de 10% de chance da Inteligência Artificial transformadora ser desenvolvida dentro de 15 anos (até 2036); aproximadamente 50% em 40 anos (até 2060); e uma chance de aproximadamente 2/3 de ser desenvolvida neste século (até 2100).

[117]Isso inclui os livros ***Superintelligence***, ***Human Compatible***, ***Life 3.0***, e ***The Alignment Problem***. A apresentação mais curta e acessível que conheço é **The case for taking AI seriously as a threat to humanity** (Artigo na Vox por Kelsey Piper). Este ***report on existential risk from power-seeking AI*** **(relatório sobre o risco existencial de uma Inteligência Artificial que visa poder)**, por Joe Carlsmith da Open Philanthropy, estabelece um conjunto detalhado de premissas que coletivamente implicariam que o problema é sério.

[118]A ordem de classificação não é absoluta, é claro. Existem versões de "Maturidade Tecnológica Adversária" que seriam piores do que "IA Desalinhada" — por exemplo, se a primeira resultar em poder indo para aqueles que deliberadamente infligem sofrimento.

[119]Parte da razão para isso é que os mais rápidos e menos cuidadosos acabariam superando rapidamente os outros e determinando o futuro da galáxia. Há também um risco de longo prazo discutido em ***The Future of Human Evolution*** **(O futuro da evolução humana)** de Nick Bostrom; veja também **esta discussão** das ideias de Bostrom no Slate Star Codex, mas também veja **este artigo de Carl Shulman** argumentando ser improvável que essa dinâmica resulte na eliminação total de coisas boas.

[120]Veja aqui

[121]Por exemplo, veja esta seção de **Pessoas digitais seriam mais importantes ainda**.

[122]Um artigo relevante: ***Public Policy and Superintelligent AI: A Vector Field Approach*****(Políticas públicas e Inteligência Artificial superinteligente: uma abordagem de campo vetorial)**de Bostrom, Dafoe e Flynn.

[123]Maturidade tecnológica adversária refere-se a um mundo onde a tecnologia altamente avançada **já foi desenvolvida**, provavelmente com a ajuda da IA, e diferentes coalizões estão competindo por influência sobre o mundo. Por outro lado, "Competição" refere-se a uma estratégia de como se comportar **antes do desenvolvimento de Inteligência Artificial avançada.** Pode-se imaginar um mundo em que algum governo ou coalizão adota um quadro de "competição", desenvolve Inteligência Artificial avançada muito antes de outros e, em seguida, toma uma série de boas decisões que impedem a Maturidade Tecnológica Adversária. (Ou, inversamente, um mundo onde o fracasso em se sair bem na "competição" aumenta os riscos de Maturidade Tecnológica Adversária.)

[124]Veja as definições deste problema na **Wikipedia** e **no Medium de Paul Christiano**.

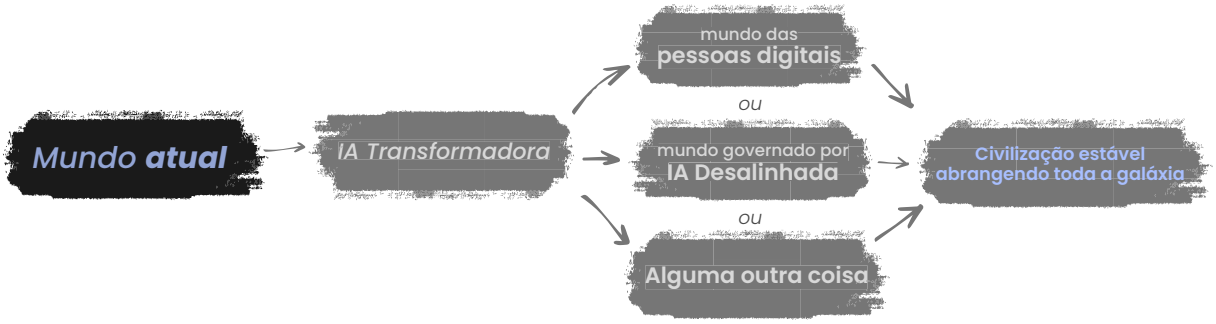
[125]Uma pesquisa privada mais detalhada feita para **este relatório**, perguntando sobre a probabilidade de "desgraça" antes de 2070 devido ao tipo de problema discutido no relatório, obteve respostas variando de <1% a >50%. Na minha opinião, há pessoas muito ponderadas que consideraram seriamente esses assuntos em ambas as extremidades desse intervalo.

[126]Alguns exemplos de tópicos técnicos aqui.

[127]Algumas discussões sobre este tópico aqui: **Distinção de definições de decolagem -** ***AI Alignment Forum***

[128]Mais algumas reflexões sobre "quando o dinheiro não é suficiente" **nesta postagem antiga da** ***GiveWell***.

Uma chamada à vigilância



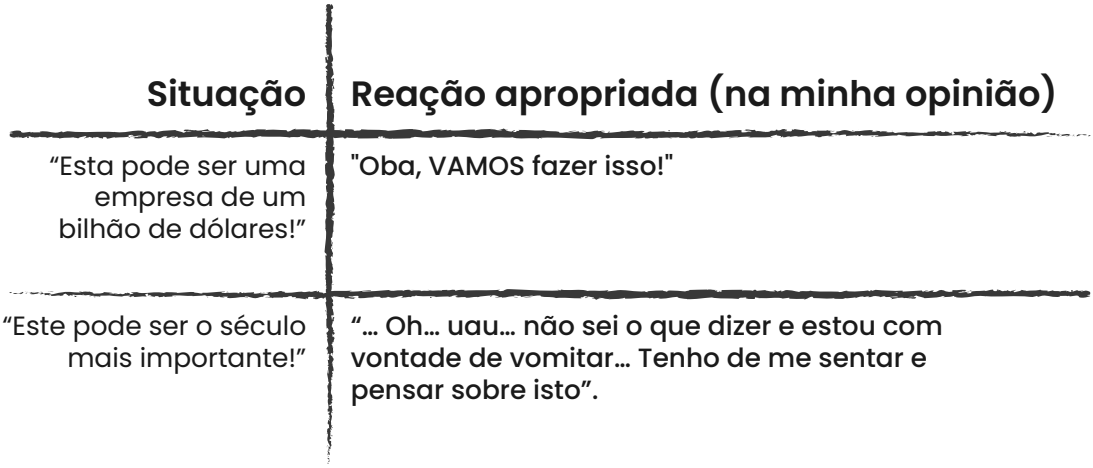
Este é o artigo final da série "o século mais importante", que argumentou haver uma alta probabilidade[129] que as próximas décadas verão:

- O desenvolvimento de uma tecnologia como o PASTA (processo para automação do avanço científico e tecnológico).
- Uma explosão de produtividade resultante, levando ao desenvolvimento de novas tecnologias transformadoras.
- Os primórdios de uma civilização estável em toda a galáxia, possivelmente povoada por pessoas digitais, ou gerida por uma Inteligência Artificial desalinhada.

Ao tentar chamar a atenção para um problema subestimado, **é comum concluir com uma "chamada à ação":** uma ação tangível e concreta que os leitores podem realizar para ajudar.

Mas isso é um desafio, porque, como argumentei anteriormente, há muitas **p**erguntas não respondidas sobre quais ações seriam úteis ou prejudiciais. (Embora possamos identificar algumas ações vigorosamente úteis atualmente.)

Isso cria uma situação um tanto embaraçosa. Ao confrontar a hipótese do "século mais importante", a minha atitude não corresponde às atitudes comuns de "entusiasmo e movimento" ou "medo e evitação". Em vez disso, sinto uma **estranha mistura de intensidade, urgência, confusão e hesitação.** Estou analisando um problema maior do que jamais esperei enfrentar, sentindo-me desqualificado e ignorante sobre o que fazer a seguir. Este é um sentimento difícil de compartilhar e espalhar, mas estou tentando.



Então em vez de uma chamada à ação, eu quero fazer uma **chamada à vigilância.** Caso você tenha sido convencido pelos argumentos desta série, então não se apresse a “fazer alguma coisa” e depois seguir em frente. Em vez disso, realize hoje todas as ações vigorosamente boas que puder e, caso contrário, coloque-se numa posição melhor para realizar ações importantes quando a hora chegar.

Isso poderia ser:

- Encontrar maneiras de interagir mais e aprender mais sobre os principais tópicos/campos/indústrias, tais como a Inteligência Artificial (por razões óbvias), ciência e tecnologia em geral (já que muitas das hipóteses do “século mais importante” analisam uma explosão no avanço científico e tecnológico), e áreas relevantes de política e segurança nacional.
- Aproveitar as oportunidades (quando você as identificar) de direcionar sua carreira para áreas com maior probabilidade de serem relevantes (alguns pensamentos meus sobre isso estão **aqui**; veja também ***80,000 Hours*** **(80.000 horas)**).
- Conectar-se com outras pessoas interessadas nesses tópicos (acredito que isso tenha sido uma das maiores fontes de motivação para as pessoas que realizaram trabalhos de alto impacto no passado). Atualmente, acho que a comunidade do **altruísmo eficaz** é o melhor local para isso, e você pode saber mais como se conectar com as pessoas por meio do **Centro de Altruísmo Eficaz** (veja a seção “Envolva-se” do menu suspenso). Se novas formas de engajamento surgirem no futuro, provavelmente irei publicá-las no Cold Takes.
- E, claro, aproveitar todas as oportunidades que você tiver para realizar **a**ções vigorosamente úteis**.**

Botões que você pode clicar

Aqui está algo que você pode fazer agora que seria genuinamente útil, embora talvez não tão visceralmente satisfatório quanto assinar uma petição ou fazer uma doação.

No meu **trabalho**, muitas vezes estou — ou alguém com quem trabalho está — procurando por um tipo específico de pessoa (talvez para preencher uma vaga de emprego como bolsista ou para compartilhar sua experiência sobre algum tópico, ou outra coisa). Com o tempo, espero que haja cada vez mais oportunidades para pessoas com habilidades, interesses, conhecimentos específicos, etc., realizarem ações que ajudem a tirar o melhor proveito do século mais importante. E acredito que um grande desafio será simplesmente **saber quem está por aí** — quem está interessado nesta causa e quer ajudar, e quais habilidades e interesses essas pessoas têm.

Se você é uma pessoa que gostaríamos de encontrar no futuro, você poderia nos ajudar agora mesmo enviando suas informações por meio **deste formulário simples**. Garanto que suas informações não serão vendidas ou usadas para ganhar dinheiro, que suas preferências de comunicação (sobre as quais o formulário pergunta em detalhes) serão respeitadas e que você sempre poderá cancelar qualquer comunicação.

O site 80000 Horas está realizando uma pesquisa para identificar pessoas interessadas nos riscos catastróficos provenientes da IA. Ao preenchê-la você passa a fazer parte de um cadastro e pode ser contatado quando surgirem vagas de emprego adequadas ao seu perfil. Esta pesquisa é mais recente e provavelmente será mais utilizada que o formulário mencionado anteriormente. Você pode preenchê-la aqui. (LINK)

Compartilhando uma mentalidade

Em Isto não pode continuar, eu fiz uma analogia do mundo com pessoas em um avião acelerando na pista, sem saber por que estão se movendo tão rápido ou o que está por vir:

Como alguém sentado dentro deste avião, eu adoraria poder dizer que descobri exatamente o que está acontecendo e para qual futuro precisamos nos planejar. Mas eu não sei.

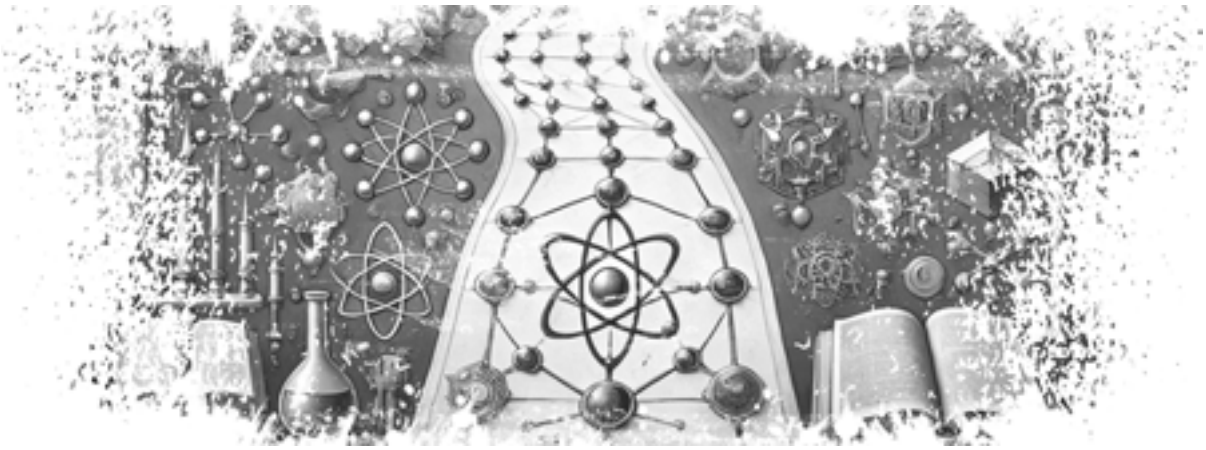
Na falta de respostas, tentei pelo menos mostrar o que vejo:

- Esboços obscuros dos eventos mais importantes no passado ou futuro da humanidade.
- Um caso em que eles estão se aproximando de nós mais rapidamente do que parece — estejamos prontos ou não.
- Uma sensação de que o mundo e as regras com as quais estamos acostumados não são confiáveis. Que precisamos dirigir nosso olhar para além da torrente diária de notícias tangíveis e relacionáveis — e tentar finalmente entender melhor assuntos mais estranhos e audaciosos, que provavelmente serão vistos como **as manchetes sobre esta era daqui a bilhões de anos.**

Há muito que não sei. Mas se este é o século mais importante, estou confiante de que nós, como civilização, ainda não estamos à altura dos desafios que ele apresenta. Para que isso mude, é preciso começar com mais pessoas vendo a situação como ela realmente é, levando-a a sério, agindo quando for possível — e quando não for, mantendo-se vigilantes.

Notas

[129]"Prevejo mais de 10% de chance de Inteligência Artificial transformadora ser desenvolvida dentro de 15 anos (até 2036); aproximadamente 50% disso acontecer dentro de 40 anos (até 2060); e aproximadamente 2/3 de chance que isso ocorra ainda neste século (até 2100)."



Apêndices

Ponto fraco de “O século mais importante”: automação completa

Agora que recebi algumas reações e críticas, achei que seria bom escrever algumas postagens cobrindo os que considero serem os pontos mais fracos da série “o século mais importante”.

Atualmente, acredito que o ponto mais fraco da série é mais ou menos assim:

- É verdade que se a Inteligência Artificial pudesse *literalmente* automatizar *tudo* o que fosse necessário para causar avanço científico e tecnológico, as consequências descritas na série (uma aceleração dramática no avanço científico e tecnológico, levando a um futuro radicalmente desconhecido) se concretizariam.

- E, se a Inteligência Artificial pudesse automatizar apenas *99%* do que é necessário para o avanço científico e tecnológico? E se os sistemas de Inteligência Artificial pudessem propor experimentos, mas não conseguissem executá-los? E, se eles pudessem propor experimentos e executá-los, mas não conseguissem autorização regulatória para executá-los? **Nesse caso, é plausível que 1% das coisas que as IAs não poderiam fazer de forma rápida e automática seriam um “gargalo” no progresso, levando a um crescimento drasticamente menor.**

- A série cita a opinião de especialistas sobre quando a Inteligência Artificial transformadora será desenvolvida. *Tecnicamente falando*, o tipo de situação que os entrevistados estão prevendo — “máquinas independentes podem realizar todas as tarefas melhor e mais barato do que trabalhadores humanos” — deveria ser suficiente para uma explosão de produtividade. Mas os respondentes da pesquisa podem estar pensando em um tipo de Inteligência Artificial *um pouco* menos poderosa do que está literalmente implícito nessa afirmação — o que poderia levar a impactos drasticamente menores. Ou elas podem estar imaginando que mesmo IAs com capacidade intelectual comparável a dos humanos, podem ainda não ter a habilidade prática para realizar tarefas cruciais, porque (por exemplo) os humanos não confiam instintivamente nelas. De qualquer forma, eles (os entrevistados da pesquisa) podem estar imaginando algo quase tão capaz — mas não tão impactante — quanto o tipo de Inteligência Artificial que discuto aqui.

- Além disso, mesmo que as IAs conseguissem fazer tudo o que os humanos fazem para automatizar o avanço científico e tecnológico, seu progresso científico e tecnológico teria que aguardar os resultados de experimentos reais, o que poderia atrasá-los muito.

Resumindo: **uma pequena lacuna no que a Inteligência Artificial pode automatizar pode levar a um impacto muito menor do que a série implica.** Automatizar "quase tudo" seria muito diferente de automatizar tudo.

Este é um contexto importante para as tentativas de prever a Inteligência Artificial transformadora: eles estão realmente prevendo algo bastante extremo.

Minha resposta

Acho que tudo o que foi dito acima está correto: de fato, precisaríamos de níveis extremos de automação para produzir as consequências que imagino. (Pode haver algumas tarefas que precisam ser realizadas por humanos, mas elas devem ser um conjunto bem pequeno e limitado para evitar que as coisas fiquem muito lentas devido a gargalos.)

Também é verdade que não expliquei como essa automação extrema poderia ser alcançada — como cada atividade necessária para promover o avanço científico e tecnológico (incluindo a execução de experimentos e a espera de sua conclusão) poderia ser realizada de maneira rápida e/ou automatizada, sem gargalos humanos ou outros que atrasam muito as coisas.

Tendo reconhecido isso, também vale a pena ressaltar que os níveis extremos de automação **não precisam se aplicar a toda a economia: a automação extrema para um conjunto relativamente pequeno de atividades seria suficiente para chegar às conclusões da série.**

Por exemplo, poderia ser o suficiente para sistemas de Inteligência Artificial desenvolver: (a) computadores cada vez mais eficientes; (b) painéis solares (para energia); (c) robôs de mineração e manufatura; (d) sondas espaciais (para construir mais computadores no espaço, onde a energia e o metal são abundantes). Isso seria suficiente (via retroalimentação) para um crescimento explosivo na energia, materiais e poder de computação disponível, e há muitas maneiras pelas quais esse crescimento seria transformador.

Por exemplo, isso poderia levar a:

- IA desalinhada com acesso a quantidades perigosas de materiais e energia.
- Pessoas digitais, se os sistemas de Inteligência Artificial também tivessem alguma forma de (a) "virtualizar" a neurociência (por meio de experimentos virtuais ou simplesmente aumentando dramaticamente a taxa de aprendizado de experimentos reais); ou (b) ter, por outro lado, uma visão sobre como criar algo que consideraríamos apropriadamente como "descendentes digitais".

Conclusão

Acho que não demonstrei completamente (ou, para leitores com forte ceticismo inicial sobre esse ponto, convincentemente) que a Inteligência Artificial avançada causará uma aceleração explosiva no avanço científico e tecnológico, sem atingir "gargalos" dependentes dos humanos ou outros. Acho que dei uma boa noção da intuição de porque eles poderiam, mas esse é certamente um tópico no qual não me aprofundei o máximo que pude; espero e suponho que alguém o faça eventualmente.

Eu realmente acredito que essa cutucada acabará por corroborar o panorama que apresentei na série "século mais importante". Isso é parcialmente baseado no raciocínio acima: o escopo relativamente limitado do que precisaria ser totalmente automatizado para sustentar minhas conclusões gerais. Também é parcialmente baseado em um processo de raciocínio semelhante ao que usei no passado para **prever algumas conclusões principais antes de fazermos todo**

o dever de casa: envolver-se em muitas conversas e formar pontos de vista sobre o quão informados estão os diferentes grupos envolvidos e o quanto eles estão entendendo sobre isso. Mas reconheço que isso não é tão satisfatório ou confiável quanto seria se eu desse uma descrição altamente detalhada de quais atividades específicas poderiam ser automatizadas.

Ponto fraco de "O século mais importante": aprisionamento/lock-in

Esta é a segunda de (por enquanto) duas postagens que cobrem os que considero serem os pontos mais fracos da série "o século mais importante". (A primeira delas está aqui.)

O ponto fraco que abordarei aqui é a discussão sobre o "aprisionamento/lock-in": a ideia de que a Inteligência Artificial transformadora poderia levar a sociedades **estáveis por bilhões de anos.** Se for verdade, isso significaria que o andamento das coisas neste século afetaria como a vida seria, de maneiras previsíveis e sistemáticas por períodos insondáveis.

Minha cobertura principal deste tópico está em uma seção do meu artigo sobre pessoas digitais. É bastante superficial, não está supercompleto e não é embasado em um relatório técnico detalhado (embora eu faça um link para **algumas observações informais** do físico **Jess Riedel** feitas por ele quando ele trabalhava na Open Philanthropy). ***Overcoming Bias* (Superando o viés)** me criticou neste ponto, causando uma **breve discussão na seção de comentários**.

Não serei dramaticamente mais completo ou convincente aqui, mas direi um pouco mais sobre como o argumento geral do "século mais importante" é afetado se ignorarmos essa parte dele e um pouco mais sobre porque acho o aprisionamento um cenário plausível.

(Observe também que o "aprisionamento" será discutido com mais detalhes em um próximo livro de Will MacAskill, What We Owe the Future (O que devemos ao futuro).)

Ao longo deste artigo, usarei aprisionamento/*lock-in* para significar que coisas importantes sobre a sociedade, como quem está no poder ou quais religiões/ideologias são dominantes, são aprisionadas no lugar indefinidamente, plausivelmente por bilhões de anos" e "dinamismo" ou "instabilidade" para significar o oposto: "tais coisas mudam em horizontes de tempo muito mais curtos, como em décadas/séculos/milênios". Como observado anteriormente, considero o aprisionamento/*lock-in* uma possibilidade assustadora por definição, embora seja imaginável que certos tipos de aprisionamento (por exemplo, de proteção aos direitos humanos) seriam bons.

"O século mais importante" sem aprisionamento

Primeiro, vejamos o que acontece se descartarmos toda essa parte do argumento e assumirmos que o aprisionamento não é uma possibilidade, mas ainda aceitar o resto das informações. Em outras palavras, presumimos que:

- Algo como o PASTA (IA avançada que automatiza o avanço científico e tecnológico) será provavelmente desenvolvida neste século.
- Isso, no que lhe concerne, levaria a um avanço científico e tecnológico explosivo, resultando em um mundo governado por pessoas digitais ou Inteligência Artificial desalinhada, ou qualquer outra coisa, de forma que será justo dizer que: "transitamos para um estado onde os humanos, tais como os conhecemos, não são mais a força principal nos eventos mundiais."
- Isso *não* levaria a que nenhum aspecto particular do mundo fosse permanentemente imutável. Restariam bilhões de anos com muitos desenvolvimentos imprevisíveis.

Nesse caso, acredito que ainda há um sentido importante em que este seria o "século mais importante para a humanidade": seria nossa última chance de moldar a transição de um mundo governado por humanos para um mundo governado por algo muito diferente dos humanos. Esta é uma das duas definições de "século mais importante" dadas aqui.

Mais amplamente, neste caso, acredito que há um sentido importante em que a série "o século mais importante" deve ser pensada como "Apontando para uma questão drasticamente subestimada; correta em suas implicações mais consequentes e controversas, senão em todos os detalhes". Quando as pessoas falam sobre as questões mais importantes de nosso tempo (na verdade, mesmo quando estão falando especificamente sobre as **prováveis consequências da Inteligência Artificial avançada**), elas raramente incluem muita discussão sobre os tipos de questões enfatizadas nesta série; e deveriam fazê-lo, quer esta série esteja correta ou não, sobre a possibilidade de "aprisionamento/lock-in".

Como observado aqui no final das contas, eu me preocupo mais se a série do "século mais importante" está correta nesse sentido — apontando para questões drasticamente subestimadas — do que sobre a probabilidade de seu título acabar descrevendo a realidade. (Embora eu me importe com ambas as coisas.) É por esta razão que acredito que a discussão relativamente leve sobre aprisionamento/*lock-in* é um "ponto fraco" menos importante do que escrevi, o que levanta questões sobre se a Inteligência Artificial avançada mudaria muito o mundo em muitos aspectos ou muito rapidamente.

Mas incluí a menção do aprisionamento/*lock-in* porque acredito ser uma possibilidade real e tornaria as apostas deste século ainda mais altas.

Dissecando o "aprisionamento"

Provavelmente existiram muitas pessoas na história (imperadores, ditadores) com enorme poder sobre sua sociedade, que gostariam de manter as coisas como estavam para sempre. Também pode ter havido momentos em que governos eleitos democraticamente teriam "aprisionado" pelo menos algumas coisas sobre sua sociedade para sempre, se pudessem.

Mas não conseguiram. Por que não?

Acho que as razões se enquadram em algumas categorias, e pessoas digitais (ou Inteligência Artificial desalinhada, mas vou me concentrar nas pessoas digitais para manter as coisas simples por enquanto) pode mudar um pouco o cenário.

Primeiro listarei os fatores que parecem particularmente suscetíveis de serem alterados pela tecnologia, depois um fator que parece menos suscetível.

Fatores que parecem particularmente suscetíveis de serem alterados pela tecnologia

Envelhecimento e morte. Qualquer pessoa poderosa tem que morrer em algum momento. Eles podem tentar transferir o poder para filhos ou aliados, mas muitas coisas mudam na transferência (e em períodos muito longos de tempo, há muitas transferências).

Pessoas digitais não precisariam morrer ou envelhecer. (Mais amplamente, avanços suficientes em ciência e tecnologia parecem prováveis de conseguir eliminar o envelhecimento e a morte, mesmo que não seja por meio de pessoas digitais.) Portanto, se algum conjunto específico deles tivesse poder sobre alguma parte específica da galáxia, a morte e o envelhecimento não precisariam interferir aqui.

Outras mudanças populacionais. Com o tempo, a composição de qualquer população muda e, em particular, uma geração substitui a anterior. Isso tende a levar a mudanças de valores e dinâmicas de poder.

Sem envelhecimento ou morte, e com produtividade extrema, acabaríamos esgotando rapidamente a capacidade de carga de qualquer área específica — de modo que essa área não sofreria nenhuma alteração na composição populacional (ou ocorreriam alterações muito menores e mais controladas do que estamos acostumados atualmente — não há casos em que toda uma geração é substituída por uma nova). A rotatividade geracional parece um grande impulsionador do dinamismo até hoje.

Caos. Até hoje, mesmo quando algum governo é oficialmente "responsável" por uma sociedade, ele tem uma capacidade muito limitada de monitorar e intervir em tudo o que está acontecendo. Mas acredito que o avanço tecnológico até hoje já aumentou muito a capacidade de um governo de exercer controle sobre inúmeras pessoas e um vasto território. Uma explosão no avanço científico e tecnológico aumentaria ainda mais radicalmente o controle, na prática, dos governos sobre o que acontece.

(Pessoas digitais fornecem um exemplo extremo: controlar o servidor que roda um ambiente virtual significaria conseguir monitorar e controlar tudo sobre as pessoas naquele ambiente. E figuras poderosas conseguiriam criar muitas cópias de si mesmas para monitoramento e execução.)

Eventos naturais Todos os tipos de coisas podem perturbar uma sociedade humana: mudanças no tempo/clima, falta de recursos, etc. Avanços suficientes em ciência e tecnologia podem levar esse tipo de disrupção a níveis extremamente baixos (e, em particular, pessoas digitais têm necessidades de recursos bastante limitadas, de modo que não precisariam de recursos por bilhões de anos).

Busca por aperfeiçoamento Enquanto alguns ditadores e imperadores prefeririam manter as coisas como estão para sempre, a maioria dos governos atuais não tende a ter isso como uma aspiração: os funcionários eleitos se consideram responsáveis perante grandes populações cujas vidas estão tentando melhorar.

Mas avanços dramáticos em ciência e tecnologia significariam dramaticamente mais controle sobre o mundo, bem como potencialmente menos espaço para melhorias *adicionais* (suponho que a velocidade das melhorias deva diminuir em algum momento). Isso tornaria cada vez mais provável que algum governo ou organização decida que prefere manter as coisas como estão.

Mas esses fatores poderiam ser eliminados tão completamente a ponto de causar estabilidade por bilhões de anos? Acho que sim, se uma parcela suficiente da sociedade fosse digital (por exemplo, pessoas digitais**)**, de modo que aqueles que buscam estabilidade usassem a correção digital de erros (essencialmente, fazer várias cópias de qualquer coisa importante, que seria usada para reverter qualquer mudança que aconteça, por qualquer motivo — para saber mais, veja as **observações informais de Jess Riedel**, que defende que a correção digital de erros poderia ser usada para atingir níveis bastante extremos de estabilidade).

Um exemplo tangível aqui seria ambientes virtuais rigidamente controlados, contendo pessoas digitais, programados para reiniciar totalmente (ou reiniciar propriedades principais) qualquer elemento principal que tivesse sido alterado. Estes exemplos representam uma maneira hipotética de eliminar, essencialmente, todos os fatores acima como fontes de mudança.

Mas mesmo que prefiramos evitar pensar em tais cenários específicos, penso que há casos mais amplos para avanços científicos e tecnológicos explosivos, reduzindo radicalmente o papel de cada um desses fatores, conforme descrito acima.

É claro que só porque algum governo *conseguiu* conquistar o "aprisionamento" isso não quer dizer que ele *o faria*. Mas após um longo tempo, as sociedades contra-aprisionamento simplesmente teriam cada vez mais oportunidades de se tornarem sociedades pró-aprisionamento,

enquanto mesmo alguns poucos anos de uma sociedade "pró-aprisionamento" resultaria em aprisionamento por tempo indefinido. (E em um mundo de pessoas digitais operando muito mais rápido que os humanos, muito "tempo" poderia passar até o final deste século.)

Um fator que parece particularmente suscetível de ser alterado pela tecnologia: competição entre sociedades

Mesmo que um governo tivesse controle total sobre sua sociedade, isso não garantiria estabilidade, porque ele sempre poderia ser atacado por agentes externos. E **ao contrário dos fatores acima, isso não é nada que avanços radicais em ciência e tecnologia parecem particularmente propensos a mudar:** em um mundo de pessoas digitais, diferentes governos ainda conseguiriam se atacar reciprocamente e negociar entre si, tendo a ameaça de ataque ao fundo.

Isso causaria uma instabilidade permanente, de modo que o mundo estaria em constante mudança. Este é o ponto enfatizado pela **crítica do *Overcoming Bias* (Superando o viés)**.

Acho que essa dinâmica poderia — ou não — ser uma fonte duradoura de dinamismo. Algumas razões pelas quais isso pode não acontecer:

- Se a Inteligência Artificial causar uma explosão no avanço científico e tecnológico, quem a desenvolvesse primeiro rapidamente se tornaria muito poderoso — ser "o primeiro a desenvolver o PASTA em alguns meses" significaria, efetivamente, o mesmo que desenvolver o equivalente à conquista de vários séculos de liderança em ciência e tecnologia depois disso. Isso levaria à consolidação do poder na Terra, e não há sinais de vida inteligente fora da Terra — então seria o fim da dinâmica do "ataque" como força de instabilidade.
- A consciência do risco mencionado acima, faria com que as principais potências negociassem e dividissem explicitamente a galáxia, comprometendo-se (talvez forçosamente,

dependendo de como o panorama tecnológico tiver sido abalado) para nunca invadir o território um do outro. Nesse caso, qualquer parte específica da galáxia estaria livre de ataques.

- Pode acontecer que os assentamentos espaciais sejam geralmente mais fáceis de defender do que atacar, de modo que, uma vez que alguém os estabeleça, eles essencialmente não estarão sujeitos a ataques.

Qualquer um dos itens acima, ou uma combinação deles (por exemplo, ataques são possíveis, mas arriscados e caros; as potências mundiais optam por não atacar umas às outras para não desencadear uma guerra) levariam ao desaparecimento permanente da competição militar como um fator, e abririam a possibilidade de alguns governos "aprisionarem" características-chave de suas sociedades.

Três categorias de futuro de longo prazo

Acima, listei alguns fatores que poderiam — ou não — continuar sendo fontes de dinamismo mesmo após um avanço científico e tecnológico explosivo. Acho que comecei a entender por que, no mínimo, as fontes de dinamismo seriam muito *reduzidas* no caso de pessoas digitais ou outras tecnologias radicalmente avançadas, em comparação com o que temos atualmente.

Agora quero dividir os diferentes futuros possíveis em três grandes categorias:

Aprisionamento discricionário total. É aqui que um determinado governo (ou coalizão, ou configuração negociada) consegue aprisionar, essencialmente, quaisquer propriedades que escolher para sua sociedade, indefinidamente.

Isso pode acontecer se essencialmente todas as fontes de dinamismo descritas acima desaparecerem e os governos optarem por buscar o aprisionamento.

Dinâmica competitiva previsível. Acho que a fonte de dinamismo com maior probabilidade de persistir (em um mundo de pessoas digitais ou ciência e tecnologia comparativamente avançadas) é a última discutida na seção acima: competição militar entre sociedades avançadas.

No entanto, acredito que persistiria de uma forma que tornaria **os resultados de longo prazo previsíveis.** Na verdade, acho que "resultados de longo prazo previsíveis de uma forma importante" fazem parte da visão implícita na **crítica do *Overcoming Bias* (Superando o viés)**, que argumenta que o mundo precisará ser quase exclusivamente povoado por seres que passem quase toda a existência trabalhando (já que a população crescerá a ponto de ser preciso trabalhar constantemente para sobreviver).

Se todos os seres digitais controlassem tudo no mundo, exceto se precisassem lutar contra outros países, isso poderia causar grandes problemas para os seres digitais que são mais ambiciosos, produtivos, trabalhadores e agressivos. Estes seriam seres que fariam pouco mais do que apenas lutar por recursos.

Dinamismo verdadeiro. Em vez de um mundo onde governos confiscam propriedades arbitrariamente ou seres digitais competem com resultados previsíveis, poderíamos ter um mundo com verdadeira liberdade e dinamismo. Isso poderia ser alcançado por meio de medidas que impeçam o confisco e a competição desenfreada, promovendo a diversidade e até mesmo a aleatoriedade.

E acabarmos vivendo em qualquer um desses três mundos, e este século determinar em qual (ou qual mistura) deles viveremos, isso constitui um bom argumento a favor da ideia de que este século teria impactos especialmente notáveis, e, portanto, ele seria o século mais importante de todos os tempos para a vida inteligente.

Por exemplo, do ponto de vista atual, temos a mesma probabilidade de (a) viver em um mundo onde governos poderosos utilizariam o aprisionamento, (b) viver em um mundo onde a competição desenfreada levaria a galáxia a dominação pelos seres mais fortes/produtivos/agressivos, ou (c) um mundo verdadeiramente dinâmico onde eventos futuros seriam imprevisíveis e importantes.

Nesse caso, se terminarmos com (c), e os eventos futuros acabarem sendo extremamente interessantes e consequentes, eu pensaria que ainda haveria um sentido importante em que o desenvolvimento mais importante de todos os tempos *fosse o estabelecimento dessa mesma dinâmica.* (Dado que uma das outras duas alternativas acabaria determinando a forma da civilização em toda a galáxia, a longo prazo.) Outra forma de dizer isso: se o aprisionamento/*lock-in* (e/ou previsivelmente a dinâmica competitiva) for uma possibilidade séria a partir deste século, a oportunidade de *evitá-lo* tornaria este século o mais importante.

Resumindo

Forneci muitos detalhes a respeito de futuros radicalmente desconhecidos, e os leitores podem estar com a sensação neste ponto, que as coisas ficaram muito específicas e complexas para serem consideradas seriamente. Mas acho que as intuições gerais aqui são bastante simples e sólidas, então darei um resumo de alto nível:

- Avanço científico e tecnológico reduziria ou eliminaria muitas das fontes atuais de instabilidade, desde o envelhecimento e a morte até o caos e eventos naturais. Uma explosão no avanço científico e tecnológico poderia, portanto, levar a uma grande redução do dinamismo. (E como um exemplo vívido, as pessoas digitais configurariam ambientes virtuais rigidamente controlados com correção de erros muito robusta, algo que considero uma possibilidade assustadora por padrão, conforme observado na introdução.)

- O dinamismo permaneceria ou não, dependendo de uma série de fatores sobre como o poder acabaria sendo consolidado e como diferentes governos/sociedades se relacionariam entre si. O "é possível ou não é" seria determinado neste século.

- Acho que esta é uma possibilidade séria o suficiente para aumentar as apostas a favor do "século mais importante", mas não estou muito confiante nesse pensamento aqui, e acredito que a maior parte do espírito da hipótese do "século mais importante" sobrevive mesmo se nos esquecermos de tudo isso.

Espero que essas considerações adicionais tenham fornecido um contexto útil sobre o que quero dizer, mas continuo reconhecendo que esta é uma das partes menos desenvolvidas da série e estou interessado em explorar mais o assunto.

"Âncoras biológicas" é sobre delimitar e, não, apontar, cronologias da IA

Anteriormente resumi o método das "âncoras biológicas" para a previsão da Inteligência Artificial transformadora de Ajeya Cotra, também conhecido como **"Bio-âncoras."** Aqui, quero tentar esclarecer por que considero esse método tão útil, *embora* concorde com a maioria das coisas específicas que ouvi as pessoas dizerem sobre seus pontos fracos (às vezes, pessoas que não conseguem entender por que eu daria qualquer importância a ela).

Algumas considerações preliminares:

- Provavelmente, esta postagem é de interesse, principalmente, dos céticos a respeito da teoria das Bio-âncoras e/ou pessoas que se sentem muito confusas/agnósticas sobre seu valor e gostariam de ler uma resposta dada aos céticos.

- Não quero dar a impressão de que estou fazendo novas críticas ao "Bio-âncoras" e pressionando por uma nova reinterpretação. Acho que a autora de "Bio-âncoras" concorda com a maioria do que digo sobre os pontos fracos do relatório e sobre como usá-lo melhor

Um resumo do que se trata o método

Apenas para restabelecer o contexto, aqui estão algumas citações importantes da minha postagem principal sobre o método das âncoras biológicas:

A ideia básica é:

Os modelos modernos de IA podem "aprender" a realizar tarefas por meio de um processo (financeiramente caro) conhecido como "treinamento". Você pode pensar no treinamento como uma enorme quantidade de tentativa e erro. Por exemplo, os modelos de IA de reconhecimento de voz recebem um arquivo de áudio de alguém falando, adivinham o que a pessoa está dizendo e recebem a resposta certa. Ao fazer isso milhões de vezes, eles "aprendem" a traduzir de forma confiável a fala em texto. Mais: **Treinamento**

Quanto maior for um modelo de IA e mais complexa a tarefa, mais caro será o processo de

treinamento [ou **“rodada de treinamento”**]. Alguns modelos de IA são maiores que outros; até o momento, nenhum deles chega nem perto de ser “tão grande quanto o cérebro humano” (o que isso significa será explicado a seguir).

Mais: Tamanho do modelo e tipo de tarefa

O método das âncoras biológicas pergunta: **“Com base nos padrões usuais, quanto custaria treinar um modelo de Inteligência Artificial tão grande quanto um cérebro humano para executar as tarefas mais difíceis que os humanos conseguem fazer? E quando isso será barato o suficiente para que alguém faça esse treinamento?”**

Mais: **Estimando os custos**

… O método propõe uma maneira de pensar sobre como pode ser simultaneamente verdadeiro que (a) os sistemas de Inteligência Artificial de uma década atrás não pareciam muito surpreendentes; (b) os sistemas de Inteligência Artificial atuais conseguem fazer muitas coisas surpreendentes, mas parecem ainda muito aquém do que os humanos conseguem fazer; (c) o desenvolvimento da Inteligência Artificial transformadora poderá, facilmente, acontecer nas próximas décadas - ou mesmo nos próximos 15 anos.

Além disso, acredito que vale a pena observar **alguns argumentos de alto nível** do Bio-âncoras que **não dependem de tantas estimativas e suposições:**

- Na próxima década, provavelmente veremos - pela primeira vez - modelos de Inteligência Artificial com “tamanho” comparável ao do cérebro humano.
- Se os modelos de Inteligência Artificial continuarem a se tornar maiores e mais eficientes nas taxas que o Bio-âncoras estima, provavelmente se tornará econômico **neste século atingir alguns marcos bastante extremos - o “ponto alto” do que a Bio-âncoras acha que seria necessário.** Estes são difíceis de resumir, mas veja as estruturas de “rede neural de horizonte longo” e “âncora de evolução” no relatório.
- Uma maneira de pensar sobre isso é que no próximo século passaremos provavelmente de “computação insuficiente para executar um modelo de tamanho humano” para “computação extremamente abundante, tanto quanto estimativas bastante conservadoras do que podemos precisar.” O poder computacional não é o único fator no progresso da IA, mas a medida que outros fatores (algoritmos, processos de treinamento) forem se tornando novos gargalos, provavelmente haverá incentivos poderosos (e várias décadas) para resolvê-los.

Pontos com os quais concordo sobre as fraquezas/limitações do método

O Bio-âncoras “age como se” a Inteligência Artificial fosse desenvolvida de uma maneira específica, e quase certamente ela não será

O Bio-âncoras, em um certo sentido, “age como se” a Inteligência Artificial transformadora fosse construída de uma maneira específica: **simples tentativa e erro de força bruta de tarefas computacionalmente intensivas** (conforme descrito aqui). Suas principais previsões são baseadas nesse panorama: o método estima quando haverá computação suficiente para executar uma certa quantidade de tentativa e erro e chama isso de “estimativa de quando a Inteligência Artificial transformadora será desenvolvida”.

Acho improvável que, se, e quando, a Inteligência Artificial transformadora for desenvolvida, como ela será desenvolvida se assemelhará a esse tipo de tentativa e erro cego de tarefas de horizonte longo.

Se eu tivesse que prever como a Inteligência Artificial transformadora será desenvolvida, seria mais parecido com isso aqui:

- Primeiro, os sistemas de Inteligência Artificial restrito provam seu valor na realização de um conjunto limitado de tarefas. (Isso já está acontecendo, em um grau limitado, por exemplo, nas tarefas de reconhecimento de voz, tradução e buscas).
- Isso levaria a (a) **mais atenção e financiamento para IA**; (b) maior integração da Inteligência Artificial na economia, de modo que seja mais fácil coletar **dados sobre como os humanos interagem com as IAs,** que poderiam ser usados para treinamento adicional; (c) maior conscientização geral sobre o que seria necessário para a Inteligência Artificial automatizar tarefas principais de maneira útil e, portanto, **maior conscientização (e atenção) das maiores barreiras que impedem a Inteligência Artificial de ser mais ampla e capaz.**
- Diferentes tipos de AIs restritas se integrariam em diferentes setores da economia. Com o tempo, o aumento de dados de treinamento, financiamento e atenção conduziria a IAs cada vez menos restritas, assumindo partes cada vez mais amplas das tarefas que elas já estivessem realizando. Essas mudanças não acontecem apenas por meio de modelos de Inteligência Artificial (e execuções de treinamento) crescentes; elas também são impulsionadas por inovações em como as IAs são projetadas e treinadas.
    - Em algum momento, alguma combinação de IAs conseguiria automatizar o suficiente o avanço científico e tecnológico para ser transformadora.

Não haveria uma única "rodada mestre" na qual uma única Inteligência Artificial seria treinada para executar as tarefas mais difíceis e amplas por meio de tentativa e erro cego.

O Bio-âncoras "age como se" a disponibilidade de poder computacional fosse o único grande obstáculo ao desenvolvimento da Inteligência Artificial transformadora, e ele, provavelmente, não é

Como observado na minha postagem anterior:

O Bio-âncoras pode ser considerado agressivo demais devido à sua suposição de que "o poder computacional é o gargalo":

Ele pressupõe que, se alguém pudesse pagar por todo o poder computacional necessário para fazer o "treinamento" de força bruta, descrito acima para as tarefas principais,(por exemplo, automatizar o trabalho científico), a IA transformadora (provavelmente) seria desenvolvida em seguida.

Treinar um modelo de IA não requer apenas comprar poder computacional. Requer a contratação de pesquisadores, realização de experimentos e, talvez o mais importante, encontrar uma maneira de configurar o processo de "tentativa e erro" para que a IA possa obter um grande número de "tentativas" na tarefa principal. Pode acontecer que fazer isso seja proibitivamente difícil.

É muito fácil imaginar mundos onde o desenvolvimento da Inteligência Artificial transformadora leva muito mais, ou menos tempo do que o Bio-âncoras sugere para acontecer, por razões que, essencialmente, não foram modeladas no Bio-âncoras.

Conforme implícito acima, a Inteligência Artificial transformadora pode levar muito tempo para ser desenvolvida por motivos como "é extremamente difícil obter dados e ambientes de treinamento para algumas tarefas cruciais" ou "algumas tarefas simplesmente não podem ser aprendidas mesmo com abundantes rodadas de tentativa e erro."

A Inteligência Artificial transformadora também poderia ser desenvolvida muito mais *rapidamente* do que o Bio-âncoras sugere. Por exemplo, algum avanço na forma como projetamos algoritmos de Inteligência Artificial - talvez inspirados pela neurociência - poderia levar ao desenvolvimento de IAs que conseguiriam fazer, praticamente, tudo o que o cérebro humano consegue, *sem* precisar da enorme quantidade de tentativa e erro que a Bio-âncoras estima (com base na extrapolação dos sistemas *atuais* de aprendizado de máquina).
Listei mais considerações como estas aqui.

O Bio-âncoras não está "apontando" o ano mais provável em que a Inteligência Artificial transformadora será desenvolvida

Minha compreensão dos modelos de mudança climática é que eles: tentam analisar **cada fator principal** que poderia aumentar ou abaixar a temperatura no futuro; então, produzir a melhor estimativa para cada um deles; e compilar todas essas previsões em uma previsão de como será a temperatura.

De certa forma, você pode considerá-las como **"apontamento da melhor estimativa"** (ou mesmo "simulação") da temperatura futura: embora elas não sejam certas ou precisas, elas estão identificando uma temperatura específica com base nos principais fatores que poderiam aumentar ou diminuir a temperatura.

Muitos outros casos em que alguém estima algo incerto (por exemplo, a população futura) têm propriedades semelhantes.

O Bio-âncoras não é assim. Existem fatores que ele ignora serem identificáveis atualmente e, que, são quase certamente importantes. Portanto, em algum sentido importante, ele não está "apontando" o ano mais provável para o desenvolvimento da Inteligência Artificial transformadora.

(Este não é o foco deste artigo) As estimativas do Bio-âncoras são muito incertas

O Bio-âncoras estima algumas coisas difíceis de prever, como:

- Qual tamanho um modelo de Inteligência Artificial teria que ter, para que ele seja "tão grande quanto o cérebro humano", em algum sentido relevante. (Para isso ele adapta **o relatório detalhado de Joe Carlsmith**.)
- Com que rapidez devemos esperar que a eficiência algorítmica, a eficiência do hardware e a "vontade de gastar muito com a IA" aumentem no futuro — tudo isso afeta a questão de "qual será o tamanho de uma rodada de Inteligência Artificial econômica". Suas estimativas aqui são muito simples e acredito que há muito espaço para melhorias, embora não espere que o panorama qualitativo mude radicalmente.

Reconheço uma incerteza significativa nessas estimativas e reconheço que (desde que tudo seja igual) quando há incerteza isso significa que devemos ser céticos.

Dito isto:

- Acho que essas estimativas provavelmente estão razoavelmente próximas do melhor que podemos fazer atualmente com as informações que temos em nosso poder.
- Acho que elas são boas o suficiente para os propósitos do que direi abaixo sobre cronologias da Inteligência Artificial transformadora.

Não planejo defender mais esta posição aqui, mas posso fazê-lo no futuro caso receba muitas críticas.

O Bio-âncoras como uma forma de delimitar cronologias de IA

Após reconhecer todas as fraquezas acima, aqui estão alguns pontos em que acredito sobre as cronologias de IA, os quais são amplamente baseados na análise do Bio-âncoras:

- **Eu ficaria pelo menos levemente surpreso se a Inteligência Artificial transformadora não fosse desenvolvida até 2060.** Estimo em 50% a probabilidade do desenvolvimento da Inteligência Artificial transformadora até então, (explico abaixo como "leve surpresa" e "50%" se relacionam); eu poderia simpatizar com alguém que estimasse uma probabilidade de 25% ou 75%, mas teria dificuldade em entender as razões de alguém que estimasse algo fora desse intervalo. Mais
- **Eu ficaria muito surpreso se a Inteligência Artificial transformadora não fosse desenvolvida até 2100.** Estimo a probabilidade de isso acontecer até 2100 de 2/3; eu seria solidário com alguém que dissesse que seria 1/3 ou 80–90%, mas teria dificuldade em entender as razões de alguém que estimasse algo fora desse intervalo. Mais
- **O desenvolvimento da Inteligência Artificial transformadora até 2036 parece plausível e concretamente imaginável, mas não parece uma boa expectativa padrão.** Acho que a probabilidade disso acontecer até 2036 é de pelo menos 10%; eu seria solidário com alguém que dissesse que seria 40–50%, mas teria dificuldade em entender as razões de alguém que dissesse que seria<10% ou >50%. Mais

Eu ficaria pelo menos levemente surpreso se a Inteligência Artificial transformadora não fosse desenvolvida até 2060.

Principalmente porque, segundo o Bio-âncoras, será econômico realizar algumas rodadas de treinamento *absurdamente* grandes. Sem dúvida, as maiores que alguém pudesse imaginar um dia precisar fazer, com base no uso de modelos de Inteligência Artificial com 10x o tamanho dos cérebros humanos e tarefas que requerem inúmeros cálculos para serem feitas até mesmo uma única vez.

Em algum sentido importante, disporemos de poder computacional em abundância. (Mais sobre esta intuição em ***Fun with +12 OOMs of compute* (Diversão com mais de doze ordens de magnitude de computação**.)

Mas *também* é importante que 2060 seja daqui a 40 anos, ou seja, que tenhamos 40 anos para:

- Desenvolver algoritmos de Inteligência Artificial cada vez mais eficientes, alguns dos quais seriam grandes avanços.
- Aumentar o número de empresas e negócios centrados em IA, coletando dados sobre a interação humana e concentrando cada vez mais atenção nas coisas que atualmente são obstáculo para aplicações mais amplas.

Dada a quantidade crescente de investimentos, talentos e potenciais aplicações para os sistemas de Inteligência Artificial atuais, 40 anos parece um tempo muito longo para ter grandes progressos nessas frentes. Para contextualizar, 40 anos é o tempo decorrido entre o **lançamento do Apple IIe** e agora.

Quando se trata de traduzir minha "sensação de leve surpresa" em uma probabilidade (veja **aqui** para ter uma noção do que estou tentando fazer ao falar sobre probabilidades; espero escrever mais sobre este tópico no futuro):

- Na maioria dos tópicos, eu igualo "eu ficaria um pouco surpreso se X não acontecesse" com algo como uma probabilidade de 60-65% de X. Mas, neste tópico, acredito que há um ônus da prova (o que considero importante, embora não seja excessivo), e estou inclinado a abaixar um pouco as minhas estimativas. Então, estou dizendo haver cerca de 50% de probabilidade de termos a Inteligência Artificial transformadora até 2060.

- Eu seria solidário se alguém dissesse "40 anos não parece o suficiente para mim; acredito que há mais de 25% de probabilidade de termos uma Inteligência Artificial transformadora até 2060." Mas se alguém colocasse a probabilidade em menos de 25%, eu começaria a pensar: "É mesmo? De onde você tirou isso? Em meio ao boom atual da Inteligência Artificial, é razoável se perguntar por que há menos de 25% de probabilidade de desenvolvermos a Inteligência Artificial transformadora em um ano. Considerando que, segundo nossas melhores estimativas, teremos poder computacional suficiente para realizar as maiores rodadas necessárias, e ainda teremos 40 anos até 2060 para resolver muitos dos outros obstáculos, é importante reavaliar essa probabilidade.

- Por outro lado, eu seria solidário com alguém que dissesse "Esta estimativa parece muito conservadora; 40 anos devem ser suficientes; Acho que há mais de 75% de probabilidade de termos uma Inteligência Artificial transformadora até 2060." Mas se alguém colocasse a probabilidade em mais de 75%, eu começaria a pensar: "É mesmo? De onde você tirou isso? A Inteligência Artificial transformadora não parece próxima, então me parece muita confiança em um evento daqui a 40 anos."

Eu ficaria muito surpreso se a Inteligência Artificial transformadora não fosse desenvolvida até 2100

Até 2100, o Bio-âncoras prevê que será econômico, não apenas realizar rodadas de treinamento quase comicamente grandes (novamente com base no tamanho hipotético dos modelos e custo por tentativa das tarefas), *mas para realizar tantos cálculos quanto todos já feitos por todos os animais da história juntos, a fim de recriar o progresso feito pela seleção natural.*

Além disso, 2100 é daqui a 80 anos — mais do que o tempo decorrido desde que os computadores digitais programáveis foram **desenvolvidos inicialmente**. É *muito tempo* para encontrar novas abordagens para algoritmos de IA, integrar a Inteligência Artificial na economia, coletar dados de treinamento, lidar com casos em que os atuais sistemas de Inteligência Artificial não conseguem aprender tarefas específicas, etc.

Para mim, parece que 2100 é algo como "O ponto mais longe no tempo para contar uma história aparentemente razoável, e mais um pouco". Consequentemente, eu ficaria muito surpreso se a Inteligência Artificial transformadora não fosse desenvolvida até então, e atribuo cerca de 2/3 de probabilidade de que ela será. E:

- Eu seria solidário com alguém que dissesse "Bem, há muito que não sabemos e muito que ainda precisa acontecer — só acredito que há 50% de probabilidade de termos uma Inteligência Artificial transformadora até 2100". Eu seria até *um pouco* compreensivo se eles dessem uma probabilidade de 1/3. Mas se alguém colocasse a probabilidade em menos de 1/3, eu realmente teria problemas para entender as suas razões para isso.

- Eu seria solidário se alguém colocasse a probabilidade do desenvolvimento da "IA transformadora até 2100" em mais ou menos 80–90%, mas dada a dificuldade de prever esse tipo de coisa, eu realmente teria problemas para entender as suas razões se eles passassem de 90%.

A Inteligência Artificial transformadora até 2036 parece plausível e concretamente imaginável, mas não parece uma boa expectativa padrão.

O Bio-âncoras apresenta cenários concretos e plausíveis nos quais há suficiente poder computacional econômico para treinar a Inteligência Artificial transformadora até 2036 (link). Conheço alguns pesquisadores de Inteligência Artificial que acham que esses cenários são mais do que plausíveis — suas intuições dizem a eles que as rodadas gigantes de treinamento imaginadas pelo Bio-âncoras são desnecessárias e que as âncoras mais agressivas no relatório estão sendo subestimadas.

Também acredito que o Bio-âncoras subestima um pouco o caso para a "IA transformadora até 2036", porque é difícil dizer quais consequências o atual boom de investimento e interesse em Inteligência Artificial terá. Se a Inteligência Artificial está prestes a se tornar uma parte visivelmente maior da economia (definitivamente um "se", mas compatível com as **tendências recentes do mercado**), isso resultaria em rápidas melhorias em muitas dimensões possíveis.

Uma relação de retroalimentação positiva poderia existir na qual novos aplicativos lucrativos de Inteligência Artificial estimulassem maiores investimentos em IA. Isso, por sua vez, poderia levar a melhorias mais rápidas do que o esperado na eficiência dos algoritmos e da computação de IA, o que, em última análise, resultaria em aplicativos mais lucrativos.…

Considerando todos esses fatores, acredito **que a probabilidade do desenvolvimento da Inteligência Artificial transformadora até 2036 é de pelo menos 10%**, e não tenho muita simpatia por alguém que diga que ela é menor.

E dito isto, tudo o que foi dito acima é um conjunto de "poderia ser" e "poderia acontecer" – todos os casos que ouvi de "IA transformadora até 2036" parecem exigir uma série de etapas incertas a serem resolvidas.

- Se tarefas de "longo-prazo" se tornarem importantes, o Bio-âncoras mostra que é difícil imaginar que haverá computação suficiente para as rodadas de treinamento necessárias.

- Mesmo que haja muito poder computacional disponível, 15 anos podem não ser tempo suficiente para resolver desafios como reunir os dados e ambientes de treinamento corretos.

- É certamente possível que algum paradigma completamente diferente surja - talvez inspirado pela neurociência - e a Inteligência Artificial transformadora seja desenvolvida de maneiras que não requeiram «rodadas de treinamento» do tipo Bio-âncoras. Mas não vejo nenhuma razão em particular para esperar que isso aconteça nos próximos 15 anos.

Portanto, também não simpatizo muito com as pessoas que pensam que há mais de 50% de probabilidade de uma Inteligência Artificial transformadora ser desenvolvida até 2036.

Conclusão

O método das Bio-âncoras é um pouco diferente da abordagem "comum" para estimar as coisas. Ele não "aponta" datas prováveis para a Inteligência Artificial transformadora; ele não modela todos os fatores-principais.

Mas acredito que ele é muito útil — em conjunto com o raciocínio informal sobre os fatores que ele não modela — para "delimitar" cronologias da Inteligência Artificial transformadora: fazer uma variedade de declarações do tipo "Seria surpreendente se a Inteligência Artificial transformadora não fosse desenvolvida até a data X" ou "Você poderia defender uma porcentagem de probabilidade X até tal data, mas acredito que seria difícil simpatizar com uma porcentagem de probabilidade Y".

E, esse tipo de "delimitação" parece bastante útil para o propósito que mais me interessa: decidir com que seriedade levar a possibilidade do século mais importante. Minha opinião é que essa possibilidade é muito séria, embora longe de ser uma certeza, e o Bio-âncoras é uma parte importante desse panorama para mim.

Mais sobre "Múltiplas economias de tamanho mundial por átomo":

Uma continuação de "Isto não pode continuar" para os céticos.

Em Isto não pode continuar, eu **argumentei** que mais 8.200 anos da taxa de crescimento de hoje exigiriam que sustentássemos "múltiplas economias tão grandes quanto a economia mundial inteira de hoje *por átomo*".

O feedback sobre este trecho foi dividido entre "Isso é obviamente impossível, 8200 anos de crescimento de 2% é uma ideia absurda — o crescimento terá que diminuir muito antes disso" e "Por que isso seria impossível? Com criatividade crescente, poderíamos aumentar cada vez mais a qualidade de vida, sem a necessidade de usar cada vez mais recursos materiais."

Aqui responderei ao último ponto, o que significa explicar porque 8.200 anos de crescimento de 2% não parece uma coisa razoável de se esperar. Farei muitas suposições extremamente audaciosas e falarei sobre todos os tipos de possibilidades estranhas, apenas para poder abordar até mesmo maneiras rebuscadas para que o crescimento de 2% continue.

Se você já está no time dos que dizem "Sim, não vejo a economia mundial crescendo tanto assim", você deve pular esta postagem, a menos que goste de ver o caso construído com uma quantidade razoável de detalhes.

Como PODERÍAMOS sustentar "múltiplas economias de tamanho mundial por átomo"

Eu realmente acho que seja concebível que poderíamos sustentar "múltiplas economias de tamanho mundial por átomo". Aqui está uma forma de fazer isso:

Digamos que descobrimos alguma nova atividade, ou experiência, ou droga, que as pessoas realmente, realmente, REALMENTE valorizam.

Especificamente, o mercado avalia essa descoberta em 10^{85} dos dólares americanos de hoje (isso é dez trilhões de trilhões de trilhões de trilhões de trilhões de trilhões de dólares). Isso significa que ela vale cerca de, 10^{71} vezes mais do que tudo o que o mundo produz em um ano atualmente (junto).[130]

Então, uma pessoa tendo essa experiência[131] significaria que o tamanho da economia é de pelo menos US$ 10^{85}. E isso seria, de fato, o equivalente a múltiplos das economias mundiais atuais por átomo.[132]

Sendo claro, não é que teríamos amontoado várias economias mundiais atuais em cada átomo. É que teríamos amontoado algo 10^{71} vezes mais valioso que a economia mundial atual em meros 10^{28} átomos que compõem um ser humano.

O que significaria, porém, valorizar uma única experiência, 10^{71} vezes mais do que toda a economia mundial atual?

Uma maneira de pensar sobre isso pode ser:

- "Uma probabilidade de 1 em 10^{71} dessa coisa ser vivenciada ser tão valiosa

quanto toda a economia mundial atual."

- Ou, para tornar isso um pouco mais fácil de intuir (embora precise simplificar demais), "Se eu fosse neutro em relação ao risco, ficaria emocionado em aceitar uma aposta em que morreria imediatamente, com quase certeza, em troca de 1 em 10^{71} chance de ter essa experiência."[133]

- O quanto a morte seria quase certa? Bem, para começar, se todas as pessoas que já viveram até hoje aceitassem essa aposta, seria quase certo que todas perderiam e acabariam mortas imediatamente.[143]

- Isso realmente não chega nem perto de transmitir o quanto as chances de ganhar essa aposta seriam ruins. É mais como: se houvesse uma pessoa para cada átomo na galáxia, e cada um deles fizesse a aposta, provavelmente **todos** perderiam.[135]

- Então, para fazer uma aposta pessoal com esse tipo de probabilidade… é melhor que a experiência seja REALMENTE boa para compensar.

- Não estamos falando de uma experiência do nível "a melhor experiência que você já teve" aqui — não seria sensato valorizar isso mais do que uma vida inteira, e a ideia de que vale tanto quanto a economia mundial atual parece claramente errada.

- Estamos falando de algo insondavelmente além de qualquer coisa que qualquer ser humano já experimentou.

Aumentando os números mais ainda

Imagine o melhor segundo da sua vida, o tipo de coisa evocada por ***Letter from Utopia* (Carta da Utopia):**

Você já experimentou um momento de felicidade? Nas corredeiras da inspiração, talvez, sua mente traçando as formas da verdade e da beleza? Ou talvez no êxtase pulsante do amor? Ou em um triunfo glorioso alcançado com amigos verdadeiros? Ou em uma conversa em um terraço coberto de videiras em uma noite estrelada? Ou talvez uma melodia se tenha infiltrado em seu coração, encantando-o e incendiando-o com emoções caleidoscópicas? Ou quando você rezou e se sentiu ouvido?

Se você experimentou tal momento – experimentou *o melhor tipo* de tal momento – então você pode ter descoberto dentro dele um certo pensamento ocioso, mas sincero: "Oh céus, sim! Eu não sabia que poderia ser assim. Isso está tão certo, em um nível totalmente diferente de certo; tão real, em um nível totalmente diferente de real. Por que não pode ser assim sempre? Antes eu estava dormindo; agora estou acordado."

No entanto, um pouco mais tarde, apenas uma hora se passou, e a fuligem em queda permanente da vida cotidiana já está cobrindo tudo. A prata e o ouro da exuberância perdem o brilho, e o mármore fica sujo.

Agora imagine, implausivelmente, que este único segundo valesse tanto quanto toda a produção da economia mundial atual em um ano. (Não parece possível que valesse mais, já que a economia mundial naquele ano *incluía* aquele segundo da sua vida, mais o resto do seu ano e os anos de muitas outras pessoas.)

E agora imagine um *ano inteiro* no qual *cada segundo* é tão bom quanto *aquele segundo.* Chamaremos isso de "ano perfeito". Segundo as suposições acima, o ano perfeito não seria mais do que cerca de $3X10^8$ vezes mais valioso que a economia mundial (existem cerca de $3X10^8$ segundos em um ano).

E agora imagine que *cada átomo da galáxia* pode ser uma pessoa tendo o ano perfeito. Isso seria agora cerca de 10^{70} X (3 X 10^8) = $3X10^{78}$ tanto valor quanto a economia mundial atual. **Um crescimento de 2% nos levaria até lá em 9150 anos.**

(Uma suposição crucial e talvez contraintuitiva que estou fazendo aqui é que "2% de crescimento" significa "2% de crescimento *verdadeiramente real*" - que o que quer que seja valioso, falando de forma holística, sobre a produção mundial anual atual, ganharemos 2% a mais disso a cada ano. Acho que esse já é o tipo de suposição que muitas pessoas estão fazendo quando dizem que não precisamos de mais material para ter uma riqueza crescente. Se você acha que o crescimento de 2% do passado recente é mais "falso" do que isso e que continuará de forma "falsa", isso seria um debate para outra hora). Esse parêntese desse parágrafo é mesmo necessário?

E 1200 anos depois *disso*, se cada ano ainda tivesse 2% de crescimento, a economia seria aproximadamente 20 bilhões de vezes maior. Então agora, para cada átomo na galáxia, deve haver alguém cujo ano seja, em certo sentido, 20 bilhões de vezes *melhor*. Ou "mais valioso" do que o ano perfeito.

Ainda estamos falando de aproximadamente 10.000 anos de crescimento de 2%.

Novas formas de vida

Isso ainda é concebível! Quem sabe o que o futuro trará?

Mas, neste ponto, é muito intuitivo para mim que não estamos falando sobre nada que se pareça com "Humanos em corpos humanos tendo tipos humanos de diversão e satisfação". Uma economia desse valor parece exigir fundamentalmente a reengenharia de algo sobre a experiência humana - encontrar alguma maneira de organizar a matéria que crie muito mais felicidade, ou realização, ou algo que valorizaríamos tão astronomicamente mais do que até mesmo as alturas da experiência humana hoje.

E acredito que a maneira mais natural de isso acontecer seria algo como: "Descobrir princípios fundamentais por trás do que valorizamos e princípios fundamentais de como organizar a matéria para tirar o máximo dela." O que, por sua vez, sugere algo mais como "Uma vez que tivéssemos esse nível de compreensão, começaríamos a organizar a matéria na galáxia de maneira otimizada e rapidamente nos aproximaríamos dos limites do que é possível" do que algo semelhante a "Cresceríamos 2%, todos os anos, por milhares de anos contínuos, mesmo quando (como aconteceria com, por exemplo, pessoas digitais) nos tornássemos seres que conseguissem fazer tanto em um ano quanto os humanos fizessem em centenas ou milhares de anos."

Mas isso ainda pode acontecer?

Acho que sim? Isso nunca foi uma prova matemática da impossibilidade de crescimento de 2% ao ano. É possível em teoria.

Mas, neste ponto, vendo que galáxia excêntrica e fundamentalmente transformada seria necessária dentro de 10.000 anos, qual é a razão *afirmativa para* esperar um crescimento de 2% ao ano por um longo período? É que "Esta é a linha de tendência e, por padrão, espero que a linha de tendência continue?"

Mas essa linha de tendência tem apenas algumas centenas de anos — por que esperar que continue por mais 10.000?

Por que não, em vez disso, esperar que o padrão de longo prazo de aceleração do crescimento econômico continue, até que nos aproximemos de algum tipo de limite fundamental sobre quanto valor podemos atribuir a uma determinada quantidade de matéria? Ou, esperar que o crescimento diminua gradualmente a partir daqui e nunca mais atinja o nível de hoje?

Os últimos dois séculos foram uma grande aventura, com a riqueza e as condições de vida melhorando a uma taxa historicamente alta. Mas não acredito que isso nos dê motivos para pensar que essa tendência continuará até o infinito. Acredito que os limites estão em algum lugar, e parece que em algum momento nos próximos 10.000 anos, teremos que nos aproximar desses limites, ou estagnar, ou crescer, ou entrar em colapso.

Espero ter dado uma ideia do porquê parece tão improvável que haverá mais 10.000 anos no futuro, com 2% ou mais de crescimento a cada ano. O que sugere que *cada* um dos últimos 100 anos ou mais será um dos 10.000 anos de crescimento mais rápido de todos os tempos.

Se quiser comentar esta postagem, **este** seria um bom lugar para fazê-lo.

Uma nota sobre o crescimento econômico histórico

Como o argumento do "século mais importante" é afetado se nossa imagem da história

econômica de longo prazo mudar.

Algumas vezes na série o século mais importante (particularmente em O Duplicador), digo que o crescimento econômico nos últimos milhares de anos se ajusta razoavelmente ao padrão (descrito aqui) de crescimento acelerado, impulsionado por uma retroalimentação: "mais ideias → mais resultados → mais pessoas → mais ideias→ …"

Este argumento é o assunto de um debate em curso (veja **esta postagem no Fórum do Altruísmo Eficaz feita por Ben Garfinkel**, e as extensas trocas de mensagens nos comentários).

Meu melhor palpite é que os dados anteriores são, de fato, um ajuste razoável (embora ambíguo) ao padrão de crescimento acelerado. No entanto, estou longe de confiar nisso e quero abordar como isso afetaria meus argumentos se, dados futuros melhores acabassem por minar decisivamente esse ajuste.

Extrapolando o crescimento econômico futuro com base em (uma visão de longo prazo) do crescimento econômico passado

Citei a projeção, feita em ***Modeling the Human Trajectory*** **(Modelando a trajetória humana)**, que a economia está "no caminho" de atingir tamanho infinito neste século se o padrão observado no passado continuar. Se os dados anteriores acabassem sendo inconsistentes com o crescimento acelerado, isso prejudicaria o Modeling the Human Trajectory (Modelando a trajetória humana), e uma nova extrapolação seria necessária. **No entanto, meu melhor palpite é que uma boa extrapolação de substituição ainda mostraria uma boa chance de crescimento explosivo (até "infinito") neste século.** Segue o raciocínio para essa suposição.

Ao discutir o padrão de crescimento passado, a principal alternativa que vi para acelerar o crescimento (inclusive na postagem do Fórum EA vinculada acima e nos comentários) *é uma série de "modos de crescimento" fundamentalmente diferentes, cada um com sua própria dinâmica de crescimento e/ou taxa de crescimento.* Por exemplo, talvez — em vez de pensar na história econômica como uma aceleração gradual — pode-se pensar nela como faseada:

- Uma fase pré-agricultura (com início há alguns milhões de anos), na qual o crescimento provavelmente foi extremamente lento e talvez bastante caótico.
- Uma fase após o desenvolvimento da agricultura (com início há cerca de 10.000 anos), durante a qual o crescimento foi provavelmente mais rápido do que antes, mas ainda bastante lento para os padrões atuais e talvez bastante caótico também.
- A fase moderna pós-Revolução Industrial (começando há aproximadamente 200 anos), com, de longe, o crescimento mais rápido.

Parece indiscutível para mim que a terceira fase é muito mais curta (no tempo histórico) e tem um crescimento dramaticamente mais rápido, em comparação com as duas primeiras fases. Isso poderia ser o resultado de uma aceleração contínua ou o surgimento de um modo de crescimento fundamentalmente novo. Este último levantaria, então, a questão de saber se uma transição para outro "modo de crescimento" ainda mais rápido seria possível.

O artigo de 2000 de Robin Hanson, ***Long-Term Growth As A Sequence of Exponential Modes* (Crescimento de Longo Prazo como uma Sequência de Modos Exponenciais)**, é a principal tentativa que conheço de explorar essa questão. Ele tenta modelar a história econômica de longo prazo usando algumas abordagens diferentes, ambas projetadas em torno da ideia de "modos de crescimento" e (nas páginas 14–17) para extrapolar os padrões observados até o momento no futuro. Ele afirma que:

Em resumo, se alguém levar a sério o modelo de crescimento econômico como uma série de modos de crescimento exponencial, e se os parâmetros de mudança relativa de uma nova transição provavelmente forem semelhantes aos parâmetros que descrevem as transições anteriores, seria difícil escapar da conclusão de que a economia mundial poderá passar por uma mudança muito dramática no próximo século, para um novo modo de crescimento econômico com um tempo de duplicação de aproximadamente duas semanas ou menos...

Se o próximo modo tivesse um tempo de duplicação "lento" de dois anos, e se durasse vinte vezes de duplicação, mais do que qualquer modo visto até agora, ainda assim ele duraria apenas quarenta anos. Depois disso, não está claro quantos modos de crescimento ainda mais rápidos são possíveis antes de atingir os limites fundamentais. Mas é difícil entender como tais limites fundamentais não seriam alcançados dentro de algumas décadas, no máximo.

Isso é qualitativamente muito semelhante à projeção que dei nas postagens do blog: ambos implicam uma aceleração econômica dramática no século XXI e ambos implicam "crescimento infinito" ou "atingir limites fundamentais" não muito tempo depois (embora o atraso potencial seja mais longo na abordagem de Hanson e poderia modestamente chegar até o século XXII, dependendo de qual projeção de Hanson se usa).

Essa extrapolação é menos direta do que a extrapolação feita em ***Modeling the Human Trajectory* (Modelando a trajetória humana)**. Não há razões muito fortes para pensar que a série de modos de crescimento seguirá um padrão específico, em termos de como eles são cronometrados e que tipo de crescimento eles trarão. A extrapolação de Hanson é apenas uma estimativa sobre o que esperar se eles seguirem um padrão relativamente regular. Ainda assim, parece razoável para mim como uma melhor estimativa.

Outras implicações se os dados econômicos anteriores não se encaixarem em um padrão de "crescimento acelerado"

- Ao longo da série, defendo que **várias tecnologias** (O duplicador , pessoas digitais, o "PASTA"[136]) **levariam a um padrão de "aceleração" que levaria a um crescimento explosivo.** Esta é uma implicação da maioria dos modelos teóricos dominantes na economia do crescimento, conforme discutido em ***Report on Whether AI Could Drive Explosive Economic Growth* (Relatório sobre se a Inteligência Artificial pode impulsionar o crescimento econômico explosivo)**.[137] Cito que os dados anteriores parecem se encaixar nessa dinâmica como evidência adicional de que tal coisa é plausível. Se os dados anteriores não se ajustassem à dinâmica, isso não afetaria este caso teórico para a expectativa de um crescimento explosivo, mas tornaria a solidez geral do argumento um pouco mais fraca.

- Também argumentarei contra a ideia de que "se a Inteligência Artificial transformadora fosse desenvolvida neste século, isso quebraria o padrão de crescimento econômico constante que observamos; portanto, devemos ter um ônus de prova muito alto para previsões de Inteligência Artificial transformadora neste século". Para esse propósito, a dinâmica de "crescimento acelerado" ou "série de diferentes modos de crescimento" parece suficiente para o meu caso de que devemos considerar plausível uma futura explosão de crescimento, embora eu ache que o caso seria um pouco mais fraco se tivesse que confiar neste último em oposição ao primeiro.

Conclusão

No geral, se ficar claro que a história econômica contém muito pouca aceleração (e, em vez disso, é mais bem pensada como uma série de "modos de crescimento" distintos), acredito que minhas afirmações e conclusões restantes ainda pareceriam corretas, embora os argumentos fossem um pouco mais fracos.

Também é possível que, se tivéssemos informações perfeitas sobre a história econômica de longo prazo, veríamos uma mistura: *algumas* instâncias/períodos da dinâmica de "crescimento acelerado" descrita aqui, alguns períodos que se parecem mais com "modos distintos de crescimento".

Alguns detalhes adicionais sobre o que quero dizer com "século mais importante"

Aqui está um pouco mais de detalhes sobre o que quero dizer quando falo sobre o "século mais importante para a humanidade".

Existem dois sentidos diferentes em que acredito que este pode ser o "século mais importante", um deles de maior risco e menos provável do que o outro:

Significado nº 1: O século mais importante de todos os tempos para a humanidade, devido à transição para um estado no qual os humanos, tais como os conhecemos, não serão mais a força principal nos eventos mundiais.

Aqui, a ideia é que:

- Durante este século, a civilização poderia terminar completamente ou mudar tão dramaticamente que "os humanos tais como os conhecemos hoje" não existiriam mais ou seriam pelo menos uma parte muito pequena da população.
- Acho que o futuro que descrevo em Pessoas digitais seriam mais importantes ainda provavelmente atingiria esse nível de estranheza muito rapidamente.
- A possibilidade dos sistemas de Inteligência Artificial se expandirem pela galáxia com base em seus próprios objetivos - com os humanos se tornando bastante irrelevantes em comparação - também se qualificaria.
- Este século é a nossa chance de moldar como isso acontece.
- Se desenvolvermos pessoas digitais, o conjunto inicial de pessoas digitais poderia rapidamente começar a fazer muitas cópias de si mesmas, multiplicando-se e trabalhando a uma taxa muito mais rápida do que os humanos normais conseguiriam rastrear ou acompanhar. Considerando esses argumentos, o conjunto inicial de pessoas digitais — e as condições virtuais em que são colocadas — seriam cruciais de forma duradoura.
- Se, em vez disso, desenvolvêssemos sistemas de Inteligência Artificial que se expandiriam pela galáxia com base em seus próprios objetivos, isso faria com que perdêssemos permanentemente a oportunidade de fazer com que a força principal nos eventos mundiais seja algo parecido com os humanos.

Com base nesses argumentos, este seria o "século mais importante" para os humanos tais como somos agora, no sentido de que essa será a melhor oportunidade que os humanos terão de influenciar um grande futuro pós-humanos-tais-como-são-agora.[138]

Isso pode ser consistente com outros séculos sendo "mais importantes" para outras "espécies".

- Algum século passado pode ter sido o século mais importante para os chimpanzés. (Isso pode ter ocorrido em algum século durante o qual os humanos começaram a surgir.)
- Algum século futuro pode ser o século mais importante para o que quer que "venha depois dos humanos". (Embora este século seja o mais importante para eles também.)

Quero dizer, aproximadamente, que *se* algo como o PASTA for desenvolvido neste século, ele tem pelo menos 50/50 de probabilidade de ser o "século mais importante" no sentido dito acima.

Significado nº 2: o século mais importante de todos os tempos para toda a vida inteligente em nossa galáxia.

É possível, pelas razões descritas aqui, que independentemente da força principal nos eventos mundiais (talvez pessoas digitais, Inteligência Artificial desalinhada ou qualquer outra coisa), ela criará civilizações altamente estáveis com valores "aprisionados", que povoarão toda a nossa galáxia por bilhões de anos por vir.

Se o suficiente desse "aprisionamento/lock-in" acontecer neste século, isso poderá torná-lo o século mais importante de todos os tempos para toda a vida inteligente em nossa galáxia.

Quero dizer, aproximadamente, que *se* algo como o PASTA for desenvolvido neste século, ele tem pelo menos 25% de chance de ser o "século mais importante" no sentido dito acima. Isso é metade da probabilidade da versão anterior do "século mais importante". Não pretendo ser preciso aqui. Estou dando uma indicação aproximada de quão provável seria tal desenvolvimento.

Para se ter uma melhor ideia disso, vale notar que o mundo parece ter "acelerado" – no sentido de mudar mais rapidamente – ao longo da história, e poderia continuar assim se algo como o PASTA for desenvolvido. Considerando essas afirmações:

- Se as primeiras bactérias estivessem conversando entre si, uma delas poderia ter afirmado que elas estavam no "período mais importante de 5 bilhões de anos," aquele no qual as bactérias evoluiriam para animais complexos.
- Os primeiros animais complexos poderiam ter afirmado estarem no "**éon**," mais importante"(???), aquele no qual os humanos emergiriam.
- Alguém vivendo a **Revolução Científica** poderia afirmar que estava no "milênio mais importante", aquele no qual o progresso científico e tecnológico decolaria.
- Se a Inteligência Artificial transformadora levar a uma civilização gerida por pessoas digitais por volta de, digamos, 2080, alguma pessoa digital em 2080 poderá alegar que está na "década mais importante". Uma década pode parecer para eles o mesmo que um século (ou mais) se parece para nós.
- Essas pessoas digitais poderiam criar pessoas digitais mais avançadas que afirmariam estar no "dia mais importante", imaginando que evoluirão para algo ainda mais estranho durante esse vasto intervalo de tempo.
- E todos eles poderiam estar certos!

Intenção holística da frase "século mais importante". Em grande parte, escolhi a frase "século mais importante" **como um chamado de atenção** sobre o quão elevados são os riscos que podemos enfrentar.

Embora eu tenha tentado dar um significado um pouco mais preciso acima, minha principal intenção é chamar a atenção para o sentimento **de perplexidade diante da possibilidade de desenvolver algo como o** PASTA ***neste século, que por sua vez levaria a um futuro radicalmente desconhecido, possivelmente envolvendo uma civilização estável em toda a galáxia.***

Se eu estiver certo sobre esse cenário, mas errado sobre o "século mais importante" por algum motivo (por exemplo, talvez algo ainda mais notável aconteça daqui a 5 bilhões de anos, ou talvez aconteça que a hipótese da simulação esteja correta), eu ainda acredito que a ideia geral desta série está correta.

Por que falar agora sobre daqui a 10.000 anos?

Parece que uma reação comum a Isto não pode continuar é algo como: «Ok, então você está dizendo que o atual nível de crescimento econômico não pode continuar por mais 10.000 anos… Então me ligue daqui a alguns milhares de anos, talvez."

Em geral, este blog costuma falar sobre períodos "longos" (décadas, séculos, milênios) como se fossem "curtos" (em comparação com os bilhões de anos que nosso universo existe, milhões de anos que nossa espécie existe e bilhões de anos que poderiam estar no futuro de nossa civilização). Eu meio que tento me **imaginar como um observador de bilhões de anos,** olhando para gráficos como este e pensando coisas como "O atual nível de crescimento econômico acabou de começar!" mesmo que tenha começado várias vidas atrás.

Por que pensar desta forma?

Uma razão é que é apenas uma maneira de pensar sobre o mundo que parece (para mim) revigorante/diferente.

Mas aqui estão mais algumas razões importantes.

Altruísmo eficaz

Minha principal obsessão é com o **altruísmo eficaz**, ou fazer o máximo de bem possível. Geralmente tento prestar mais atenção às coisas quando elas "importam mais" e acredito que as coisas "importam mais" quando afetam um número maior de pessoas.[139]

Acho que haverá MUITO mais pessoas[140] nos próximos bilhões de anos do que nas próximas gerações ou poucas. Portanto, acredito que o futuro de longo prazo, em certo sentido, "importa mais" do que o que quer que aconteça na próxima geração ou em algumas poucas gerações. Talvez não importe mais para mim e meus entes queridos, mas importa mais do ponto de vista de "todos os seres são igualmente importantes".[141]

Uma resposta óbvia é "Mas não há nada que possamos fazer que afete TODAS as pessoas que viverão nos próximos bilhões de anos. Devemos nos concentrar no que realmente podemos mudar - isso é na próxima geração ou poucas mais adiante."

Mas não estou convencido disso.

Acho que poderíamos estar no século mais importante de todos os tempos, e acredito que as coisas que fazemos hoje podem acabar importando por bilhões de anos (um exemplo óbvio é **reduzir o risco de catástrofes existenciais**).

E, de maneira mais ampla, se eu não *conseguisse* pensar em maneiras específicas pelas quais nossas ações poderiam ser importantes por bilhões de anos, ainda assim eu estaria muito interessado em *descobrir essas maneiras.* Eu ainda acharia útil tentar dar um passo para trás e perguntar: "O que estou lendo nas notícias é importante no final das contas? Esses eventos importariam se no futuro terminarmos com uma economia em explosão, estagnação ou colapso? Para que tipo de civilização digital serão nossas criações de longo prazo? E se não pudermos criar… o que poderia?"

Apreciando a estranheza do tempo em que vivemos

Acho que vivemos em uma época muito estranha. Ela parece muito estranha em vários gráficos (como este, este, e este). A maioria do avanço científico e tecnológico e do crescimento da economia aconteceu em uma pequena fração de tempo em que estamos vivendo. E daqui a bilhões de anos, provavelmente ainda será o caso de que esta pequena fração de tempo seja um ponto fora da curva em termos de crescimento e mudança.

Mais uma vez, não *parece* uma pequena fração de tempo, parece uma vida inteira São centenas de anos. Mas isso é dentre milhões (para nossa espécie) ou bilhões (para a vida na Terra).

Às vezes, quando ando na rua, olho em volta e penso: "Isso tudo é TÃO ESTRANHO. Ao meu redor, um grupo de pessoas dirige carros de aço a mais de 60 km/h, com tranquilidade. Vejo também um guindaste gigante a construir um arranha-céu, operado calmamente por um grupo de pessoas. No céu, um avião voa. Após bilhões de anos de vida na Terra, somos nós — os humanos dos últimos cem ou mais anos — os únicos capazes de realizar tais feitos. Praticamente tudo o que vejo é uma tecnologia futurista maluca que acabamos de inventar e para a qual não tivemos tempo de nos adaptar, e não teremos nos adaptado antes que a próxima coisa maluca apareça.

"E todo mundo está sendo muito monótono em seus carros, arranha-céus e aviões, mas isso *não* é normal, isso *não é* como geralmente é,' isso não faz parte de um plano ou um padrão bem estabelecido, isso é louco, estranho e de curta duração, e ninguém sabe para onde isso vai a seguir."

Acho que muitos de nós somos instintiva e intuitivamente desdenhosos de afirmações audaciosas sobre o futuro. Acho que naturalmente imaginamos que exista mais estabilidade, solidez e sabedoria oculta em "como as coisas têm sido por gerações" do que realmente existe.

Ao tentar **imaginar a perspectiva de alguém que esteve vivo durante toda a história —** bilhões de anos, não dezenas — talvez possamos estar mais abertos a estranhas possibilidades futuras. E então, talvez possamos perceber melhor as que realmente podem acontecer e que podem ser afetadas pelas nossas ações de hoje.

É por isso que costumo tentar dizer coisas como «X já está acontecendo há 200 anos e talvez possa durar mais alguns milhares de anos — ah, isso é um piscar de olhos!»

Notas de fim

1 Começaremos com a economia:

Ano	nº. de pessoas	Recursos produzidos por pessoa (como milho no diagrama)	Recursos produzidos (total)	Novas réplicas adicionadas à população (1 para cada 10 unidades de recursos)	n.º de novas ideias (inicialmente 1 para cada 20 pessoas, mas ideias são difíceis de ter)	Melhora da produtividade (1% por ideia - aumento de recursos produzidos por pessoa)
1	100	1.00	100	10	5	5.00%

100 pessoas produzem 100 unidades de recursos (1 por pessoa). Para cada 10 unidades de recursos, elas podem criar mais 1 duplicata (isso é apenas para capturar a ideia de que duplicatas são "caras" de criar). E essas 100 pessoas têm 5 novas ideias, levando a um crescimento de produtividade de 5%.

Aqui está o ano 2:

Ano	nº. de pessoas	Recursos produzidos por pessoa (como milho no diagrama)	Recursos produzidos (total)	Novas réplicas adicionadas à população (1 para cada 10 unidades de recursos)	n.º de novas ideias (inicialmente 1 para cada 20 pessoas, mas ideias são difíceis de ter)	Melhora da produtividade (1% por ideia - aumento de recursos produzidos por pessoa)
1	100	1.00	100	10	5	5.00%
2	110	1.05	116	12	5	5.24%

Agora cada pessoa produz 1,05 widgets em vez de 1, graças ao crescimento da produtividade. E a produtividade cresce mais 5%.

Essa dinâmica leva algum tempo para "decolar", mas decola:

Ano	nº. de pessoas	Recursos produzidos por pessoa (como milho no diagrama)	Recursos produzidos (total)	Novas réplicas adicionadas à população (1 para cada 10 unidades de recursos)	n.º de novas ideias (inicialmente 1 para cada 20 pessoas, mas ideias são difíceis de ter)	Melhora da produtividade (1% por ideia - aumento de recursos produzidos por pessoa)
1	100	1.00	100	10	5	5.00%
2	110	1.05	116	12	5	5.24%
3	122	1.11	134	13	6	5.50%
4	135	1.17	157	16	6	5.79%
5	151	1.23	186	19	6	6.11%
6	169	1.31	222	22	6	6.47%
7	191	1.39	267	27	7	6.87%
8	218	1.49	325	32	7	7.32%
9	251	1.60	401	40	8	7.84%
10	291	1.72	501	50	8	8.43%
11	341	1.87	637	64	9	9.12%
12	404	2.04	825	82	10	9.92%
13	487	2.24	1091	109	11	10.86%
14	596	2.48	1481	148	12	11.99%
15	744	2.76	2071	207	13	13.37%
16	961	3.15	3001	300	15	15.08%
17	1251	3.63	4543	454	17	17.23%
18	1706	4.26	7269	726	20	20.04%
19	2432	5.11	12422	1242	24	23.80%
20	3674	6.32	23236	2324	29	29.04%
21	5997	8.16	48947	4895	37	36.74%
22	10892	11.16	121568	12156	49	48.80%
23	23048	16.61	382737	38274	69	69.40%
24	61322	28.13	1724982	172498	109	109.60%
25	233820	58.79	13746471	1374647	199	198.86%
26	1608467	175.70	282608848	28260885	458	457.73%
27	29869352	979.93	29269998551	2926999855	1524	1524.65%
28	2956869207	15,914.61	47057420260847	4705742026085	9290	9289.79%
29	47066980952	1,494,349.21	70364404901460	7036440490146061	157560	157560.15%
30	70364875771	2,355,843,796.	16576865610267	1657686561026930	14934113	14934113.18%
31	16576866680	351,826,734,7(	58321845239871	5832184523987101	23658280320	23658280320.49%
32	58321845239	82,864,328,754	48339669944559	4833966994455971	35182673320712600	35182673320712600.69%
33	48339669944	2,916,092,261,(	14096293747277(	1409629374727701	82884328754996400	82884328754996400000000
34	14096293747	24,169,834,972	34070509359272(	3407050935927281	29160922619937200	29160922619937200000000
35	34070509359	7,048,146,873,	24013395403236	2401339540323629	24169834972296500	24169834972296500000000
36	24013395402	17,035,254,675	#NUM!	#NUM!	70481468736384900	70481468736384900000000
37	#NUM!	12,006,697,791	#NUM!	#NUM!	#NUM!	#NUM!
38	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!

Os #NUM! na parte inferior significam que o Google Planilhas engasga com os números grandes.

Minha **planilha inclui** uma versão com população simplesmente crescente exponencialmente; esse aqui continua por aproximadamente 1000 anos sem desafiar o Google Planilhas. Portanto, a dinâmica populacional é fundamental aqui.

Sem corpos humanos – e dependendo de quais tipos de robôs estiverem disponíveis

- as pessoas digitais poderiam não ser bons substitutos para os humanos quando se trata de trabalhos que dependem fortemente de habilidades físicas humanas ou trabalhos que exigem interação pessoal com humanos biológicos.

No entanto, as pessoas digitais provavelmente conseguiriam fazer tudo o que fosse necessário para causar um crescimento econômico explosivo, mesmo que não conseguissem fazer *tudo*. Em particular, parece que elas conseguiriam fazer tudo o que fosse necessário para aumentar a oferta de computadores e, assim, aumentar a população de pessoas digitais.

Criar mais poder computacional requer (a) matérias-primas - principalmente metal;

1. pesquisa e desenvolvimento - para projetar os computadores; (c) manufatura - executar o projeto e transformar matérias-primas em computadores; (e) energia. Pessoas digitais tornariam todas essas coisas muito mais baratas e abundantes:

- **Matérias-primas.** A mineração poderia, em princípio, ser feita inteiramente com robôs. Pessoas digitais projetariam e instruiriam esses robôs para extrair matérias-primas da maneira mais eficiente possível.

- **Pesquisa e desenvolvimento.** Minha impressão é que este é um fator importante no custo da computação hoje: o trabalho necessário para projetar microprocessadores cada vez melhores e outras peças de computador. Pessoas digitais fariam isso totalmente virtualmente.

- **Manufatura.** Minha impressão é que esta é outro fator importante no custo da computação hoje. Como a mineração, isso poderia, em princípio, ser feito inteiramente com robôs.

- **Energia.** Os painéis solares também estão sujeitos a (a) melhor pesquisa e desenvolvimento; (b) manufatura acionada por robôs. Um bom design e fabricação de painéis solares acarretariam uma energia radicalmente mais barata e abundante.

Exploração espacial. Matérias-primas, energia e “locais para habitar” são superabundantes fora da Terra. Se as pessoas digitais projetassem e fabricassem naves espaciais, juntamente com robôs que construíssem painéis solares e fábricas de computadores, elas poderiam tirar proveito de recursos massivos em comparação com o que temos na Terra.

Notas

[130]A economia de hoje vale um pouco menos de US$ 10^{14} por ano (**fonte**). $\$10^{85} = \$10^{14} \times 10^{71}$.

[131](E pagando o preço total por isso, de uma forma que fica registrada pelas estatísticas do PIB, que podem ficar um pouco complicadas.)

[132]Veja uma estimativa anterior de 10^{70} átomos na galáxia.

[133]Isso pressupõe que a pessoa valorize a própria vida não muito mais do que o valor de um ano da produção da economia mundial. Não espero ver discordância suficiente sobre este ponto para querer escrever outra postagem sobre o assunto, mas é possível. Também está supondo duvidosa sobre "neutralidade de risco". Na realidade, alguém pode valorizar pessoalmente essa experiência muito menos do que, 10^{71} vezes mais do que sua própria vida, enquanto ainda paga recursos suficientes para salvar um número extraordinariamente grande de vidas de outras pessoas. É difícil transmitir o mesmo tipo de magnitude apelando para a imparcialidade, então escolhi usar com bomba de intuição de qualquer maneira; acredito que transmite o sentido básico correto de quão incompreensivelmente grande seria o valor dessa experiência.

[134]O cálculo aqui seria: se houver 10^{10} pessoas vivas hoje (isso é "arredondar para cima" de aproximadamente 8 bilhões até 10 bilhões) e cada uma tiver 10^{-71} (1 em 10^{71}) chance de ganhar o aposta, então cada um tem uma chance $(1-10^{-71})$ de perder a aposta. Portanto, a probabilidade de que **todos** percam a aposta é $(1-10^{-71})^{(10^{10})}$, que é, quase exatamente, 100%.

[135]Cálculo similar ao da nota de rodapé anterior, mas com uma população de 10^{70} (um para cada átomo na galáxia), então a probabilidade de que todos eles percam a aposta é $(1-10^{-71})^{(10^{70})}$, que acredito ser cerca de 90% (o Excel não pode realmente lidar com números tão grandes, mas é isso que cálculos semelhantes sugerem).

[136]Processo para automação do desenvolvimento científico e tecnológico - a ser discutido em um artigo futuro.

[137]Mais precisamente, a maioria dos modelos sugere que a automação total tanto da P&D quanto da produção de bens levaria a um crescimento explosivo. E quanto ao crescimento antes da automação total de ambas as coisas? Primeiro, se a automação prosseguir mais rapidamente do que sua taxa histórica antes da automação total, os modelos de crescimento geralmente sugerem que o crescimento começará a acelerar antes de atingirmos a automação total (por exemplo, consulte a seção 6.1.4.2 do relatório). Em segundo lugar, se a P&D, mas não a produção de bens, for totalmente automatizada, acredito que isso seria suficiente para um crescimento explosivo (veja a seção 6.1.6 do relatório).

[138]Você poderia dizer que as ações dos séculos passados também tiveram efeitos em cascata que influenciarão este futuro. Mas eu responderia que os efeitos dessas ações foram altamente caóticos e imprevisíveis, em comparação com os efeitos de ações mais próximas no tempo do ponto em que ocorre a transição.

[139]Geralmente emprego o termo "seres" em vez de "pessoas" para indicar que estou tentando me referir a todas as pessoas, animais ou coisas (IA?) cujo bem-estar devemos nos preocupar.

[140]Ainda mais do que você imagina intuitivamente, conforme descrito aqui

[141]Escrevi um pouco sobre essa perspectiva há vários anos, **aqui**.