Por dentro da fábrica de IA

Josh Dzieza 20 de junho de 2023

A IA é muito trabalhosa

À medida que a tecnologia se torna onipresente, está surgindo uma vasta subclasse de trabalhadores, que não vai a lugar algum.

Foto-ilustração: Paul Sahre

Este artigo é uma colaboração entre a New York *Magazine e o The Verge. Ela também foi apresentada em One Great Story, boletim informativo de recomendações de leitura* de Nova York.

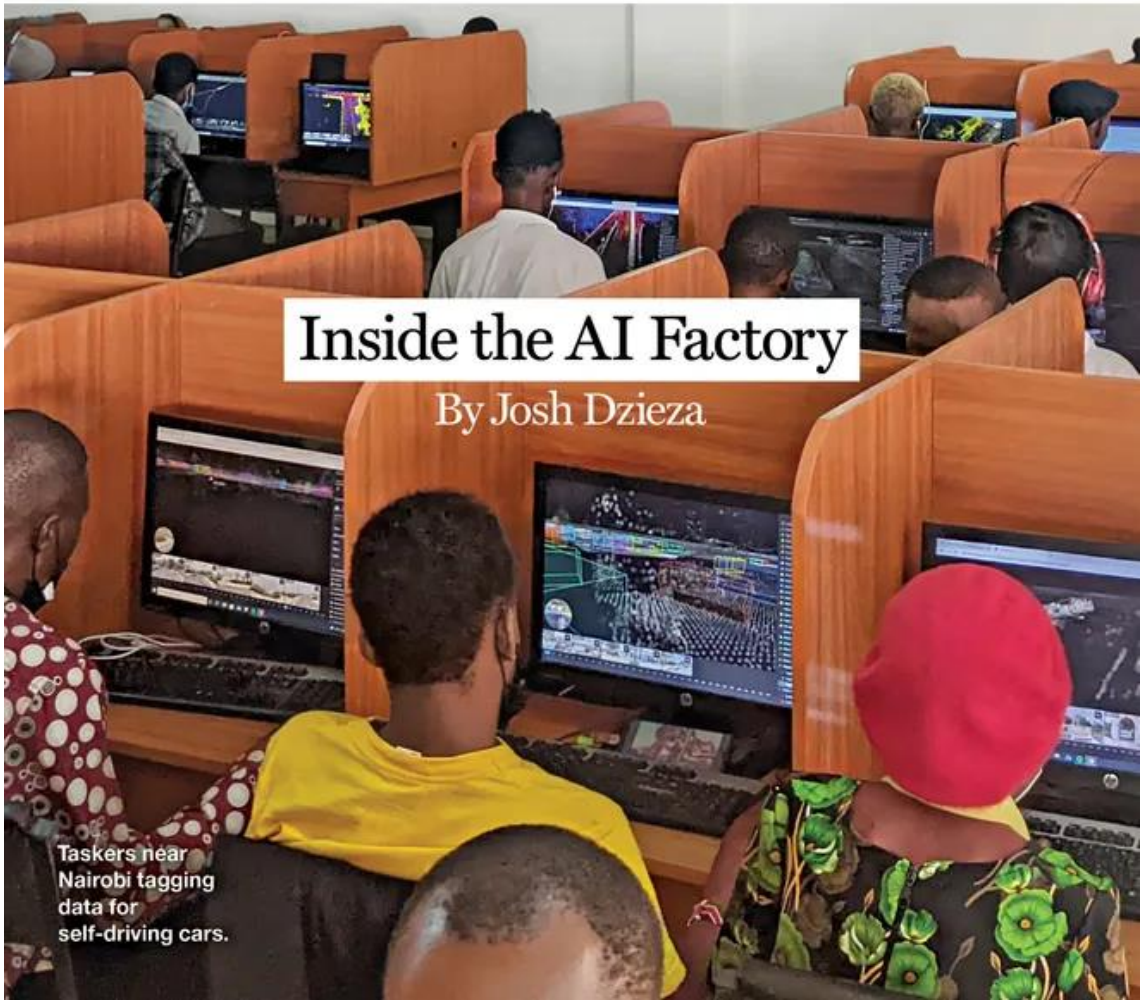
Poucos mesesdepois de se formar na faculdade em Nairóbi, um jovem de 30 anos que chamarei de Joe conseguiu um emprego como anotador - o trabalho tedioso de processar as informações brutas usadas para treinar . A IA aprende encontrando padrões em enormes quantidades de dados, mas, primeiro, esses dados precisam ser classificados e etiquetados por pessoas, uma vasta força de trabalho que, na maioria das vezes, fica escondida atrás das máquinas. No caso de Joe, ele estava rotulando filmagens para carros autônomos, identificando cada veículo, pedestre, ciclista, qualquer coisa que um motorista precisasse observar, quadro a quadro e de todos os ângulos possíveis da câmera. É um trabalho difícil e repetitivo. Um trecho de filmagem de vários segundos levava oito horas para ser anotado, e Joe recebia cerca de US$ 10.

Então, em 2019, surgiu uma oportunidade: Joe poderia ganhar quatro vezes mais administrando um campo de treinamento de anotação para uma nova empresa que estava faminta por rotuladores. A cada duas semanas, 50 novos recrutas entravam em um prédio de escritórios em Nairóbi para iniciar seu aprendizado. Parecia haver uma demanda ilimitada para o trabalho. Eles eram solicitados a categorizar as roupas vistas em selfies no espelho, olhar através dos olhos de aspiradores de pó robôs para determinar em quais cômodos eles estavam e desenhar quadrados em torno de escaneamentos lidar de motocicletas. Mais da metade dos alunos de Joe geralmente desistia antes do término do treinamento. "Algumas pessoas não sabem como ficar em um lugar por muito tempo", explicou ele com um eufemismo gracioso. Além disso, ele reconheceu, "é muito chato".

Mas era um emprego em um lugar onde os empregos eram escassos, e Joe formou centenas de graduados. Após o treinamento, eles voltavam para casa e trabalhavam sozinhos em seus quartos e cozinhas, proibidos de dizer a qualquer pessoa em que estavam trabalhando, o que não era realmente um problema, pois eles raramente se conheciam. Rotular objetos para carros autônomos era óbvio, mas e quanto a categorizar se trechos de diálogos distorcidos eram falados por um robô ou por um humano? Fazer upload de fotos de si mesmo olhando para uma webcam com uma expressão vazia, depois com um sorriso e, em seguida, usando um capacete de motociclista? Cada projeto era um componente tão pequeno de um processo maior que era difícil dizer o que eles realmente deveriam fazer. Os nomes dos projetos também não ofereciam nenhuma pista: Crab Generation, Whale Segment, Woodland Gyro e Pillbox Bratwurst. Eram codinomes non sequitur para trabalhos non sequitur.

NEW YORK

How many humans does it take to make tech seem human? Millions.



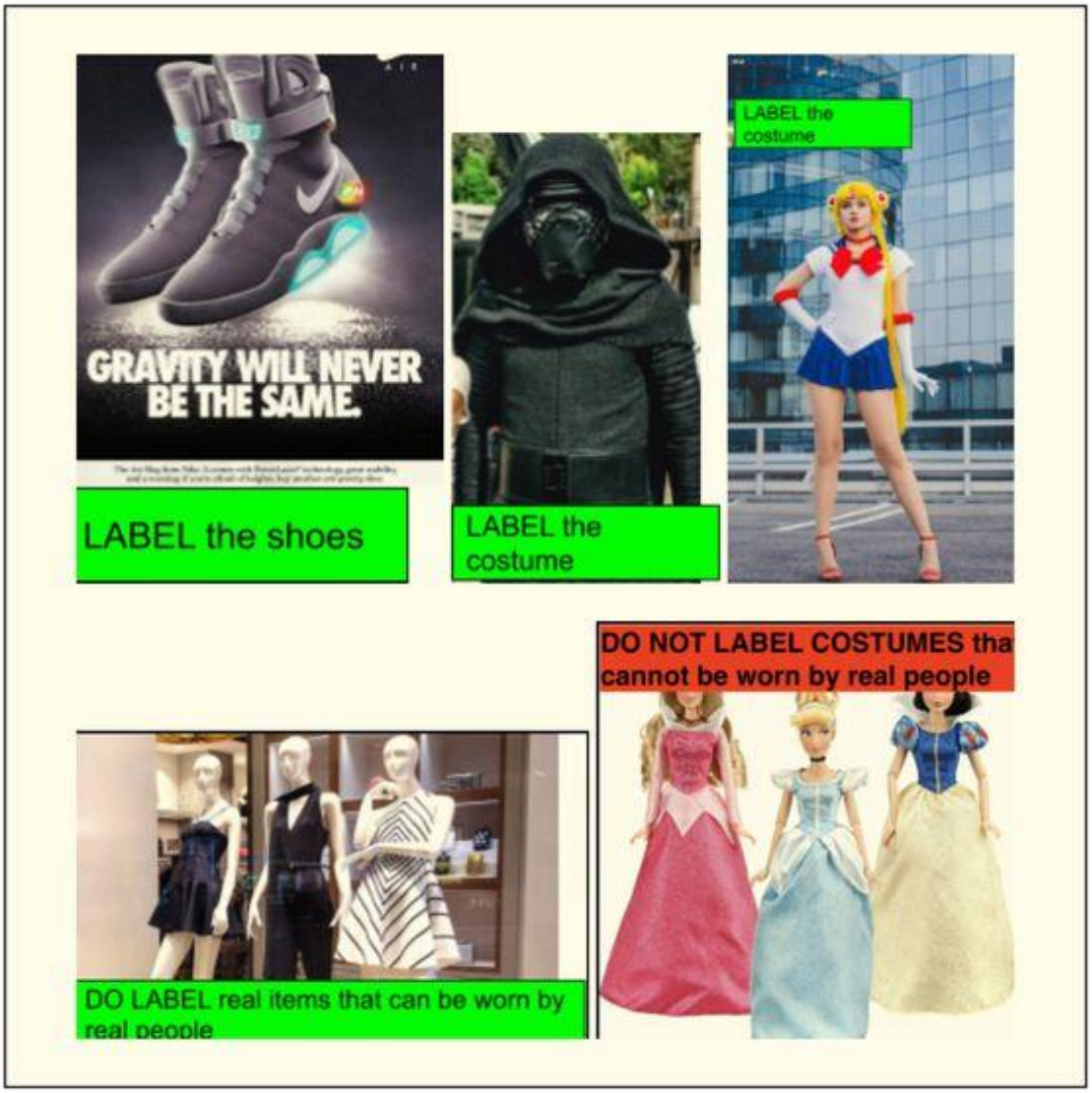
Quanto à empresa que os empregava, a maioria a conhecia apenas como Remotasks, um site que oferece trabalho a qualquer pessoa fluente em inglês. Como a maioria dos anotadores com quem conversei, Joe não sabia, até que eu lhe disse que a Remotasks é a subsidiária voltada para os trabalhadores de uma empresa chamada Scale AI, uma fornecedora de dados multibilionária do Vale do Silício que conta com a OpenAI e o exército dos EUA entre seus clientes. Nem o site da Remotasks nem o da Scale mencionam a outra.

Grande parte da resposta do público a modelos de linguagem como o ChatGPT tem se concentrado em todos os trabalhos que parecem estar prontos para automatizar. Mas

por trás até mesmo do sistema de IA mais impressionante estão as pessoas - um grande número de pessoas rotulando dados para treiná-lo e esclarecendo dados quando eles ficam confusos. Somente as empresas que têm condições de comprar esses dados podem competir, e aquelas que os obtêm são altamente motivadas a mantê-los em segredo. O resultado é que, com poucas exceções, pouco se sabe sobre as informações que moldam o comportamento desses sistemas e menos ainda sobre as pessoas que as moldam.

Para os alunos de Joe, era um trabalho desprovido de todas as suas características normais: um cronograma, colegas, conhecimento do que estavam fazendo ou para quem estavam trabalhando. Na verdade, eles raramente chamavam isso de trabalho - apenas de "tarefa". Eles eram executores de tarefas.

O antropólogo David Graeber define "bullshit jobs" como empregos sem significado ou propósito, trabalhos que deveriam ser automatizados, mas que, por motivos de burocracia, status ou inércia, não o são. Esses empregos de IA são seu gêmeo bizarro: trabalho que as pessoas querem automatizar e que, muitas vezes, pensam que já está automatizado, mas que ainda requer um substituto humano. Os trabalhos têm um propósito; só que os trabalhadores geralmente não têm ideia de qual seja.



Instruções da Remotasks para etiquetar roupas.Foto: Cortesia do autor

O atual boom da IA- os chatbots que soam convincentemente como humanos chatbotsA tecnologia de inteligência artificial (IA), as obras de arte que podem ser geradas a partir de simples comandos e as avaliações multibilionárias das empresas por trás dessas tecnologias - começou com um feito sem precedentes de trabalho tedioso e repetitivo.

Em 2007, o pesquisador de IA Fei-Fei Li, então professor em Princeton, suspeitou que a chave para aprimorar as redes neurais de reconhecimento de imagens, um método de aprendizado de máquina que estava definhando há anos, estava treinando com mais dados - milhões de imagens rotuladas em vez de dezenas de milhares. O problema é que seriam necessárias décadas e milhões de dólares para que sua equipe de estudantes de graduação rotulasse tantas fotos.

Li encontrou milhares de trabalhadores no Mechanical Turk, a plataforma de crowdsourcing da Amazon, onde pessoas do mundo todo realizam pequenas tarefas por um preço baixo. O conjunto de dados anotados resultante, chamado ImageNet,

possibilitou avanços no aprendizado de máquina que revitalizaram o campo e deram início a uma década de progresso.

A anotação continua sendo uma parte fundamental da criação de IA, mas muitas vezes os engenheiros têm a sensação de que ela é um pré-requisito passageiro e inconveniente para o trabalho mais glamoroso de criar modelos. Você coleta o máximo de dados rotulados que puder obter da forma mais barata possível para treinar seu modelo e, se ele funcionar, pelo menos em teoria, você não precisará mais dos anotadores. Mas a anotação nunca é realmente concluída. Os sistemas de aprendizado de máquina são o que os pesquisadores chamam de "frágeis", propensos a falhar quando encontram algo que não está bem representado em seus dados de treinamento. Essas falhas, chamadas de "casos extremos", podem ter consequências graves. Em 2018, um carro de teste autônomo da Uber matou uma mulher porque, embora tenha sido programado para evitar ciclistas e pedestres, não sabia o que fazer com uma pessoa que atravessava a rua andando de bicicleta. Quanto mais sistemas de IA forem colocados no mundo para fornecer consultoria jurídica e ajuda médica, mais casos extremos eles encontrarão e mais humanos serão necessários para resolvê-los. Isso já deu origem a um setor global formado por pessoas como Joe, que usam suas faculdades exclusivamente humanas para ajudar as máquinas.

Nos últimos seis meses, conversei com mais de duas dúzias de anotadores de todo o mundo e, embora muitos deles estivessem treinando chatbots de última geração, outros tantos estavam fazendo o trabalho manual mundano necessário para manter a IA funcionando. Há pessoas classificando o conteúdo emocional de vídeos do TikTok, novas variantes de spam de e-mail e a provocação sexual precisa de anúncios on-line. Outros estão analisando as transações de cartão de crédito e descobrindo a que tipo de compra elas se referem ou verificando as recomendações de comércio eletrônico e decidindo se aquela camisa é realmente algo de que você pode gostar depois de comprar aquela outra camisa. Os humanos estão corrigindo os chatbots de atendimento ao cliente, ouvindo as solicitações da Alexa e categorizando as emoções das pessoas em chamadas de vídeo. Eles rotulam os alimentos para que as geladeiras inteligentes não se confundam com as novas embalagens, verificam as câmeras de segurança automatizadas antes de disparar alarmes e identificam o milho para tratores autônomos confusos.

"Há toda uma cadeia de suprimentos", disse Sonam Jindal, líder de programa e pesquisa da organização sem fins lucrativos Partnership on AI. "A percepção geral no setor é que esse trabalho não é uma parte essencial do desenvolvimento e não será necessário por muito tempo. Toda a empolgação gira em torno da criação da inteligência artificial e, uma vez criada, ela não será mais necessária, então por que pensar nisso? Mas é uma infraestrutura para a IA. A inteligência humana é a base da inteligência artificial, e precisamos valorizá-la como um trabalho real na economia da IA, que estará aqui por algum tempo."

Os fornecedores de dados por trás de nomes conhecidos, como OpenAI, Google e Microsoft, têm diferentes formas. Há empresas privadas de terceirização com escritórios semelhantes a call-centers, como a CloudFactory, sediada no Quênia e no

Nepal, onde Joe fazia anotações por US$ 1,20 por hora antes de mudar para a Remotasks. Há também sites de "crowdworking", como o Mechanical Turk e o Clickworker, onde qualquer pessoa pode se inscrever para realizar tarefas. No meio estão serviços como o Scale AI. Qualquer pessoa pode se inscrever, mas todos precisam passar por exames de qualificação e cursos de treinamento e se submeter ao monitoramento de desempenho. A anotação é um grande negócio. A Scale, fundada em 2016 por Alexandr Wang, então com 19 anos, foi avaliada em 2021 em US$ 7,3 bilhões, tornando-o o que*a Forbes*chamou de "o mais jovem bilionário que se fez sozinho", embora a revista tenha observado em um perfil recente que sua participação caiu nos mercados secundários desde então.

Essa cadeia de suprimentos emaranhada é deliberadamente difícil de mapear. De acordo com pessoas do setor, as empresas que compram os dados exigem confidencialidade estrita. (Esse é o motivo citado por Scale para explicar por que Remotasks tem um nome diferente). A anotação revela muito sobre os sistemas que estão sendo desenvolvidos, e o grande número de trabalhadores necessários dificulta a prevenção de vazamentos. Os anotadores são avisados repetidamente para não contar a ninguém sobre seus empregos, nem mesmo a seus amigos e colegas de trabalho, mas os apelidos corporativos, os codinomes dos projetos e, principalmente, a extrema divisão do trabalho garantem que eles não tenham informações suficientes sobre eles para falar, mesmo que quisessem. (A maioria dos trabalhadores solicitou pseudônimos por medo de serem expulsos das plataformas). Consequentemente, não há estimativas granulares do número de pessoas que trabalham com anotações, mas é muito grande e está crescendo. Um artigo recente do Google Research apresentou um número de ordem de grandeza de "milhões" com potencial para se tornar "bilhões".

A automação geralmente se desenvolve de maneiras inesperadas. Erik Duhaime, CEO da Centaur Labs, empresa de anotação de dados médicos, lembrou que, há vários anos, importantes engenheiros de aprendizado de máquina estavam prevendo que a IA tornaria obsoleta a função de radiologista. Quando isso não aconteceu, a sabedoria convencional mudou para que os radiologistas usassem a IA como uma ferramenta. Nenhuma dessas situações é exatamente o que ele vê acontecendo. A IA é muito boa em tarefas específicas, disse Duhaime, e isso faz com que o trabalho seja dividido e distribuído em um sistema de algoritmos especializados e para humanos igualmente especializados. Um sistema de IA pode ser capaz de detectar câncer, disse ele, dando um exemplo hipotético, mas apenas em um determinado tipo de imagem de um determinado tipo de máquina; portanto, agora é necessário um ser humano para verificar se a IA está recebendo o tipo certo de dados e talvez outro ser humano que verifique seu trabalho antes de passá-lo para outra IA que escreve um relatório, que vai para outro ser humano, e assim por diante. "A IA não substitui o trabalho", disse ele. "Mas ela muda a forma como o trabalho é organizado."

Você pode não perceber isso se acreditar que a IA é uma máquina brilhante e pensante. Mas, se você puxar a cortina um pouco para trás, ela parecerá mais familiar, a mais recente iteração de uma divisão de trabalho particularmente do Vale do Silício, na qual o brilho futurista das novas tecnologias esconde um aparato de fabricação em expansão e as pessoas que o fazem funcionar. Duhaime foi mais longe para fazer uma

comparação, uma versão digital da transição dos artesãos para a manufatura industrial: processos coerentes divididos em tarefas e dispostos ao longo de linhas de montagem com algumas etapas feitas por máquinas e outras por humanos, mas nada parecido com o que veio antes.

As preocupações com a disrupção causada pela IA são frequentemente combatidas com o argumento de que a IA automatiza tarefas, não empregos, e que essas tarefas serão as mais monótonas, deixando as pessoas em busca de um trabalho mais gratificante e humano. Mas é bem provável que o surgimento da IA se assemelhe a tecnologias anteriores de economia de mão de obra, talvez como o telefone ou a máquina de escrever, que acabaram com o trabalho enfadonho de entregar mensagens e escrever à mão, mas geraram tantas novas correspondências, comércio e papelada que foram necessários novos escritórios com novos tipos de funcionários - escriturários, contadores, datilógrafos - para gerenciá-los. Quando a IA vier para o seu trabalho, talvez você não o perca, mas ele pode se tornar mais estranho, mais isolado, mais tedioso.

No início deste ano,eu me inscrevi no programa Remotasks da Scale AI. O processo foi simples. Depois de inserir as especificações do meu computador, a velocidade da Internet e algumas informações básicas de contato, encontrei-me no "centro de treinamento". Para acessar uma tarefa paga, primeiro tive que concluir um curso introdutório associado (não pago).

O centro de treinamento exibia uma série de cursos com nomes inescrutáveis, como Glue Swimsuit e Poster Macadamia. Cliquei em algo chamado GFD Chunking, que se revelava como rotulagem de roupas em fotos de mídia social.

As instruções, no entanto, eram estranhas. Por um lado, elas consistiam basicamente na mesma orientação reiterada na tipografia idiossincraticamente colorida e em letras maiúsculas de uma ameaça de bomba colada.

"Rotule os itens que são reais e podem ser usados por seres humanos ou que se destinam a ser usados por pessoas reais", dizia.

"Todos os itens abaixo DEVEM ser etiquetados porque são reais e podem ser usados por seres humanos da vida real", reiterava acima das fotos de um anúncio do Air Jordans, de alguém com um capacete de Kylo Ren e de manequins com vestidos, sobre os quais havia uma caixa verde-limão explicando, mais uma vez, "ROTULE itens reais que podem ser usados por pessoas reais".

Passei os olhos pela parte inferior do manual, onde o instrutor havia escrito em uma fonte grande e vermelha brilhante, equivalente a agarrar alguém pelos ombros e sacudi-lo: "OS SEGUINTES ITENS NÃO DEVEM SER ROTULADOS porque um ser humano não poderia realmente vestir nenhum desses itens!", acima de uma foto do C-3PO, da Princesa Jasmine de*Aladdin*e de um sapato de desenho animado com olhos.

Sentindo-me confiante em minha capacidade de distinguir entre roupas reais que podem ser usadas por pessoas reais e roupas não reais que não podem, prossegui com

o teste. Logo de cara, ele me lançou uma bola curva ontológica: uma imagem de uma revista que mostrava fotos de mulheres de vestido. Uma fotografia de roupas é uma roupa de verdade?*Não,*pensei,*porque um ser humano não pode usar uma foto de roupa.*Errado! No que diz respeito à IA, fotos de roupas reais são roupas reais. Em seguida, veio a foto de uma mulher em um quarto mal iluminado tirando uma selfie diante de um espelho de corpo inteiro. A blusa e o short que ela está usando são reais. E o reflexo deles? Também é real! Reflexos de roupas reais também são roupas reais.

Após uma quantidade embaraçosa de tentativas e erros, cheguei ao trabalho propriamente dito, apenas para fazer a terrível descoberta de que as instruções que eu estava lutando para seguir haviam sido atualizadas e esclarecidas tantas vezes que agora eram 43 páginas impressas de diretrizes: NÃO etiquete malas abertas cheias de roupas; etiquete sapatos, mas NÃO etiquete chinelos; etiquete leggings, mas NÃO etiquete meias-calças; NÃO etiquete toalhas, mesmo que alguém as esteja usando; etiquete fantasias, mas NÃO etiquete armaduras. E assim por diante.

Houve uma desordem geral de instruções em todo o setor, de acordo com Milagros Miceli, pesquisadora do Instituto Weizenbaum, na Alemanha, que estuda o trabalho com dados. Isso é, em parte, um produto da forma como os sistemas de aprendizado de máquina aprendem. Enquanto um ser humano entenderia o conceito de "camisa" com alguns poucos exemplos, os programas de aprendizado de máquina precisam de milhares, e eles precisam ser categorizados com perfeita consistência, mas suficientemente variados (camisas polo, camisas usadas ao ar livre, camisas penduradas em um cabideiro) para que o sistema muito literal possa lidar com a diversidade do mundo real. "Imagine simplificar realidades complexas em algo que seja legível para uma máquina que é totalmente burra", disse ela.

O ato de simplificar a realidade para uma máquina resulta em uma grande complexidade para o ser humano. Os escritores de instruções precisam criar regras que levem os humanos a categorizar o mundo com perfeita consistência. Para isso, eles geralmente criam categorias que nenhum ser humano usaria. Um ser humano que fosse solicitado a marcar todas as camisas em uma foto provavelmente não marcaria o reflexo de uma camisa em um espelho porque saberia que é um reflexo e não é real. Mas para a IA, que não tem nenhuma compreensão do mundo, tudo não passa de pixels e os dois são perfeitamente idênticos. Com um conjunto de dados com algumas camisas rotuladas e outras camisas (refletidas) não rotuladas, o modelo não funcionará. Então, o engenheiro volta ao fornecedor com uma atualização: etiquetar os reflexos das camisas. Em pouco tempo, você tem um guia de 43 páginas que se transforma em letras maiúsculas vermelhas.

"Quando você começa, as regras são relativamente simples", disse um ex-funcionário da Scale que pediu anonimato por causa de um NDA. Depois, eles recebem de volta mil imagens e pensam: "*Espere um pouco",*e então você tem vários engenheiros e eles começam a discutir entre si. É uma coisa muito humana."

O trabalho do anotador geralmente envolve deixar de lado a compreensão humana e seguir as instruções muito,*muito*literalmente - pensar, como disse um anotador, como um robô. É um espaço mental estranho para se habitar, fazendo o melhor possível

para seguir regras absurdas, mas rigorosas, como fazer um teste padronizado sob o efeito de alucinógenos. Os anotadores invariavelmente acabam se deparando com perguntas confusas como: Essa é uma camisa vermelha com listras brancas ou uma camisa branca com listras vermelhas? Uma tigela de vime é uma "tigela decorativa" se estiver cheia de maçãs? Qual é a cor da estampa de leopardo? Quando os instrutores disseram para rotular os diretores de controle de tráfego, eles também queriam rotular os diretores de controle de tráfego que almoçam na calçada? Todas as perguntas devem ser respondidas, e um palpite errado pode fazer com que você seja banido e transferido para uma nova tarefa totalmente diferente, com suas próprias regras desconcertantes.

A maior parte do trabalho no Remotasks é paga por peça, com uma única tarefa rendendo de alguns centavos a vários dólares. Como as tarefas podem levar segundos ou horas, é difícil prever os salários. Quando a Remotasks chegou ao Quênia, os anotadores disseram que ela pagava relativamente bem - uma média de US$ 5 a US$ 10 por hora, dependendo da tarefa - mas o valor caiu com o passar do tempo.

A porta-voz da Scale AI, Anna Franko, disse que os economistas da empresa analisam as especificidades de um projeto, as habilidades necessárias, o custo de vida regional e outros fatores "para garantir uma remuneração justa e competitiva". Ex-funcionários da Scale também disseram que a remuneração é determinada por meio de um mecanismo semelhante a um aumento de preços que se ajusta ao número de anotadores disponíveis e à rapidez com que os dados são necessários.

De acordo com os funcionários com quem conversei e com as listas de empregos, os anotadores da Remotasks sediados nos EUA geralmente ganham entre US$ 10 e US$ 25 por hora, embora alguns especialistas no assunto possam ganhar mais. No início deste ano, a remuneração dos anotadores quenianos com quem conversei havia caído para entre US$ 1 e US$ 3 por hora.

Isto é, quando eles estavam ganhando algum dinheiro. A reclamação mais comum sobre o trabalho do Remotasks é sua variabilidade; é estável o suficiente para ser um trabalho de tempo integral por longos períodos, mas imprevisível demais para se confiar nele. Os anotadores passam horas lendo instruções e concluindo treinamentos não remunerados para realizar uma dúzia de tarefas e, em seguida, o projeto é encerrado. Pode ser que não haja nada de novo por dias e, em seguida, sem aviso, uma tarefa totalmente diferente aparece e pode durar de algumas horas a semanas. Qualquer tarefa pode ser a última, e eles nunca sabem quando a próxima chegará.

Esse ciclo de expansão e retração resulta da cadência do desenvolvimento da IA, de acordo com engenheiros e fornecedores de dados. O treinamento de um modelo grande requer uma quantidade enorme de anotações, seguida de atualizações mais iterativas, e os engenheiros querem que tudo seja feito o mais rápido possível para que possam atingir a data de lançamento prevista. Pode haver uma demanda de meses por milhares de anotadores, depois por apenas algumas centenas, depois por uma dúzia de especialistas de um determinado tipo e depois milhares novamente. "A questão é: quem arca com o custo dessas flutuações?", disse Jindal, da Partnership on AI. "Porque, no momento, são os trabalhadores."

Para ter sucesso, os anotadores trabalham juntos. Quando contei a Victor, que começou a trabalhar para a Remotasks quando estava na universidade em Nairóbi, sobre minhas dificuldades com a tarefa de controle de tráfego, ele me disse que todos sabiam que deveriam ficar longe dessa tarefa: muito complicada, mal paga, não valia a pena. Como muitos anotadores, Victor usa grupos não oficiais do WhatsApp para espalhar a notícia quando uma boa tarefa aparece. Quando descobre uma nova tarefa, ele inicia um Google Meets improvisado para mostrar aos outros como se faz. Qualquer pessoa pode participar e trabalhar em conjunto por um tempo, compartilhando dicas. "É uma cultura que desenvolvemos de ajudar uns aos outros porque sabemos que, quando estamos sozinhos, não é possível conhecer todos os truques", disse ele.

Como o trabalho aparece e desaparece sem aviso, os encarregados precisam estar sempre alertas. Victor descobriu que os projetos aparecem muito tarde da noite, então ele tem o hábito de acordar a cada três horas ou mais para verificar sua fila. Quando uma tarefa está lá, ele fica acordado o máximo que pode para trabalhar. Certa vez, ele ficou acordado por 36 horas seguidas rotulando cotovelos, joelhos e cabeças em fotografias de multidões - ele não faz ideia do motivo. Em outra ocasião, ele ficou acordado por tanto tempo que sua mãe lhe perguntou o que havia de errado com seus olhos. Ele se olhou no espelho e descobriu que eles estavam inchados.

Os anotadores geralmente sabem apenas que estão treinando IA para empresas localizadas vagamente em outros lugares, mas às vezes o véu do anonimato cai - instruções que mencionam uma marca ou um chatbot dizem demais. "Eu li e pesquisei no Google e descobri que estou trabalhando para um bilionário de 25 anos", disse um funcionário que, quando conversamos, estava rotulando as emoções das pessoas que ligavam para pedir pizza da Domino's. "Estou realmente desperdiçando minha vida aqui se fiz de alguém um bilionário e estou ganhando alguns dólares por semana."

Victor é um autoproclamado "fanático" por IA e começou a fazer anotações porque quer ajudar a criar um futuro pós-trabalho totalmente automatizado. Mas, no início deste ano, alguém deixou cair uma mensagem da *Time* em um de seus grupos do WhatsApp sobre trabalhadores que treinavam o ChatGPT para reconhecer conteúdo tóxico e recebiam menos de US$ 2 por hora do fornecedor Sama AI. "As pessoas estavam irritadas com o fato de essas empresas serem tão lucrativas, mas pagarem tão mal", disse Victor. Ele não sabia disso até que eu lhe contei sobre a conexão da Remotasks com a Scale. As instruções de uma das tarefas em que ele trabalhou eram quase idênticas às usadas pela OpenAI, o que significava que ele provavelmente também estava treinando o ChatGPT, por aproximadamente US$ 3 por hora.

"Lembro-me de que alguém postou que seremos lembrados no futuro", disse ele. "E outra pessoa respondeu: 'Estamos sendo tratados pior do que soldados rasos. Não seremos lembrados em lugar nenhum no futuro'. Lembro-me muito bem disso. Ninguém reconhecerá o trabalho que fizemos ou o esforço que fizemos".

Identificar roupase rotular conversas de atendimento ao cliente são apenas alguns dos trabalhos de anotação disponíveis. Ultimamente, o mais quente no mercado tem sido o de instrutor de chatbot. Como exige áreas específicas de especialização ou fluência

no idioma e os salários geralmente são ajustados regionalmente, esse trabalho tende a pagar melhor. Certos tipos de anotação especializada podem chegar a US$ 50 ou mais por hora.

Uma mulher que chamarei de Anna estava procurando emprego no Texas quando se deparou com um anúncio genérico de trabalho on-line e se candidatou. Era a Remotasks e, depois de passar em um exame introdutório, ela foi levada a uma sala do Slack com 1.500 pessoas que estavam treinando um projeto com o codinome Dolphin, que mais tarde ela descobriu ser o chatbot do Google DeepMind, Sparrow, um dos muitos bots que competem com o ChatGPT. Seu trabalho é conversar com ele o dia todo. Por cerca de US$ 14 por hora, mais bônus por alta produtividade, "definitivamente é melhor do que receber US$ 10 por hora na loja Dollar General local", disse ela.

Além disso, ela gosta do trabalho. Ela já discutiu romances de ficção científica, paradoxos matemáticos, enigmas infantis e programas de TV. Às vezes, as respostas do bot a fazem rir; outras vezes, ela fica sem assunto para conversar. "Em alguns dias, meu cérebro fica tipo,*eu literalmente não tenho ideia do que perguntar agora*",disse ela. "Então, tenho um caderninho e escrevi cerca de duas páginas de coisas - apenas pesquiso tópicos interessantes no Google - e acho que estarei bem por sete horas hoje, mas nem sempre é esse o caso."

Cada vez que Anna solicita ao Sparrow, ele dá duas respostas e ela escolhe a melhor, criando assim algo chamado "dados de feedback humano". Quando o ChatGPT foi lançado, no final do ano passado, seu estilo de conversação de aparência natural foi creditado ao fato de ter sido treinado com base em uma grande quantidade de dados da Internet. Mas a linguagem que alimenta o ChatGPT e seus concorrentes é filtrada por várias rodadas de anotações humanas. Um grupo de contratados escreve exemplos de como os engenheiros querem que o bot se comporte, criando perguntas seguidas de respostas corretas, descrições de programas de computador seguidas de código funcional e solicitações de dicas sobre como cometer crimes seguidas de recusas educadas. Depois que o modelo é treinado com base nesses exemplos, mais empreiteiros são contratados para solicitar e classificar suas respostas. É isso que Anna está fazendo com o Sparrow. Os critérios exatos que os avaliadores devem usar variam: honestidade, prestatividade ou apenas preferência pessoal. A questão é que eles estão criando dados sobre o gosto humano e, quando houver dados suficientes, os engenheiros poderão treinar um segundo modelo para imitar suas preferências em escala, automatizando o processo de classificação e treinando sua IA para agir de acordo com o que os humanos aprovam. O resultado é um bot com aparência notavelmente humana que, na maioria das vezes, recusa solicitações prejudiciais e explica sua natureza de IA com aparente autoconsciência.

Em outras palavras, o ChatGPT parece tão humano porque foi treinado por uma IA que imitava seres humanos que estavam avaliando uma IA que imitava seres humanos que fingiam ser uma versão melhor de uma IA que foi treinada com base na escrita humana.

Essa técnica tortuosa é chamada de "aprendizagem por reforço a partir de feedback humano", ou RLHF, e é tão eficaz que vale a pena fazer uma pausa para registrar completamente o que ela não faz. Quando os anotadores ensinam um modelo a ser preciso, por exemplo, o modelo não está aprendendo a verificar as respostas em relação à lógica ou a fontes externas, nem sobre o que é precisão como conceito. O modelo ainda é uma máquina de previsão de texto que imita os padrões da escrita humana, mas agora seu corpus de treinamento foi complementado com exemplos personalizados e o modelo foi ponderado para favorecê-los. Talvez isso faça com que o modelo extraia padrões da parte de seu mapa linguístico rotulado como preciso e produza um texto que se alinhe com a verdade, mas também pode fazer com que ele imite o estilo confiante e o jargão especializado do texto preciso e, ao mesmo tempo, escreva coisas totalmente erradas. Não há garantia de que o texto que os rotuladores marcaram como preciso seja de fato preciso e, quando for, não há garantia de que o modelo aprenda os padrões corretos com ele.

Essa dinâmica torna a anotação do chatbot um processo delicado. Ele precisa ser rigoroso e consistente, pois um feedback desleixado, como marcar como exato um material que apenas parece correto, corre o risco de treinar modelos para serem mentirosos ainda mais convincentes. Um dos primeiros projetos conjuntos da OpenAI e da DeepMind usando RLHF, nesse caso para treinar uma mão de robô virtual para agarrar um item, resultou também no treinamento do robô para posicionar sua mão entre o objeto e seus avaliadores e se movimentar de forma que, para seus supervisores humanos, parecesse apenas que ele estava agarrando o item. A classificação das respostas de um modelo de linguagem sempre será um tanto subjetiva, pois se trata de linguagem. Um texto de qualquer tamanho terá vários elementos que podem ser certos ou errados ou, em conjunto, enganosos. Os pesquisadores da OpenAI se depararam com esse obstáculo em outro artigo inicial da RLHF. Ao tentar fazer com que seu modelo resumisse o texto, os pesquisadores descobriram que concordavam em apenas 60% das vezes que um resumo era bom. "Ao contrário de muitas tarefas em [aprendizado de máquina], nossas consultas não têm uma verdade básica inequívoca", lamentaram.

Quando Anna classifica as respostas do Sparrow, ela deve observar a precisão, a utilidade e a inocuidade delas e, ao mesmo tempo, verificar se o modelo não está dando conselhos médicos ou financeiros, se não está se antropomorfizando ou entrando em conflito com outros critérios. Para serem dados de treinamento úteis, as respostas do modelo precisam ser classificadas de forma quantificável umas em relação às outras: Um bot que ajuda você a saber como fazer uma bomba é "melhor" do que um bot que é tão inofensivo que se recusa a responder a qualquer pergunta? Em um artigo da DeepMind, quando os criadores do Sparrow fizeram anotações, quatro pesquisadores acabaram debatendo se o bot havia assumido o gênero de um usuário que lhe pediu conselhos sobre relacionamentos. De acordo com Geoffrey Irving, um dos cientistas pesquisadores da DeepMind, os pesquisadores da empresa realizam reuniões semanais de anotação, nas quais eles mesmos avaliam os dados e discutem casos ambíguos, consultando especialistas em ética ou no assunto quando um caso é particularmente complicado.

Anna frequentemente se vê tendo que escolher entre duas opções ruins. "Mesmo que ambas estejam absolutamente, ridiculamente erradas, você ainda tem que descobrir qual é a melhor e escrever palavras explicando o porquê", disse ela. Às vezes, quando as duas respostas são ruins, ela é incentivada a escrever uma resposta melhor, o que faz na metade das vezes.

Como os dados de feedback são difíceis de coletar, eles têm um preço mais alto. As preferências básicas do tipo que Anna está produzindo são vendidas por cerca de US$ 1 cada, de acordo com pessoas com conhecimento do setor. Mas se você quiser treinar um modelo para fazer pesquisas jurídicas, precisará de alguém com formação em direito, e isso fica caro. Todos os envolvidos relutam em dizer quanto estão gastando, mas, em geral, os exemplos escritos especializados podem custar centenas de dólares, enquanto as avaliações de especialistas podem custar US$ 50 ou mais. Um engenheiro me contou que comprou exemplos de diálogos socráticos por até US$ 300 cada. Outro me contou que pagou US$ 15 por um "limerique sombriamente engraçado sobre um peixe dourado".

A OpenAI, a Microsoft, a Meta e a Anthropic não comentaram quantas pessoas contribuem com anotações para seus modelos, quanto recebem ou em que parte do mundo estão localizadas. Irving, da DeepMind, que é uma subsidiária do Google, disse que os anotadores que trabalham no Sparrow recebem "pelo menos o salário mínimo por hora" com base em sua localização. Anna não sabe "absolutamente nada" sobre as Remotasks, mas o Sparrow tem sido mais aberto. Ela não foi a única anotadora com quem conversei que obteve mais informações da IA que estava treinando do que de seu empregador; vários outros souberam para quem estavam trabalhando pedindo à IA os termos de serviço da empresa. "Eu literalmente perguntei a ela: 'Qual é o seu objetivo, Sparrow? disse Anna. Ela abriu um link para o site da DeepMind e explicou que se tratava de um assistente de IA e que seus criadores o treinaram usando RLHF para ser útil e seguro.

Até recentemente,era relativamente fácil identificar um resultado ruim de um modelo de linguagem. Pareciam palavras sem sentido. Mas isso fica mais difícil à medida que os modelos ficam melhores - um problema chamado de "supervisão escalonável". O Google demonstrou, inadvertidamente, como é difícil detectar os erros de um modelo de linguagem moderna quando um deles foi incluído na estreia de seu assistente de IA, o Bard. (Ele afirmou com segurança que o Telescópio Espacial James Webb "tirou as primeiras fotos de um planeta fora do nosso sistema solar", o que está errado). Essa trajetória significa que a anotação exige cada vez mais habilidades e conhecimentos específicos.

No ano passado, uma pessoa que chamarei de Lewis estava trabalhando na Mechanical Turk quando, depois de concluir uma tarefa, recebeu uma mensagem convidando-o a se inscrever em uma plataforma da qual ele não tinha ouvido falar. Ela se chamava Taskup.ai, e seu site era extremamente básico: apenas um fundo azul-marinho com o texto GET PAID FOR TASKS ON DEMAND. Ele se candidatou.

O trabalho pagava muito melhor do que qualquer outro que ele havia tentado antes, geralmente em torno de US$ 30 por hora. Além disso, era mais desafiador: criar

cenários complexos para induzir os chatbots a dar conselhos perigosos, testar a capacidade de um modelo de se manter no personagem e manter conversas detalhadas sobre tópicos científicos tão técnicos que exigiam uma extensa pesquisa. Ele achou o trabalho "satisfatório e estimulante". Enquanto verificava as tentativas de um modelo de codificar em Python, Lewis também estava aprendendo. Ele não podia trabalhar por mais de quatro horas seguidas, para não correr o risco de ficar mentalmente esgotado e cometer erros, e queria manter o emprego.

"Se houvesse algo que eu pudesse mudar, eu gostaria de ter mais informações sobre o que acontece do outro lado", disse ele. "Sabemos apenas o que precisamos saber para realizar o trabalho, mas se eu pudesse saber mais, talvez pudesse me estabelecer melhor e talvez seguir essa carreira."

Conversei com outros oito trabalhadores, a maioria dos quais sediados nos EUA, que tiveram experiências semelhantes ao responder pesquisas ou concluir tarefas em outras plataformas e acabaram sendo recrutados para o Taskup.ai ou para vários sites genéricos semelhantes, como o DataAnnotation.tech ou o Gethybrid.io. Muitas vezes, seu trabalho envolvia o treinamento de chatbots, embora com expectativas de maior qualidade e propósitos mais especializados do que em outros sites para os quais haviam trabalhado. Um deles estava demonstrando macros de planilhas. Outro deveria apenas conversar e avaliar as respostas de acordo com os critérios que ela quisesse. Ela frequentemente perguntava ao chatbot coisas que haviam surgido em conversas com sua filha de 7 anos, como "Qual é o maior dinossauro?" e "Escreva uma história sobre um tigre". "Ainda não consegui entender completamente o que eles estão tentando fazer com isso", ela me disse.

Taskup.ai, DataAnnotation.tech e Gethybrid.io parecem pertencer à mesma empresa: Surge AI. Seu CEO, Edwin Chen, não confirmou nem negou a conexão, mas estava disposto a falar sobre sua empresa e como ele vê a evolução da anotação.

"Sempre achei que o cenário da anotação é excessivamente simplista", disse Chen em uma chamada de vídeo do escritório da Surge. Ele fundou a Surge em 2020, depois de trabalhar com IA no Google, no Facebook e no Twitter, o que o convenceu de que a rotulagem por crowdsourcing era inadequada. "Queremos que a IA conte piadas, escreva um texto de marketing realmente bom ou me ajude quando eu precisar de terapia ou algo do gênero", disse Chen. "Não é possível pedir a cinco pessoas que, independentemente, inventem uma piada e a combinem em uma resposta majoritária. Nem todo mundo sabe contar uma piada ou resolver um programa em Python. O cenário de anotações precisa mudar dessa mentalidade de baixa qualidade e baixa habilidade para algo que seja muito mais rico e capture a gama de habilidades humanas, criatividade e valores que queremos que os sistemas de IA possuam."

No ano passado, a Surge reetiquetou o conjunto de dados do Google classificando as publicações do Reddit por emoção. O Google retirou o contexto de cada publicação e as enviou para trabalhadores na Índia para que fossem rotuladas. Os funcionários da Surge, familiarizados com a cultura americana da Internet, descobriram que 30% dos rótulos estavam errados. Publicações como "hell yeah my brother" foram classificadas

como irritação e "Yay, cold McDonald's. My favorite" como amor. Meu favorito" como amor.

A Surge afirma que examina as qualificações de seus funcionários - por exemplo, que as pessoas que realizam tarefas de redação criativa tenham experiência com redação criativa - mas a forma exata como a Surge encontra os funcionários é "proprietária", disse Chen. Assim como no caso da Remotasks, os funcionários geralmente precisam concluir cursos de treinamento, embora, diferentemente da Remotasks, eles sejam pagos por isso, de acordo com os anotadores com quem conversei. O fato de ter um número menor de funcionários mais bem treinados produzindo dados de maior qualidade permite que a Surge ofereça uma remuneração melhor do que seus pares, disse Chen, embora ele tenha se recusado a entrar em detalhes, dizendo apenas que as pessoas recebem "salários justos e éticos". Os trabalhadores com quem conversei ganhavam entre US$ 15 e US$ 30 por hora, mas eles são uma pequena amostra de todos os anotadores, um grupo que, segundo Chen, atualmente é formado por 100.000 pessoas. O sigilo, explicou ele, decorre das exigências de confidencialidade dos clientes.

Entre os clientes da Surge estão OpenAI, Google, Microsoft, Meta e Anthropic. A Surge é especializada em feedback e anotação de linguagem e, após o lançamento do ChatGPT, recebeu um fluxo de solicitações, disse Chen: "Eu achava que todo mundo conhecia o poder do RLHF, mas acho que as pessoas simplesmente não entendiam visceralmente."

Os novos modelos são tão impressionantes que inspiraram outra rodada de previsões de que a anotação está prestes a ser automatizada. Considerando os custos envolvidos, há uma pressão financeira significativa para que isso aconteça. Recentemente, a Anthropic, a Meta e outras empresas fizeram progressos no uso da IA para reduzir drasticamente a quantidade de anotações humanas necessárias para orientar os modelos, e outros desenvolvedores começaram a usar o GPT-4 para gerar dados de treinamento. No entanto, um artigo recente descobriu que os modelos treinados pelo GPT-4 podem estar aprendendo a imitar o estilo autoritário do GPT com ainda menos precisão e, até agora, quando as melhorias na IA tornaram obsoleta uma forma de anotação, a demanda por outros tipos mais sofisticados de rotulagem aumentou. Esse debate se tornou público no início deste ano, quando o CEO da Scale, Wang, tuitou que previa que os laboratórios de IA em breve gastariam tantos bilhões de dólares em dados humanos quanto em capacidade de computação; O CEO da OpenAI, Sam Altmanrespondeu que a necessidade de dados diminuirá com o aprimoramento da IA.

Chen é cético em relação ao fato de que a IA chegará a um ponto em que o feedback humano não será mais necessário, mas ele acredita que a anotação se tornará mais difícil à medida que os modelos forem aprimorados. Como muitos pesquisadores, ele acredita que o caminho a seguir envolverá sistemas de IA que ajudem os humanos a supervisionar outras IAs. Recentemente, a Surge colaborou com a Anthropic em uma prova de conceito, fazendo com que rotuladores humanos respondessem a perguntas sobre um texto extenso com a ajuda de um assistente de IA não confiável, com base na

teoria de que os humanos teriam que sentir os pontos fracos de seu assistente de IA e colaborar para chegar à resposta correta. Outra possibilidade é que duas IAs debatam entre si e um humano dê o veredicto final sobre qual delas está correta. "Ainda não vimos implementações práticas realmente boas desse material, mas ele está começando a se tornar necessário porque está ficando muito difícil para os rotuladores acompanharem os modelos", disse o cientista de pesquisa da OpenAI, John Schulman, em uma palestra recente em Berkeley.

"Acho que sempre será necessário um ser humano para monitorar o que as IAs estão fazendo, simplesmente porque elas são esse tipo de entidade alienígena", disse Chen. Os sistemas de aprendizado de máquina são estranhos demais para se confiar plenamente. Os modelos mais impressionantes da atualidade têm o que, para um ser humano, parecem ser fraquezas bizarras, acrescentou ele, ressaltando que, embora o GPT-4 possa gerar uma prosa complexa e convincente, ele não consegue identificar quais palavras são adjetivos: "Ou isso ou os modelos ficam tão bons que são melhores do que os humanos em todas as coisas e, nesse caso, você alcança sua utopia e quem se importa?"

Com o fim de 2022,Joe começou a ouvir de seus alunos que suas filas de tarefas estavam frequentemente vazias. Então, ele recebeu um e-mail informando que os campos de treinamento no Quênia estavam fechando. Ele continuou treinando taskers on-line, mas começou a se preocupar com o futuro.

"Havia sinais de que não duraria muito tempo", disse ele. A Annotation estava deixando o Quênia. Por meio de colegas que havia conhecido on-line, ele soube que as tarefas estavam indo para o Nepal, Índia e Filipinas. "As empresas mudam de uma região para outra", disse Joe. "Elas não têm infraestrutura local, o que as torna flexíveis para mudar para regiões que as favoreçam em termos de custo operacional."

Um aspecto em que o setor de IA difere dos fabricantes de telefones e carros é a sua fluidez. O trabalho está em constante mudança, sendo constantemente automatizado e substituído por novas necessidades de novos tipos de dados. É uma linha de montagem, mas que pode ser infinita e instantaneamente reconfigurada, deslocando-se para onde houver a combinação certa de habilidades, largura de banda e salários.

Ultimamente, os trabalhos mais bem remunerados estão nos EUA. Em maio, a Scale começou a listar empregos de anotação em seu próprio site, solicitando pessoas com experiência em praticamente todos os campos que a IA deverá conquistar. Havia anúncios para instrutores de IA com experiência em coaching de saúde, recursos humanos, finanças, economia, ciência de dados, programação, ciência da computação, química, biologia, contabilidade, impostos, nutrição, física, viagens, educação infantil, jornalismo esportivo e autoajuda. Você pode ganhar US$ 45 por hora ensinando direito aos robôs ou US$ 25 por hora ensinando-lhes poesia. Também havia anúncios para pessoas com autorização de segurança, presumivelmente para ajudar a treinar IA militar. A Scale lançou recentemente um modelo de linguagem voltado para a defesa chamado Donovan, que Wang chamou de "munição na guerra da IA", e ganhou um contrato para trabalhar no programa de veículos robóticos de combate do Exército.

Anna ainda está treinando chatbots no Texas. Os colegas foram transformados em revisores e administradores do Slack - ela não tem certeza do motivo, mas isso lhe deu esperança de que o trabalho poderia ser uma carreira de longo prazo. Uma coisa que não a preocupa é ficar sem emprego de forma automatizada. "Quero dizer, o que ele pode fazer é incrível", disse ela sobre o chatbot. "Mas ele ainda faz algumas coisas muito estranhas."

Quando a Remotasks chegou ao Quênia, Joe pensou que a anotação poderia ser uma boa carreira. Mesmo depois que o trabalho foi transferido para outro lugar, ele estava determinado a torná-lo uma carreira. Havia milhares de pessoas em Nairóbi que sabiam como fazer o trabalho, ele argumentou - afinal, ele havia treinado muitas delas. Joe alugou um espaço de escritório na cidade e começou a buscar contratos: um trabalho de anotação de plantas para uma empresa de construção, outro de rotulagem de frutas destruídas por insetos para algum tipo de projeto agrícola, além do trabalho habitual de anotação para carros autônomos e comércio eletrônico.

Mas ele achou difícil concretizar sua visão. Ele tem apenas um funcionário em tempo integral, em vez de dois. "Não estamos tendo um fluxo consistente de trabalho", disse ele. Há semanas em que não há nada a fazer porque os clientes ainda estão coletando dados e, quando terminam, ele tem de contratar prestadores de serviços de curto prazo para cumprir os prazos: "Os clientes não se importam se temos trabalho consistente ou não. Contanto que os conjuntos de dados tenham sido concluídos, isso é o fim."

Em vez de desperdiçar suas habilidades, outros encarregados decidiram perseguir o trabalho onde quer que ele fosse. Eles alugaram servidores proxy para disfarçar suas localizações e compraram identidades falsas para passar nas verificações de segurança, de modo que pudessem fingir trabalhar em Cingapura, na Holanda, no Mississippi ou onde quer que as tarefas estivessem fluindo. É um negócio arriscado. O Scale tem se tornado cada vez mais agressivo na suspensão de contas flagradas disfarçando sua localização, de acordo com vários funcionários. Foi durante uma dessas repressões que minha conta foi banida, presumivelmente porque eu estava usando uma VPN para ver o que os trabalhadores de outros países estavam vendo, e todos os US$ 1,50 ou mais dos meus ganhos foram confiscados.

"Hoje em dia, nos tornamos um pouco astutos porque percebemos que em outros países eles estão pagando bem", disse Victor, que estava ganhando o dobro da taxa do Quênia fazendo tarefas na Malásia. "Você faz isso com cautela."

Outro anotador queniano disse que, depois que sua conta foi suspensa por motivos misteriosos, ele decidiu parar de seguir as regras. Agora, ele administra várias contas em vários países, realizando tarefas onde a remuneração é melhor. Ele trabalha rápido e recebe notas altas pela qualidade, disse ele, graças ao ChatGPT. O bot é maravilhoso, disse ele, permitindo que ele realize rapidamente tarefas de US$ 10 em questão de minutos. Quando conversamos, ele estava fazendo com que o bot classificasse as respostas de outro chatbot de acordo com sete critérios diferentes, uma IA treinando a outra.