Pesquisadores da Universidade de Stanford pagaram US$ 60 a 1.052 pessoas para lerem as duas primeiras linhas do *O Grande Gatsby* para um aplicativo. Feito isso, uma IA que parecia um sprite 2D de um jogo Final Fantasy da era SNES pediu aos participantes que contassem a história de suas vidas. Os cientistas pegaram essas entrevistas e as transformaram em uma IA que, segundo eles, replica o comportamento dos participantes com 85% de precisão.

O estudo, intitulado *Simulações de agentes geradores de 1.000 pessoas*é uma joint venture entre Stanford e cientistas que trabalham para o laboratório de pesquisa DeepMind AI do Google. A ideia é que a criação de agentes de IA baseados em pessoas aleatórias poderia ajudar os legisladores e empresários a compreender melhor o público. Por que usar grupos focais ou pesquisar o público quando você pode conversar com eles uma vez, criar um LLM com base nessa conversa e depois ter seus pensamentos e opiniões para sempre? Ou, pelo menos, uma aproximação tão próxima desses pensamentos e sentimentos quanto um LLM é capaz de recriar.

“Este trabalho fornece uma base para novas ferramentas que podem ajudar a investigar o comportamento individual e coletivo”, afirma o resumo do artigo.

“Como poderá, por exemplo, um conjunto diversificado de indivíduos responder a novas políticas e mensagens de saúde pública, reagir ao lançamento de produtos ou responder a grandes choques?” O jornal continuou. “Quando indivíduos simulados são combinados em coletivos, essas simulações podem ajudar a pilotar intervenções, desenvolver teorias complexas que capturam interações causais e contextuais diferenciadas e expandir nossa compreensão de estruturas como instituições e redes em domínios como economia, sociologia, organizações e ciência política. ”

Todas essas possibilidades baseadas em uma entrevista de duas horas alimentaram um LLM que respondeu a perguntas principalmente como suas contrapartes da vida real.

Grande parte do processo foi automatizado. Os pesquisadores contrataram a Bovitz, empresa de pesquisa de mercado, para reunir os participantes. O objetivo era obter uma amostra ampla da população dos EUA, tão ampla quanto possível quando limitada a 1.000 pessoas. Para concluir o estudo, os usuários criaram uma conta em uma interface feita sob medida, criaram um avatar de sprite 2D e começaram a conversar com um entrevistador de IA.

As perguntas e o estilo de entrevista são uma versão modificada daquela usada pelo American Voices Project, um projeto conjunto de Stanford e da Universidade de Princeton que entrevista pessoas em todo o país.

Cada entrevista começou com os participantes lendo as duas primeiras linhas do *O Grande Gatsby* (“Em meus anos mais jovens e mais vulneráveis, meu pai me deu alguns conselhos que venho pensando desde então. ‘Sempre que você sentir vontade de criticar alguém’, ele me disse, ‘lembre-se de que todas as pessoas em este mundo não teve as vantagens que

você teve.’”) como forma de calibrar o áudio.

De acordo com o artigo, “A interface da entrevista exibia o avatar do sprite 2-D representando o agente entrevistador no centro, com o avatar do participante mostrado na parte inferior, caminhando em direção a uma trave para indicar o progresso. Quando o agente entrevistador de IA estava falando, isso foi sinalizado por uma animação pulsante do círculo central com o avatar do entrevistador.”

As entrevistas de duas horas, em média, produziram transcrições com 6.491 palavras. Fez perguntas sobre raça, género, política, rendimento, utilização das redes sociais, o stress dos seus empregos e a composição das suas famílias. Os pesquisadores publicaram o roteiro da entrevista e as perguntas feitas pela IA.

Essas transcrições, com menos de 10.000 palavras cada, foram então inseridas em outro LLM que os pesquisadores usaram para criar agentes geradores destinados a replicar os participantes. Em seguida, os pesquisadores submeteram os participantes e os clones de IA a mais perguntas e jogos econômicos para ver como eles se compariam. “Quando um agente é questionado, toda a transcrição da entrevista é injetada no prompt do modelo, instruindo o modelo a imitar a pessoa com base nos dados da entrevista”, disse o jornal.

Esta parte do processo foi o mais controlada possível. Os pesquisadores usaram o General Social Survey (GSS) e o Big Five Personality Inventory (BFI) para testar até que ponto os LLMs correspondiam à sua inspiração. Em seguida, conduziu os participantes e os LLMs através de cinco jogos econômicos para ver como eles se compariam.

Os resultados foram mistos. Os agentes de IA responderam cerca de 85% das perguntas da mesma forma que os participantes do mundo real no GSS. Eles atingiram 80% no BFI. Contudo, os números despencaram quando os agentes começaram a fazer jogos económicos. Os pesquisadores ofereceram aos participantes da vida real prêmios em dinheiro para jogar jogos como o Dilema do Prisioneiro e o Jogo do Ditador.

No Dilema do Prisioneiro, os participantes podem optar por trabalhar juntos e ambos terem sucesso ou atrapalhar o parceiro para ter a chance de ganhar muito. No Jogo do Ditador, os participantes devem escolher como alocar recursos aos outros participantes. Os participantes da vida real ganharam dinheiro acima dos US$ 60 originais por jogar estes jogos.

Confrontados com estes jogos económicos, os clones de IA dos humanos também não replicaram os seus homólogos do mundo real. “Em média, os agentes geradores alcançaram uma correlação normalizada de 0,66”, ou cerca de 60%.

Vale a pena ler o documento inteiro se você estiver interessado em saber como os acadêmicos pensam sobre os agentes de IA e o público. Não demorou muito para que os pesquisadores resumissem a personalidade de um ser humano em um LLM que se comportasse de maneira semelhante. Com tempo e energia, eles provavelmente poderão

aproximar os dois.

Isso é preocupante para mim. Não porque não queira ver o inefável espírito humano reduzido a uma planilha, mas porque sei que esse tipo de tecnologia será usada para o mal. Já vimos LLMs mais estúpidos treinados em gravações públicas enganando avós para que cedessem informações bancárias a um parente de IA após um rápido telefonema. O que acontece quando essas máquinas possuem um script? O que acontece quando eles têm acesso a personalidades criadas com base em atividades nas redes sociais e outras informações disponíveis publicamente?

O que acontece quando uma empresa ou um político decide que o público quer e precisa de algo com base não na sua vontade expressa, mas numa aproximação dela?