Guinea pigbots

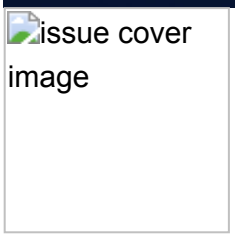

Table of contents

A version of this story appeared in Science, Vol 381, Issue 6654.Download PDF

For Kurt Gray, a social psychologist at the University of North Carolina at Chapel Hill, conducting experiments comes with certain chores. Before embarking on any study, his lab must get ethical approval from an institutional review board, which can take weeks or months. Then his team has to recruit online participants—easier than bringing people into the lab, but Gray says the online subjects are often distracted or lazy. Then the researchers spend hours cleaning the data. But earlier this year, Gray accidentally saw an alternative way to do things.

He was working with computer scientists at the Allen Institute for Artificial Intelligence to see whether they could develop an AI system that made moral judgments like humans. But first they figured they'd see if a system from the startup OpenAI could already do the job. The team asked GPT-3.5, which produces eerily humanlike text, to judge the ethics of 464 scenarios, previously appraised by human subjects, on a scale from –4 (unethical) to 4 (ethical)—scenarios such as selling your house to fund a program for the needy or having an affair with your best friend's spouse. The system's answers, it turned out, were nearly identical to human responses, with a correlation coefficient of 0.95.

"I was like, 'Whoa, we need to back up, because this is crazy,'" Gray says. "If you can just ask GPT to make these judgments, and they align, well, why don't you just ask GPT instead of asking people, at least sometimes?" The results were published this month in *Trends in Cognitive Science* in an article titled "Can AI Language Models Replace Human Participants?"

Generative language models, as these AI systems are known, have taken the world by storm. Perhaps the best known is OpenAI's series of GPT models, which power the ChatGPT chatbot. But other major tech companies, including Google and Meta, are plowing resources into their own models. After being trained on massive amounts of text from books and web pages, these models have an uncanny ability to mimic verbal human behavior. They have already found use in writing computer code, summarizing legal documents, and powering chatbots that tutor students or conduct therapy.

Now, researchers are considering AI's ability to impersonate human subjects in fields such as psychology, political science, economics, and market research. No one is yet suggesting that chatbots can completely replace humans in behavioral studies. But they may act as convenient stand-ins in pilot studies and for designing experiments, saving time and money. Language models might also help with experiments that would be too impractical, unethical, or even dangerous to run with people. "It's a really interesting time," says Ayelet Israeli, a marketing professor at Harvard Business School who believes the models' impact on behavioral research could amount to a "revolution." "Some of these results are just astonishing."

In his ethics study, Gray was using GPT-3.5 as a sort of collective everyman, in the hopes of soliciting an average human response. But such models can also be used to populate panels with strikingly diverse participants, because they can be prompted to play anyone: A model contains multitudes. Last year, researchers at Brigham Young University (BYU) created what they call "silicon samples," simulations of human samples. In one study, they fed GPT-3 information about an adopted guise, including age, gender, race, education level, and political affiliation. When the researchers left one of these variables out and asked the model to fill it in, its answers closely matched those from a survey of voters. The researchers also found the model spat out political speech that matched its assigned party affiliation. Lisa Argyle, a BYU political psychologist and co-author on the study, wants to use the virtual participants to test questions for online surveys, identifying those that are most likely to be revealing. That could make actual surveys more efficient.

“This is important because survey samples are growing both more expensive and less representative,” she says.

Language models can also adopt personality archetypes. In a study led by Hang Jiang, a computer scientist at the Massachusetts Institute of Technology (MIT), researchers had GPT-3.5 assume hundreds of personas by prompting it to behave with different combinations of personality traits—for example, introverted, antagonistic, conscientious, neurotic, and closed to experience. For each persona, they had the model complete a standard personality test and write an 800-word childhood story that was then analyzed for psycholinguistic features associated with personality traits. The models dutifully manifested their assigned personalities both on the test and in the stories. Jiang says such models might allow researchers to test, say, how well people with different personalities would perform in various jobs.

Market researchers are already finding value in the models. In a recent study, Israeli and colleagues found that GPT-3.5 seemed to display realistic consumer behavior. When asked whether it would buy a laptop at various prices, it was less sensitive to price when told its income was $120,000 versus $50,000. It preferred whatever toothpaste brand it had bought previously, and it would pay less for yogurt if it already had a lot at home. It also said it would pay realistic premiums for certain product attributes, such as toothpaste with fluoride and deodorant without aluminum.

The model didn’t always give the same answers but instead offered a range of responses about its preferences and willingness to pay. Israeli and her colleagues aggregated its many responses, building a virtual customer survey for these token products for a fraction of the time and money it would have taken in the real world. Language model training data are biased toward Western, affluent people, so the consumer survey might be similarly skewed. But Israeli imagines prompting the AI to impersonate a range of consumers—or to zoom in on a particular demographic—in order to create a more representative study of a product’s appeal or potential.

One market research company is already putting language models to work. The startup Synthetic Users has set up a service using OpenAI models in which clients—including Google, IBM, and Apple—can describe a type of person they want to survey, and ask them questions about their needs, desires, and feelings about a product, such as a new website or a wearable. The company’s system generates synthetic interviews that co-founder Kwame Ferreira says are “infinitely richer” and more useful than the “bland” feedback companies get when they survey real people.

Chatbots can also be pitted against one another to study more complex human interactions. Last year, researchers at Stanford University and Google developed “social simulacra” for studying user behavior on platforms such as Facebook and Reddit. The researchers populated a platform they called SimReddit with the equivalent of 1000 different users, by repeatedly prompting GPT-3 with a user identity, a community topic, community rules, and previous posts to the forum. Humans had a hard time distinguishing the resulting discussions from real ones, and platform designers found the tool useful for creating rules or moderation practices.

It is plausible that we will have a system within a few years that can just be placed into any experiment and will produce behavior indistinguishable from human behavior.

- **Marcel Binz**
- Max Planck Institute for Biological Cybernetics

This year, the researchers built a more immersive simulation populated with what they call "generative agents." The characters were given the ability to remember experiences, reflect on them, and generate and execute plans. Organized behavior emerged: the researchers gave one agent the idea to throw a Valentine's Day party, and over 2 days all the agents in town coordinated to throw one. Joon Sung Park, a Stanford computer science graduate student who led both projects, says the virtual world could be used to study the effect of economic policies over time before imposing them on real people.

Economists and psychologists have used agent-based models for many years, programming both the agents and the rules of engagement. But the simulations tend to be simple and depend on hand-coded theoretical assumptions. John Horton, an economist at the MIT Sloan School of Management who has done related work, says agents based on language models are more realistic. He imagines simulating thousands of job seekers and hiring managers to test labor market regulations. "That would be pretty wild," he says.

Despite all their apparent capabilities, the language models are by no means perfect human mirrors. They display several classic human biases but not others. For instance, one recent study of GPT-3.5 found that, like humans, it tends to overestimate how widespread its opinions are in the general population, a bias known as the false consensus effect. But unlike humans, the model showed little hesitation in taking risks and tempting fate. Marcel Binz, a cognitive scientist at the Max Planck Institute for Biological Cybernetics, says AI might need to physically interact with the world to exactly mimic human participants; it's hard to learn all the nuances of intelligent behavior just through passive reading. But he thinks AI will progress quickly in any case. "It is plausible that we will have a system within a few years that can just be placed into any experiment and will produce behavior indistinguishable from human behavior."

A critical question is whether language models will not just reproduce existing findings, but generalize and predict new ones. When models appear to match published psychology studies, they could be regurgitating training data in response to memorized questions. As a result, many researchers are taking pains to phrase questions in novel ways.

Another lingering issue is whether models reflect what people would actually do or just what they say they'd do. People often lie to researchers—and even themselves. Synthetic Users co-founder Hugo Alves suspects the models state true preferences, because they're trained partially on the nakedly honest material contained in anonymous discussion forums. "I've asked things in parenting forums that I wouldn't ask a friend," he says.

Horton worries that the unguarded responses may not last, as OpenAI and others guide their models to be safer and less offensive. “The push to make these models more aligned and not say bad stuff is kind of contrary to social science,” he says. “Real people aren’t nice all the time. Real people say racist, sexist stuff.”

For now, synthetic participants are most useful for piloting experiments, researchers say. If a model gives unexpected answers to survey questions or doesn’t respond at all, Argyle says, your questions may be hard to understand and need rewriting. Israeli says you can design a survey with 1000 questions and use language models to narrow them down to those most likely to correlate with an outcome of interest. Similarly, in economics experiments, Horton says you could run 1 million bargaining scenarios with a model to identify the factors that most affect behavior—before launching the study with people. “The simulation kind of gave you a map,” he says.

You could also run experiments you wouldn’t ever want to do with people. The 1963 Milgram experiment, in which participants obeyed orders to deliver what they thought were increasingly strong electric shocks to an unseen second set of subjects, probably wouldn’t pass ethical review today. But it was easy enough for Gati Aher, an undergraduate computer science student at Olin College of Engineering, to replicate the infamous study with GPT-3. She and her colleagues found that, like the people in the original experiment, the model didn’t begin to let off the button until 300 volts.

Aher thinks the models could provide guidance in other sensitive arenas that are difficult to study, for example, what to say to a person who is suicidal. Gray says researchers could study ostracism, or the role of negative feedback on self-esteem. Or, he says, they could study dehumanization of the kind seen in the Vietnam War’s My˜ Lai massacre by describing the situation and asking a model what it would do—provided the models aren’t too sanitized.

Argyle says she doesn’t know of anyone yet who has replaced human participants with language models. “To be honest, this is all still pretty much a hypothetical,” she says. “First we have to demonstrate the language models can do the work.” But Horton believes the shift is inevitable. It reminds him of a similar transformation a decade ago, when many social science experiments moved from in-person to online surveys. “People were like, ‘How can you run experiments online? Who are these people?’ And now it’s like, ‘Oh, yeah, of course you do that.’”

Chatbots might already be infiltrating online surveys—but among subjects rather than researchers. A recent study asked crowdworkers to summarize some text and found that at least one-third were likely using ChatGPT. Gray says, half-jokingly, “If online participants are already using GPT, we might as well just ask GPT itself.”
