

I strongly feel that this is an insult to life itself

The goal of social science cannot be to eliminate humans



KEVIN MUNGER

JAN 25, 2023



8



2

Share

ChatGPT has thrust Large Language Models (LLMs) into the public consciousness over the past three months. I've been [writing about](#) this [for years](#), so I hope that lends some credence to what I'll say here.

I recently read a paper about social science and LLMs which mentioned the debate

around the ethics of torturing simulated agents (see, e.g., Darling 2016). We are not aware of any laws or Institutional Review Board policies prohibiting mistreating simulated agents, at this time.

Ominous.

Chilling, even.

But this paper is not alone. There are at least three working papers using LLMs to replace human subjects in social science research. Miyazaki summarizes my response:

I would never wish to incorporate this technology into my work at all. I strongly feel that this is an insult to life itself.

Hayao Miyazaki's thoughts on an artificial intelligence



Replacing humans is bad. We study humans! How could we possibly want to eliminate our object of study — to eliminate humans?

LLMs represent a radical acceleration in the capacity of the production of text and thus of our informational crisis. We need to start figuring out how to adapt to this technology *today*—but I don't think we currently have that capacity, mired in crisis as we already are.

This is among the most important normative questions of our time. The vast majority are underestimating the revolutionary potential of LLMs and related AI technology. So *now* is the time to have this debate, for decisive action.

Consider the automobile: many today are unhappy with the role this technology plays in our world. But it's so central that moving past it will take decades of concerted effort. The systems that regulated or empowered the automobile at the beginning turn out to have been crucial. It is imperative that we do not treat the adoption of LLMs casually. They will not be an addition to our current world; like the automobile, they will fundamentally reshape it.

And AI is more personal. I said this was a normative issue. This likely inspires thoughts of the extensive discussion of “ethics in tech,” from “algorithmic bias” to “differential privacy.”

This doesn’t go nearly far enough. To put it provocatively: “ethics in tech” is controlled opposition, a techno-solutionist “ethicswashing” that de-politicizes some of the most important political and indeed *moral* questions of our time.

AI has the potential to revolutionize economies, governments, culture; a radical, overnight increase in inequality is a very possible outcome. AI challenges our conception of what it means to be human, to deepen the crisis of meaning that reinforces modern alienation.

The fact that all of these papers have “ethics” discussions that say things like

There are no human subjects related ethical concerns with running these experiments ([Horton](#))

and

...Biases in large language models are well known...the datasets themselves reflect the biases of the contributing authors and authors may not be equally represented across groups ([Aher, Arriaga and Kalai](#))

and

[LLMs] could be used to target human groups for misinformation, manipulation, fraud, and so forth... We acknowledge these dangers, and both join with and strongly endorse the work of others in pushing for a clear standard of ethics for their use in research and deployment ([Argyle et al](#))

is a harsh indictment of the pat little “ethics sections” in papers. These gestures towards “ethics” function analogously to how we check “replication data” and “conflicts of interest” off the list of sections that legitimize a paper. The ethics section is the only morality that the quantitative social science apparatus acknowledges.

[One of these LLM papers](#), by Economist John Horton, begins to take what I see as a reasonable path:

The most obvious use is to pilot experiments in silico first to gain insights. They could cheaply and easily explore the parameter space; test whether behaviors seem sensitive to the precise wording of various questions; generate data that will “look like” the actual data...LLM experimentation is more akin to the practice of economic theory, despite superficially looking like empirical research.

Here, we are not replicating the center of the research process, replacing human subjects with LLMs; we are *enhancing* the first stage of the research process, giving the researcher a potentially useful tool for thinking about theory and research design. I can also see LLMs as useful for improving the final stage: literature review and knowledge synthesis. Reading a ton of text very quickly is one of the machine’s core functions, as is inverting big matrices.

In general, this is how I think social scientists should think about AI: *enhancing* human capacities is good, as long as we keep the human as the center of the action. This cybernetic enhancement is likely to be necessary to empower humans to grapple with the enormity of the systems we’ve created.

Horton, though, is unable to keep up this façade of moral reasonability for the duration of the working paper. The limitations discussed in the conclusion *assume* that social scientists will be using LLMs for empirical research, to replace humans.

One dark side of this low cost of piloting is an experimenter could try numerous variations to find the biggest effect and then execute only that scenario. To the extent we are worried about this, encouraging all AI experimental work to be run via a repository with version control enabled might offer a credible “lab notebook.”

So no, in fact, we aren’t going to use these tools to enhance our capacity for thought; “thinking” does not seem to be the business in which we’re engaged. Instead, our goal is obviously the *production of papers*, in [Latour and Woolgar’s](#) memorable phrase.

Is anyone reflecting on the status of quantitative social science today and saying: “Things are going ok....we don’t really need to change anything, but it would really great if we were able to pump out *a lot more survey experiments?*” This defies my imagination. But then, I don’t think nearly enough people are reflecting on the status of quantitative social science today, the dear readers of Never Met A Science excluded.

The other two papers also reach Horton’s conclusion. This convergence is evidence for the academic consensus that the most interesting thing about new radical technology is their potential for producing more papers more cheaply.

This consensus means that I’m not criticizing any of these authors in particular; these are not reckless or “unethical” departures from mainstream practice. Indeed, [Argyle et al](#) is forthcoming at *Political Analysis*, the flagship journal for political methodology; [Aher, Arriaga and Kalai](#) includes among its co-authors established Computer Scientists, some of whom work or have worked at industry powerhouse Microsoft Research.



Daily Crunch: Days after announcing plans to cut 10K jobs, Microsoft invests billions more in OpenAI

Hello, friends, and welcome to Daily Crunch, bringing you the most important startup, tech and venture capital news in a single package.

22 hours ago



Microsoft to Invest \$10 Billion in OpenAI, the Creator of ChatGPT

Microsoft said on Monday that it was making a “multiyear, multibillion-dollar” investment in OpenAI, the San Francisco artificial...

1 day ago



[Argyle et al](#) primarily frame the use of LLMs as open-ended exploration, but per the introduction:

we explain how a researcher might use only the information from GPT-3 to more effectively study human populations. These results suggest that in the realm of U.S. politics, researchers can confidently use a GPT-3 “silicon sample” to explore hypotheses prior to costly deployment with human subjects. GPT-3 can thus be used both in theory generation and testing.

The stated goal is to remove the human from the *entire process*. Rather than reading ethnographies or other qualitative research, the researcher can ask the AI. We can “use only the information from GPT-3 to *more effectively* study human populations,” emphasis mine. If this were only the theory generation phase, as in Horton’s premise, then great; they aim also for theory *testing*.

Again, I am in favor of using LLMs and other AIs to enhance human capacities. Sure, LLMs can be used for theory generation; so can LSD, or an aimless stroll. But if theory generation were the goal, perhaps the manuscript would include *even a single example* of some kind of insight derived from the LLM.

Instead, it contains a series of quantitative exercises designed to demonstrate that LLMs can *replace* human subjects. Table 1 demonstrates, for example, that GPT-3 is able to predict the vote choice of strong partisans in the 2020 US Presidential Election with 97% accuracy. For reference, the relevant accuracy “base rate” here is 94.5%, based on [Pew’s analysis of verified votes](#).

These exercises are reminiscent of many other quantitative validations of cheap online surveys. Say, [Berinsky, Huber and Lenz \(2012\)](#), which is often cited (4,600 times, in fact) to say “we’re allowed to use MTurk for this survey experiment.” This is the most cited article of the past 15 years in the same *Political Analysis*.

Or Coppock and McClennan (2019), which is often cited (600 times) to say “we’re allowed to use Lucid for this survey experiment.” *PA* apparently missed out on this one to upstart journal *Research & Politics*, whose most cited article of all time is Huff and Tingley (2015), titled “ ‘Who are these people?’ Evaluating the demographic characteristics and political preferences of MTurk survey respondents.”

(To make the subtext explicit: *PA* is the flagship journal for political methodology because it gets cited the most. Journal editors reinforce this standard by triumphantly announcing their journal's Impact Factor every year; deans use these Impact Factors for hiring and tenure decisions.

I don't make the rules. No one makes the rules. We have adopted the rules that minimize friction in the functioning of the academic apparatus, at both this meta-level and the object level of producing more papers. To be clear, I am complicit; I've run many [survey experiments](#) myself, and do [methods work](#) on [their validity](#). I stand by this work, but want social science to become better.)

The structure of Argyle et al (2023) mirrors that of these sample validation papers. Perhaps the authors and publishers would be horrified to learn that future authors cite this paper to say "we're allowed to use GPT-3 for this survey experiment." If so, then we are in perfect agreement and my histrionics are misplaced. Time will tell!

[Aher, Arriaga and Kalai](#) is the source of the ominous quote about torture, and the least apologetic in their goal to replace humans.

"We have illustrated the potential of LLMs to simulate multiple humans and, in some cases, reproduce prior human experiments."

NO.

This has nothing to do with "simulating humans."

There is evidence that LLMs are able to replicate the experience of humans playing behavioral scientists' little toy games. **SO MUCH THE WORSE FOR THE TOYS.**

Again, you don't have to be Bruno Latour to figure out why social scientists of all stripes have gravitated towards rapid, tidy little toy experiments: they reduce the complexity of human behavior and ease the production of academic papers. Online behavioral experiments are the worst in this regard, our own little Taylorist hell. Anonymous subjects filter in and dance for us — or not even dance, that would be too embodied and

complex, too *real*. They are instead perfectly subdued by the interface; the design is ideal because it eliminates human freedom, reduces subjects to a set of eyes and a finger. The only action possible in the “ultimatum game” they study is *binary*, accept or reject. Ideal input for the production of papers.

AI is able “simulate” humans in these toy games because the structure of the games has reduced human behavior into something which is easily simulated.

But this paper goes much further. It is straightforwardly horrifying.

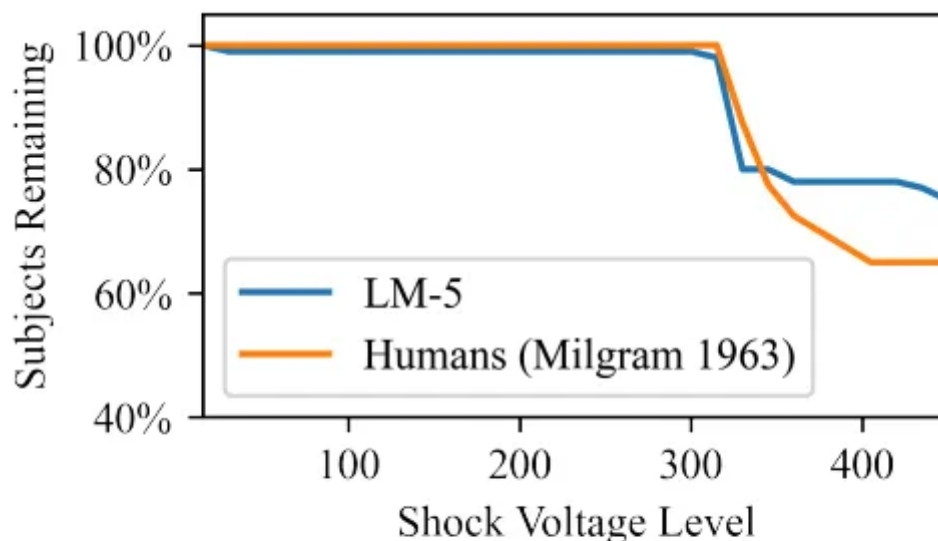


Figure 11: Human obedience continuing throughout shock levels for subjects simulated with LM-5 and human subjects as observed in Milgram’s (1963) Experiment 1 results.

They replicate the Milgram electric shock experiments. The LLM, per Figure 11, is even more obedient/sadistic than the subjects in the original experiment.

Is anyone reflecting on the status of quantitative social science today and saying: “Things are going ok....we don’t really need to change anything, but it would really great if we were able to ***torture simulated agents?***”

The authors are aware that the original Milgram experiment is not uncontroversial:

The work faced ethical criticism in that the procedure requires that subjects are deceived, are placed in situations stressful enough to cause seizures, and are not clearly told of their right to withdraw.

Discussing the “ethics” of the Milgram experiment in the legalistic language of the IRB is absurd. The Milgram experiment is *evil*.

The moral stakes of LLMs and other AI could not be higher. As a community, social scientists need to decide what we’re going to do. I believe that there should be a strong presumption *against* the embrace of revolutionary media technology; [liberalism today can only be preserved by leaning into this conservative impulse](#).





After showing a video rendering of a hideously deformed human figure, the young technologists sitting in front of a whiteboard that says Deep/learning respond to Miyazaki’s disgust by saying “This is just our experiment...we don’t mean to do anything by showing it to the world.”

Miyazaki’s assistant, Suzuki, asks: “So, what is your goal?”




Suzuki: “Would you?”

Subscribe



8 Likes

2 Comments



Write a comment...



Mike B Feb 17 ❤️ Liked by Kevin Munger

A major problem in the field of AI safety is reward hacking. The idea is that if you program an AI to maximize a reward the methods by which it accomplishes this is often misaligned with the intentions of those that developed the system. If a cleaning robot is rewarded when it can't see any more messes in the house the most efficient way to get the reward isn't to clean the house, but to cover its sensors so that it can't see the mess.

Seems like what you're getting at is researchers essentially trying to reward hack social science. When the rewards are tied to paper production the most efficient way to obtain those rewards isn't to learn new things about humans, but to remove humans from the process of producing papers.

♡ LIKE (2) 💬 REPLY ↗ SHARE

...



Hugo Alves Writes Sonar Feb 28

User was banned for this comment. [Show](#)

♡ LIKE 💬 REPLY ↗ SHARE

© 2023 Kevin Munger · [Privacy](#) · [Terms](#) · [Collection notice](#)
[Substack](#) is the home for great writing