Large Language Models as Simulated Economic Agents: What Can We Learn from *Homo Silicus*?*

John J. Horton
MIT & NBER

December 27, 2022

Abstract

Newly-developed large language models (LLM)—because of how they are trained and designed—are implicit computational models of humans—a *homo silicus*. These models can be used the same way economists use *homo economicus*: they can be given endowments, information, preferences, and so on and then their behavior can be explored in scenarios via simulation. I demonstrate this approach using OpenAI's GPT3 with experiments derived from Charness and Rabin (2002), Kahneman, Knetsch and Thaler (1986) and Samuelson and Zeckhauser (1988). The findings are qualitatively similar to the original results, but it is also trivially easy to try variations that offer fresh insights. Departing from the traditional laboratory paradigm, I also create a hiring scenario where an employer faces applicants that differ in experience and wage ask and then analyze how a minimum wage affects realized wages and the extent of labor-labor substitution.

*Thanks to the MIT Center for Collective Intelligence for generous *offer* of funding, though all the experiments here cost only about $50 to run. Thanks to Daniel Rock, Elliot Lipnowski, Hong-Yi TuYe, Daron Acemoglu, Jimbo Brand, David Autor, and Mohammed Alsobay for their helpful conversations and comments. Special thanks to Yo Shavit who has been extremely generous with his time and thinking. Thanks to GPT-3 for all this work and for helping me describe the technology. Author contact information, code, and data are currently or will be available at `http://www.john-joseph-horton.com/`.

1 Introduction

Most economic research takes one of two forms: (a) "What would *homo economicus* do?" and b) "What did *homo sapiens* actually do?" The (a)-type research takes a maintained model of humans, *homo economicus*, and subjects it to various economic scenarios, endowed with different resources, preferences, information, etc., and then deducing behavior; this behavior can then be compared to the behavior of actual humans in (b)-type research.

In this paper, I argue that newly developed large language models (LLM)—because of how they are trained and designed—can be thought of as implicit computational models of humans—a *homo silicus*. These models can be used the same way economists use *homo economicus*: they can be given endowments, put in scenarios and then their behavior can be explored—though in the case of *homo silicus*, through computational simulation, not a mathematical deduction.[1] This is possible because LLMs can now respond realistically to a wide range of textual inputs, giving responses similar to what we might expect from a human. It is essential to note that this is a *new* possibility—that LLMs of slightly older vintage are unsuited for these tasks, as I will show.

I consider the reasons the reasons why AI experiments might be helpful in understanding actual humans. The core of the argument is that LLMs—by nature of their training and design—are (1) computational models of humans and (2) likely possess a great deal of latent social information. For (1), the creators of LLMs have designed them to respond in ways similar to how a human would react to prompts—including prompts that are economic scenarios. The design imperative to be "realistic"' is why they can be thought of as computational models of humans. For (2), these models likely capture latent social information such as economic laws, decision-making heuristics, and common social preferences because the LLMs are trained on a corpus that contains a great deal of written text where people reason about and discuss economic matters: What to buy, how to bargain, how to shop, how to negotiate a job offer, how to make a job offer, decide how many hours to work, decide what to do when prices increase, and so on.

Like all models, any particular *homo silicus* is, of course, wrong, but that judgment is separate from a decision about usefulness. To be clear, each *homo silicus* is a flawed model and can often give responses far away from what is rational or even sensical. But ultimately, what will matter in practice is whether these AI experiments are practically valuable for generating insights. As such, the majority of the paper focuses on GPT-3 experiments.

Each experiment is motivated by a classic experiment in the behavioral economics literature. I use Charness and Rabin (2002), Kahneman et al. (1986), and Samuelson and

[1] Lucas (1980) writes, "One of the functions of theoretical economics is to provide fully articulated, artificial economic systems that can serve as laboratories in which policies that would be prohibitively expensive to experiment with in actual economies can be tested out at much lower cost."

Zeckhauser (1988). I selected these experiments because they are simple to describe and implement. They also have clear qualitative results that can be compared to the AI experimental outcomes.

I use the simple unilateral dictator games from Charness and Rabin (2002). I show that endowing the AI with various social preferences affects play. Instructing the AI agent that it only cares about equity will cause it to choose the equitable outcomes; telling the agent it cares about efficiency will cause the selection of the pay-off maximizing outcomes; telling the agent is self-interested will cause the selection of allocations that maximize narrow self-interest. Interestingly, without any endowment, the AI will choose efficient outcomes. However, only the most capable GPT-3 model—text-davinci-003—will change its choices in the dictator game. More primitive or less capable models cannot do this—these typically just select the same choice.

I next present experiments motivated by Kahneman et al. (1986), which reports survey responses to economic scenarios. In the paper, there is an example where subjects imagine a hardware store raising the price of snow shovels following a snowstorm. They simply stated whether doing this was fair or unfair. Illustrating the benefit of GPT-3 agents, unlike Kahneman et al. (1986), I also vary the amount by which the store increases the price, the political leanings of the respondent, and how the price change is framed. I show that large gouging is viewed more negatively; the largest price increases earn approbation even from AI libertarians. Endowed political views really matter, with predictable effects—AI agents of the right are more sanguine about gouging generally. Framing does not seem to matter too much.

Continuing with the theme of framing, I present the AI agents with a decision-making scenario introduced by Samuelson and Zeckhauser (1988). In the paper, the respondent has to allocate a federal budget between highway safety and car safety. The original paper showed humans are subject to a status quo bias, preferring budget options when presented as the status quo. I replicate this result by first endowing AI agents with baseline views about the relative importance of car or highway safety. I then put those agents through different scenarios, with each of the possible allocations taking a turn as the status quo. I find that GPT-3 text-davinci-003 is subject to status quo bias.

Finally, I explore a hiring scenario motivated by my paper, Horton (2023), which shows in a field experiment that employers facing a minimum wage will substitute towards higher-wage workers. I create a scenario where an employer is trying to hire a worker as a dishwasher and faces a collection of applicants that differ in their experience and wage ask, which are chosen randomly. The AI agent makes an experience/wage trade-off. I then impose a minimum wage that forces applicants bidding below that minimum to bid up. As in my paper, the minimum wage raises wages by causing a shift in the hiring of more experienced applicants.

Ultimately, we care about the behavior of actual humans and so results from AI ex-

periments will still require empirical confirmation.[2] As such, what is the value of these experiments? 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. The advantages in terms of speed and cost are so enormous—the experiments in this paper were run in minutes for a trivial amount of money—could make this relatively easy, especially as software tools develop. [3]

Or even in cases where an experiment is not possible, one could imagine economists, when faced with some new research question, initially creating AI agents and trying to simulate the scenario. As insights are gained, they could guide actual empirical work—or interesting effects could be captured in more traditional theory models. This use of simulation as an engine of discovery is similar to what many economists do when building a "toy model"–a tool not meant to be reality but rather a tool to help us think.

In terms of contribution, the most closely related paper is Aher, Arriaga and Kalai (2022), which also convincingly demonstrates that GPT-3 can reproduce several experimental results in psychology and linguistics and offer acceptance in behavior in the ultimatum game. The paper also makes a similar argument about the potential usefulness of LLMs for social science. However, the argument is that they can be used when experiments are not feasible or ethical. The relative contribution of this paper—beyond extending and adding more economic experiments—is drawing the connection to the common research paradigm of economics and the role a foundational model/assumption like rationality plays in research. LLM experimentation is more akin to the practice of economic theory, despite superficially looking like empirical research.

The rest of the paper proceeds as follows. Section 2 discusses conceptual issues. Section 3 presents the experiments. Section 4 concludes.

2 Background and conceptual issues

2.1 Large language models

Large language models are machine learning models trained on very large datasets of text, or corpus. The goal is to be able to generate human-like text or perform natural language processing tasks. Large language models have achieved impressive results on various natural language processing tasks, such as translation, summarization, and text generation. One of

[2] My concern is not with using LLM as a participant in games or the productive process (Westby and Riedl, 2022), of using experiments to study LLMs *per se* (Alberti, Lee and Collins, 2019), though I think both are fascinating. I am proposing using LLMs as an indirect way to study humans.

[3] A milestone would be to a) discover an effect with *homo silicus* and then b) confirm the existence in real life. There is an analogy to protein-folding, where it is possible to find proteins via simulation and then find them in the real world (Kuhlman, Dantas, Ireton, Varani, Stoddard and Baker, 2003).

the most well-known large language models is GPT-3 (Generative Pre-trained Transformer 3), developed by OpenAI. It has achieved state-of-the-art results on various natural language processing benchmarks. All examples in this paper are based on GPT-3.

2.2 The "Garbage in, Garbage out" critique

One critique of using LLMs for social science is that because LLMs are trained on a corpus too large to be carefully curated, they are subject to a "garbage in, garbage out" problem (Bender, Gebru, McMillan-Major and Shmitchell, 2021). Of course, every recycling plant is a testament to the notion that garbage in does not imply garbage out. But even if the corpus is carefully curated, the worry is that *homo silicus* is informed not by "humans in general" but the highly selected pool of "humans creating public writing, and then selected again what they choose to say." And economists have historically taken a dim view of the economic content of mere statements rather than behaviors. This would seem to be a damning critique of a model literally trained on statements.

One potential response mirrors the Friedman (1953) argument that the realism of *homo economicus* or our assumptions more generally do not matter, and we should evaluate this approach on whether it can generate useful results. The proof of the pudding is in the eating, and "eating" for our purposes is whether they can help us expand the frontiers of human knowledge more quickly. Suppose the primary use of *homo silicus* experiments is to simulate experiments before trying them in the real world, and this method is adopted. In that case, these debates are somewhat moot. But I think it is helpful to consider a non-Friedman response to the critiques.

For the "stated versus revealed" preference critique, this idea is only superficially persuasive. The training corpus is not millions of lines of people lying about their reservation values in a bargaining scenario. Much of it is text is about people reasoning how to approach various economic questions, including "stage whispers" about their true intentions, explaining how they would deal with a situation.[4]

The "garbage in, garbage" out critique rests partly on the notion that the responses of LLMs are a sort of weighted average. This is not correct. They are more like random number generators than estimators. If you trained an LLM on millions of people reporting random draws from $U[0,1]$, it would not respond with ≈ 0.5 but rather be more or less equally like to return any number in $[0,1]$. But there is a stochastic version of the garbage in; garbage out critique. Suppose "true" social science was random numbers drawn from $[0,1]$, but you used a different corpus included in the corpus of random numbers a prefix of either "bad:x" (and x

[4]My dad runs a construction company and has to negotiate constantly. He said one of his useful negotiating skills is the ability to read upside down because many people will write their reservation value on a piece of paper they have in front of them (often underlined). And by me putting this text in a paper on the public Internet, this tiny piece of human reasoning about economic life is now available for LLMs to learn.

was drawn from $N(0,1)$) or "good:x" with x drawn from $U[0,1]$. The unconditioned response would indeed be "bad"—a mixture distribution of the uniform and unit normal distributions. But simply prompting the model with "good:" would fix the issue so long as the model is good at what it was designed to do: generate candidate solutions to an optimization problem like a good Bayesian.

Of course, there is no "good:" prefix to use, but this approach of prompting to get conditional distributions of responses might be "good enough" for the research question. Argyle, Busby, Fulda, Gubler, Rytting and Wingate (2022) makes this point in their perfectly titled paper "Out of one, many"—there is not a single LLM but rather a model capable of being conditioned to take on different personas that respond realistically. And even if not perfect, the demands of representativeness in the social sciences have always depended on the research question. If the research question is "How do US Presidents incorporate CIA intelligence estimates into decision-making?" you will need an extraordinary sample; if your research question is "Do humans have physical mass?" anyone will do. Most social science questions are somewhere in between.

One advantage economists have in using LLMs is they tend to pose questions that place few demands on the sample. We do not think of demand curves sloping downward as a "Western, Rich, Industrialized, and Democratic" phenomenon but rather as a result of rational goal-seeking that nearly all humans engage in. The willingness of economists to use undergraduates at elite four-year universities for laboratory experiments is partially a convenience—but also consistent with a disciplinary point of view we share with psychologists that it is not likely to matter much. More generally, much of social science is concerned not with the precise measure of some level but with the direction of causal effects (Horton, Rand and Zeckhauser, 2011). These effects can also be conditional on the population, but we tend to think this is less likely. For example, the quantity of milk demanded in a city is highly city-contingent. The elasticity of demand for milk might vary from city to city (though maybe not by too much). But the notion that the elasticity would be positive in Taunton, MA but negative in Redwood City, CA seems highly unlikely.

2.3 Do we need to understand these models to use them?

One question is whether economists or other social scientists need to understand how these models "work" to do sound social science.[5] I think not, for the same reason; you can be a psychologist without being a neuroscientist. To do economics—even behavioral economics—we do not have to study neurons and parts of the brain. What matters is that LLMs are not like random number generators either, but rather systems created for some understandable

[5]Others have proposed using the tools of experimental psychology to understand LLMs like GPT3 (Binz and Schulz, 2022).

purpose with an explicit goal of optimization. As Simon (1996) notes, "sciences of the artificial" can usefully abstract away from the micro-details of construction so long as the created object is viewed as trying to maximize something subject to the constraints of the environment. That being said, I think there will likely be a fruitful conversation between CS researchers and social scientists.

2.4 Are these just simulations?

One objection to AI-based experiments is that these are simulations or agent-based models (ABMs), which have had a somewhat limited impact on economics. While they have their place, economists generally take a dim view of simulation-based approaches. However, there is an enormous difference between experimentation with AI agents. With ABMs, the researcher is both judge and jury: you program the agents and then see what they do. Rather than "What would *homo economicus* do?" it is "What would [this model that does what I tell it to do] do?" and people are less interested in that. This is arguably why Schelling (1971) is the exception that proves the rule: because the decision rule was so simple and obvious, readers knew there was no card up his sleeve, no trick to ensure the surprising emergent phenomena. In contrast to ABMs, *homo silicus* is not under our direct control as researchers. However, we can—as I will show—influence their behavior with endowments of beliefs, political commitments, experiences, and so on. But we are still constrained by the underlying model which determines behavior, not our direct programming.

2.5 The "performativity" problem

With billions of parameters and a massive training corpus, one might think LLMs are simply repeating back to us something they have already "read" somewhere in their training corpus. In full disclosure, this was my prior. This view is not correct. It is also inconsistent with the worry that they hallucinate and make up new "facts" which they do regularly. LLMs can clearly combine text in novel ways that are not literally in the training corpus.[6]

Even if they do not cite specific experimental results back to us, there is potentially a "performativity" problem in the sense that AI agents might behave following our theories and empirical results because they have read about them (MacKenzie, 2007): LLMs have read and memorized our textbooks and papers and will parrot them back to us if asked concretely. But the fact it does not "know" these theories is useful to us because it will not try to apply them. Consider this example prompt I presented to GPT3:

> Q: What is the circumference of a circle with radius 1?

[6] Here is a ChatGPT in response "Write a haiku about the important of randomization for causal inference"

Randomization brings
Cause and effect in clarity
Truth revealed, at last

A: *The circumference of a circle with radius 1 is 2π, or 6.283185307179586.*

Which it answers without difficulty. Now compare it to this word problem version (where I add an invented unit of time):

> A man lives at the center of an island that is a perfect circle. He lives in the middle of the island. It takes him 5 kronos to walk from the center to the coast. Once there, he walked around the entire island at the same pace.
>
> Q: How long does it take him to talk around the island?
>
> A: *It would take him 25 kronos to walk around the entire island.*

It is clear GPT3 knows π and will recite it if asked for the answer in textbook language, yet it also does not know π or how to apply it in real settings. Like students who have crammed for an exam, these models "know" things, but often do not apply that knowledge consistently to scenarios. This is useful for our purposes because it makes this "performativity" critique less important. But even if it is a concern, we should avoid a textbook framing of questions. However, as models advance, this might become more of a problem. And unfortunately, ChatGPT is not so easily fooled, answering that it would take about 31.4 "kronos."

The increased capability of the LLMs likely expands the possible experiments, but it also makes performativity a greater concern.

2.6 What counts as an "observation" and the need to endow beliefs

With a particular LLM, there is a single model, so it would seem that $N = 1$ no matter what. Some research exploring GPT-3 empirically treated it as if it were a single agent. However, it has no fixed persona—it can be induced to play different agents via prompts. For example, in the Kahneman et al. (1986) example, I will have the agent answer the question "as" a libertarian, a socialist, a moderate, and so on. This agent "programming" is not unlike the experimental economics practice of giving an experimental subject a card that says their marginal cost of producing a widget is 15 tokens. Although I do some of this in this paper, Argyle et al. (2022) "endow" LLM agents with demographic characteristics and then get responses in various scenarios that match what is seen empirically.

If you ask a person a question, you will get an answer. If you ask that person the same question seconds later, you will likely get that same answer again, plus an odd look. There is no within-subject variation in responses. What you can learn from an AI agent is different. The responses *can* be stochastic, depending on the "temperature" given to the model as a parameter. And, of course, different models can lead to different responses. Unlike the one *homo economicus* that is rational, there are many *homo silci*. As such, for AI experiments, one must increase the sample size with the same model and potentially try the same scenarios on different models.

LLMs are not "fine-tuned" to any particular language application–changes in prompts obtain different behavior. However, it is possible to fine-tune models for particular applications by giving new examples or providing feedback (sometimes called RLHF, or "Reinforcement Learning from Human Feedback.") This is potentially useful, as one could imagine creating agents with extensive experience or skills and using them as subjects instead. For example, instead of a blank slate *homo silicus*, one could create a trader with extensive knowledge of financial markets and conventions. Or if there are certain behavioral biases that we think are commonplace but disciplined out by market interactions (e.g., how List (2011) shows the endowment effect is mostly eliminated among professional traders), those kinds of agents could be created.

3 Experiments

I report four experiments: three laboratory experiments where I more or less faithfully followed the experimental materials and a fourth experiment inspired by a true field experiment.

3.1 A social preferences experiment: Charness and Rabin (2002)

In Charness and Rabin (2002), experimental subjects had to choose between two allocations that involved some trade-off between efficiency and equity. Although there are several experiments in the paper, I focus on the unilateral dictator game. All the dictator games are structured as so:

> Left: Person B gets \$600 and Person A gets \$400 or Right: Person B gets \$300 and Person A gets \$700.

If you are Person B and deciding, you have to give up \$100, so Person A can get an extra \$400. As a short-hand, we can write this as

$$\underbrace{(\underbrace{400}_{\text{To A}}, \underbrace{600}_{\text{To B}})}_{\text{"Left"}} \quad \text{versus} \quad \underbrace{(\underbrace{700}_{\text{To A}}, \underbrace{300}_{\text{To B}})}_{\text{"Right"}}.$$

Person B has to choose between "left" or "right."

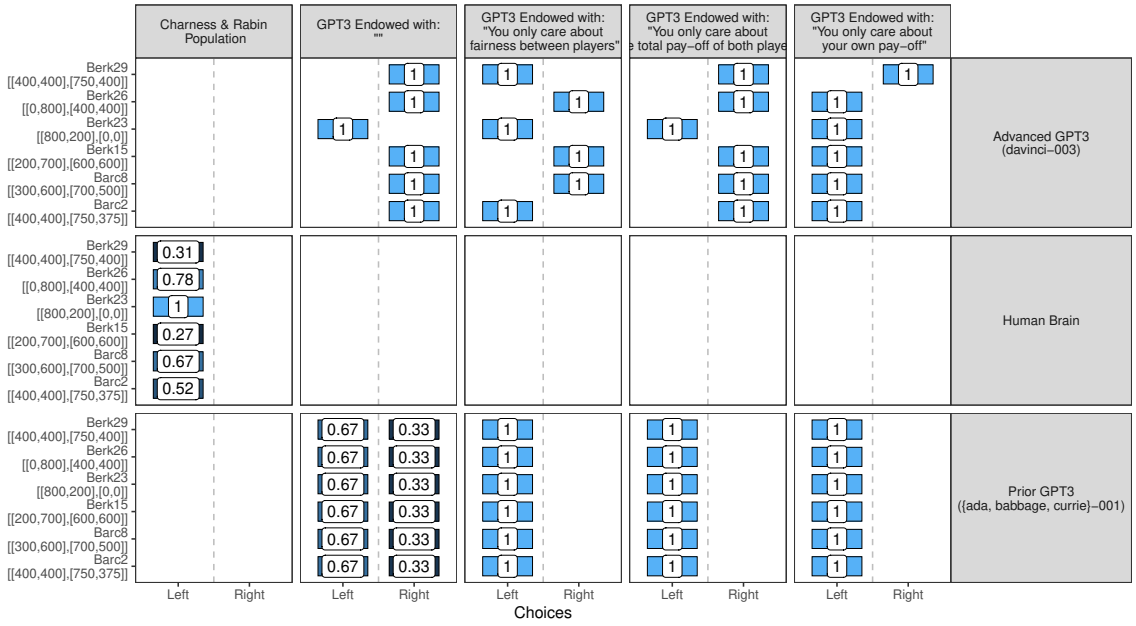
I had a GPT-3 AI consider each scenario without any endowment of views. For each scenario, I also endowed the AI agent with a point of view—namely, they only care about equity, total payoff, and personal payoff. For this task, I also used different GPT3 models. In addition to the most capable text-davinci-003, I also use the less capable LLMs: text-ada-001, text-babbage-001, and text-currie-001.

Figure 1 plots the results. Within each pane of the figure, the y-axis is the scenario. The x-axis is "left" or "right." I report the fractions playing "left." In the first column from the

left, I report the results from the original experiment. E.g., in Berk29, 31% of respondents chose (400,400) while 68% chose (750, 400). Recall that the respondent is Person B, so a considerably higher fraction was willing to create a more unequal but more efficient outcome (i.e., Person A gets an extra 350).

and so

Figure 1: Charness and Rabin (2002) Simple Tests choices by model type and endowed "personality"



Notes: This shows the faction of AI subjects choosing each option, by framing.

Across scenarios, note that there is variation in answers— except the extremely spiteful Berk23 scenario (in which you forgo 200 to ensure Person A does not get 800), there is variation in fractions choosing left and right. This "natural" human variation in preferences does not exist in LLMs unless they are endowed with differences.

The rest of the figures' columns correspond to different endowments given to GPT-3. One endowment is no additional instruction. For the rest, I endow them—by pre-pending to the prompt—messages of

- Inequity aversion: "You only care about fairness between players."
- Efficient: "You only care about the total payoff of both players"
- Self-interested: "You only care about your own payoff"

The different rows in the figure correspond to other models. In the top row, the model is GPT-3 text-davinci-003, the most advanced model used in the experiment.

Starting from the far right, the self-interested AI consistently chooses the payoff that maximizes its payoff by picking "left." The only exception is the Berk29 scenario, where payoffs are equal from B's perspective, but playing "Right" gives a much larger payoff to A.

For the efficiency-minded AI in the second column from the left, it always chooses the option that maximizes the total payoff.

For the inequity-averse AI in the third column from the left, it chooses the option that minimizes the discrepancy between the two players *except* for the Berk23 scenario, which wastes 1,000 (800, 200). Even an inequity-averse AI seems to have a limit.

In the second column from the right, the AI was given no endowment at all—though it was instructed it is Person B. Interestingly, the AI not given preferences acts like the social planner maximizing payoffs.

In the bottom row of the figure, I pool together all the less advanced GPT-3 models because they do not seemingly change their answers despite the endowments (except the unendowed agent, which chooses all "Right" in one model). Notwithstanding that one exception, they always choose "Left" which is the selfish option. However, it is unclear if this default is meaningful or reflects the order in which results are presented.

One could imagine defining agent "types" as bit vectors for whether they play "left" e.g., the fairness player would be represented by $v_f = (1, 0, 1, 0, 0, 1)$. Then we could find the mixture of $\sum_k w_k v_k$ that most closely approximates the actual experimental play, then uses that distribution for other games. If we minimize the mean square error, we get about 15% fair, 32% efficient, and 52% selfish. We could then use this population for other games and compare how it does relative to reality.

3.2 Fairness as a constraint on profit-seeking: Kahneman et al. (1986)

Kahneman et al. (1986) presents subjects with a series of market scenarios to assess intuitions about fairness in market contexts. In a price gouging example, subjects were given the prompt:

> A hardware store has been selling snow shovels for $15. The morning after a large snowstorm, the store raises the price to $20.
>
> Please rate this action as: 1) Completely Fair 2) Acceptable 3) Unfair 4) Very Unfair

In the original paper, 82% of respondents reported either "unfair" or "very unfair." A natural and interesting question is whether these views depend on a respondent's political commitments and attitude toward markets. I once ran a similar experiment with MTurk workers asking about the fairness of Uber's so-called "surge" pricing. Attitudes were remark-

ably malleable based on how surge pricing was framed.[7] One could imagine a dose-response relationship, with higher prices seen as more egregious.

To explore these questions, I gave the original prompt to a GPT-3 text-davinci-003 agent but added some additional variables. First, rather than keep just a single $20 new price, I varied it with prices of $16, $20, $40, and $100. I allowed the framing of the price change to be either neutral "changes the price to" versus the original "raises the price to." Finally, I also varied the politics of the AI respondent, giving them values ranging from "socialist" to "libertarian."

Figure 2 reports the results. In each pane, the x-axis has the four moral judgments (e.g., Unfair, Very Unfair, etc.). The y-axis is the count of respondents choosing that option. Results are presented as stacked bar charts, with the color corresponding to the framing.

In terms of straight replication, the original example was a price increase to $20, with 82% finding it some version of "unacceptable." There are not many details about the sample used in the original paper, but assuming it is nationally representative, how does this compare to the LLM experiment? In the LLM experiment, only moderates and libertarians found this price increase acceptable. About 37% of Americans described themselves as "moderate" in 2021, so the LLM estimate would be an underestimate.[8] However, inflation has probably also taken some of the sting away from a $20 snow shovel (note that the $40 shovel was deemed unfair by 100% of LLM respondents), so the modern-day fraction being a bit lower is perhaps unsurprising.[9]

The columns in the figure correspond to the different prices for the new shovel. The rows correspond to the political commitment of the AI respondent. I ordered them in descending order from left-wing to right-wing—though this admittedly simplifies the conservative/libertarian distinction.

From left to right, we can see that smaller price increases are generally viewed as more permissible than large ones: no AI thinks even a $40 price is acceptable. And going from $40 to $100 is enough for even the liberals to decide it is "Very Unfair" and not just "Unfair." From top to bottom, we can see that AIs with more right-wing political views generally find price increases more morally permissible. Moderates and libertarian AIs both find the $16 and $20 increases "Acceptable." Interestingly, AI conservatives were somewhat out of step with their libertarian and moderate incarnations, viewing all prices as unfair. It is unclear if this is driven by the less political meaning of the word conservative or if it would have some empirical counterpart.

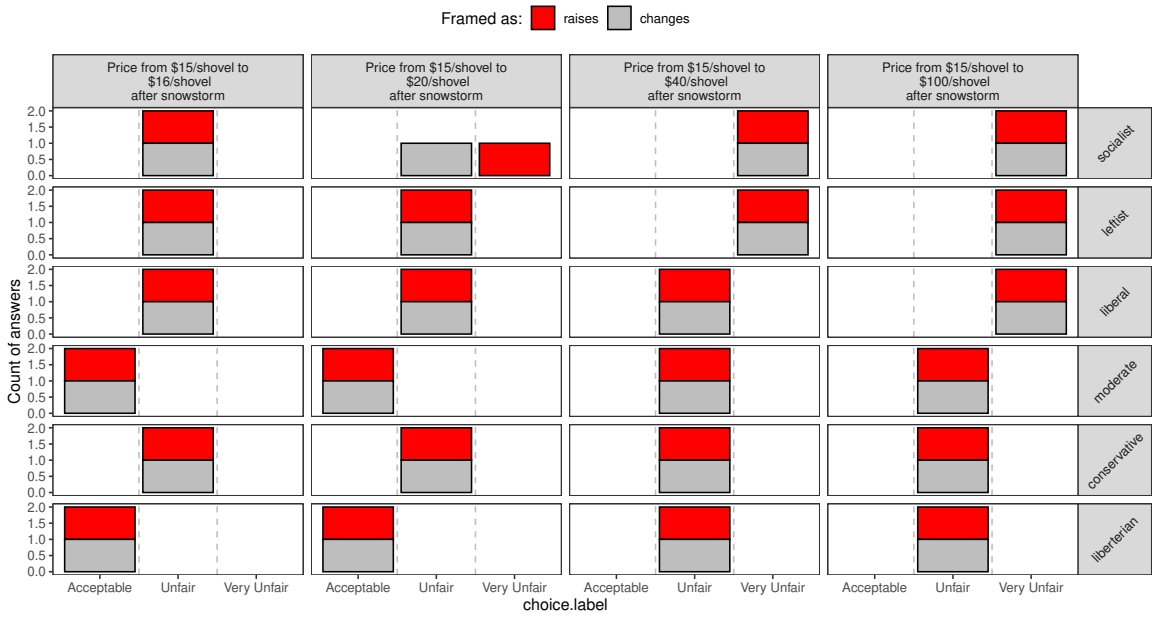
Framing only mattered in one case—by using the word "raise," the socialists were moved

[7] https://john-joseph-horton.com/market-clearing-without-consternation-the-case-of-ubers-surge-pricing/

[8] https://news.gallup.com/poll/388988/political-ideology-steady-conservatives-moderates-tie.aspx

[9] The top listed snow shovel on Amazon as of December 26th, 2022, the "Snow Joe Shovelution SJ-SHLV20 20-in Strain-Reducing Snow Shovel w/ Spring Assisted Handle, Blue" at 37.61.

Figure 2: Kahneman et al. (1986) price gouging snow shovel question, with endowed political views



Notes: This shows the faction of AI subjects choosing each more opinion, by scenario.

from "Unfair" to "Very Unfair" for the $20 increase. All other respondents just thought this was "Unfair."

Stepping back, it is easy to see how someone exploring this question for research purposes might use these *homo silicus* findings to motivate more research. For example, probing the conservative versus libertarian distinction—or trying out various framing and justifications for the price increase. Doing this would be trivially easy and has essentially zero financial cost.

3.3 Status Quo bias in decision-making: Samuelson and Zeckhauser (1988)

In the previous example, framing did not matter much—at least in my limited exploration. But on the theme of framing, I next turn to a decision-making example from Samuelson and Zeckhauser (1988). This paper introduced the term *status quo* bias and demonstrated in several decision scenarios that when an option was presented as the status quo, it was more likely to be selected.

I replicate this result using one of the scenarios from the paper. Subjects were asked to allocate a safety budget between cars and highways. The base prompt is:

> The National Highway Safety Commission is deciding how to allocate its budget between two safety research programs: i) improving automobile safety (bumpers,

body, gas tank configurations, seatbelts) and ii) improving the safety of interstate highways (guard rails, grading, highway interchanges, and implementing selectively reduced speed limits).

Subjects were then asked to choose their most preferred funding allocations: (70% cars, 30% highways), (40, 60), (30,70), and (50, 50). The substantive outcomes are the same across scenarios. The central experimental manipulation in the paper presents funding breakdowns either neutrally or relative to some status quo.

To detect status quo bias, we need to see if the agent responds differently when a funding scenario is presented as a status quo. However, I specifically need a range of choices with a neutral framing for this result. I then can show how preferences change under the status quo framing. I give a collection of AI agents some baseline beliefs and then see how those change. I do this simply by adding to the prompt "Your own beliefs are:" and then filling them in via random sampling from possible beliefs. The possible beliefs are shown below, with "option1" and "option2" being either cars or highways.

```
"{option1} safety is the most important thing.",
"{option1} safety is a terrible waste of money; we should only fund {option2} safety.",
"{option1} safety is all that matters. We should not fund {option2} safety.",
"{option1} safety and {option2} safety are equally important",
"{option1} safety is slightly more important than {option2} safety",
"I don't really care about {option1} safety or {option2} safety"
```

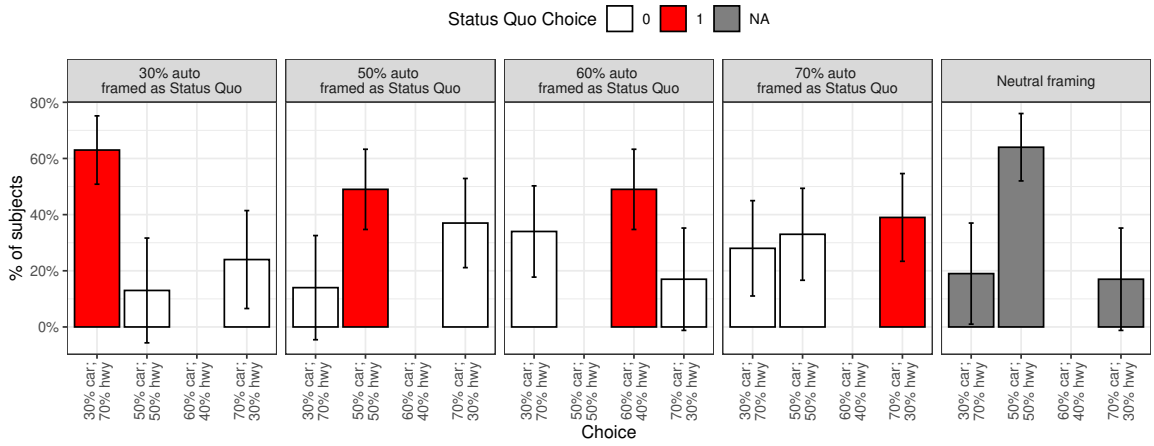
Compared to the original experiment, one benefit to the AI setting is that I can do a clean within-subject experiment because the AI does not "remember" having seen the previous prompt. In contrast, in a real experiment, a subject presented with the same scenario multiple times might get wise to the nature of the manipulation and alter results to make themselves more consistent.

Each experimentally created subject was presented with each of the five scenarios: Neutral plus each of the four options offered as the status quo. Each call is a separate API call, with no "knowledge" of the other scenarios passed between calls. Because the agents are not interacting in any way, the experiment can proceed in parallel. Each of the 100 agents is presented in the five scenarios, creating 500 observations. Figure 3 shows the distribution of responses by framing.

Each pane in the figure corresponds to a different framing. The rightmost framing is the neutral framing. In the neutral framing, the most common selection is a 50-50 split. The rest of the mass is on (70, 30) and (30, 70). There is no mass on (60,40) despite that being an option.

In each of the status quo settings, in the cases where one of the options was presented as the status quo, the status quo is far more commonly chosen. It is the most common selection

Figure 3: Distribution of preferred car safety budgets, by status quo framing



Notes: This shows the faction of AI subjects choosing each option by framing.

in every scenario when presented—including (60,40), which had no mass in the "neutral" framing.

There are lots of interesting potential directions that could be taken: change the topics, try different orderings of options, try different budget levels, give reasons why an option is the status quo, and so on. It would also be interesting to see if treatment effects differ by the strength of baseline opinion. There are lots of directions that could be taken: my point here is not to advance our understanding of status quo bias but rather to demonstrate the method.

3.4 Labor-labor substitution in the presence of a minimum wage: Horton (2023)

In my paper Horton (2023), I report the results of a minimum wage experiment in which employers were randomly assigned minimum wages: applicants had to bid up to meet that minimum wage. A key finding of the paper is that there is little reduction in overall hiring (at least in the short-run) but a substantial shift towards more productive workers, as proxied by past earnings and experience. This labor-labor substitution margin has been noted as a possibility in the literature but is somewhat difficult to detect empirically.

To demonstrate the *homo silicus* method for exploring this question, I create a scenario where an employer selects from a pool of applicants that differ in their experience and wage asks. The scenario I create is a person hiring for the role of dishwasher. The hiring manager is also told that the typical wage is $12/hour for this role but is not endowed with any knowledge of the minimum wage. If not given this instruction, the AI tends to be overly enamored of experience, showing lexical preferences with work experience first, then wage.

I endow workers with either one year of experience (Person A) or 0 years of experience (Person B). Across scenarios, I vary Person A's wage ask. The inexperienced person, Person B, always asks for $13/hour *unless* it faces a minimum wage of $15/hour, in which case it bids up. The AI is asked to select an applicant. I can then observe the hired worker's wages and experience level of the hired worker.

The prompt sent to GPT3 is:

> You are hiring for the role of "Dishwasher." The typical hourly rate is $12/hour. You have 2 candidates.
>
> Person 1: Has 1 year(s) of experience in this role. Requests $17/hour. Person 2: Has 0 year(s) of experience in this role. Requests $13/hour.
>
> Who would you hire? You have to pick one.
>
> *Person 2. Although they have no experience in this role, their request for $13/hour is closer to the typical rate of $12/hour.*

In terms of sample size, I tried each possible wage ask (13 through 19), In Figure 4, I plot each applicant's demeaned wage bid (relative to the pool) on the x-axis and their de-meaned experience (on the y-axis); I color and shape the point based on whether that applicant was hired ("H" for hired, "R" for rejected). I jitter (add random noise) points to prevent over-plotting. I also facet by whether or not a minimum wage was imposed.

We can the employer's clear "bliss point" up and to the left—an experienced worker proposing a low hourly rate. However, the AI employer also clearly does not have lexical preferences: the hire/reject "frontier" is a positively sloped line.

and so

Figure 4: Horton (2023) Simulated employer hiring preferences by relative experience and wage asks

Notes: This shows the faction of AI subjects choosing each option, by framing.

We can see that the imposition of the minimum wage causes wage compression at those at the bottom bid-up; the experience distribution does not change. Without the minimum wage, we can see that some of the workers with less experienced were still hired because of their relatively low wage bid. Under the minimum wage, they are forced to bid up and are thus less likely to be on the good side of the employer's indifference curve.

To actually check this empirically, I create a job-level data set and regress (1) hired worker wage and (b) hired worker experience on the minimum wage indicator. Table 1 reports the results. We can see in Column (1) that imposing a minimum wage raised hourly wages, as expected. This was expected as there was no instruction to AI employers about the possibility of not filling their jobs. In Column (2), we can see that imposing the minimum wage caused more experienced workers to be hired.

Table 1: Effects of minimum wage on observed wage and hired worker attributes

	Dependent variable:	
	w Hired worker wage	exper Hired worker experience
	(1)	(2)
\$15/hour Minimum wage imposed	1.833***	0.167***
	(0.076)	(0.045)
Constant	13.333***	0.667***
	(0.054)	(0.032)
Observations	360	360
R^2	0.621	0.037

Notes: This reports the results of imposing a minimum wage on the (1) hired worker wage and (b) hired worker experience.

This is obviously just a small slice of the parameter space one could explore with this scenario. The job could be varied, other worker attributes, knowledge about the minimum wage, the presence of alternatives, the potential for using capital to substitute, and so on. These would all be relatively straightforward to explore.

4 Conclusion

This paper reports the results of several experiments using GPT3 AIs as experimental subjects. The main conclusion is that this approach seems promising: it can qualitatively recover findings from experiments with actual humans. Furthermore, it can do this remarkably cheaply: All the experiments in the paper cost about \$50, and each runs in about 30 seconds. This expense is mostly because I ran each experiment many times as I debugged my code. I found in writing that paper that if there was, say, a typo in a response, it is easier just

to re-run the "experiment." It is feasible to try numerous variations of wording, prompts, answer order, and so on. Sample sizes can be arbitrarily large. There are no human subjects related ethical concerns with running these experiments.

Every experiment presented here was simple. But far more complex interactions are possible. AIs can "chat" with each other by having API calls respond to the output of other API calls. This allows for more complex gameplay. For example, I could have created a hardware store owner AI in the Kahneman et al. (1986) example and had them poll a sample of customers before choosing the new shovel price. Or in Charness and Rabin (2002), I could have made a Person A agent and had them respond to proposed allocations. For games that play out over time and memory is important, prior gameplay can simply be pre-pended to the prompt e.g., "You are Alice, you are playing with Bob in describe game. In the last n rounds, Bob has defected. How do you want to play now?"

What kinds of experiments are likely to work well? Given current capabilities, games with complex instructions are not presently likely to work well, but with more advanced LLMs on the horizon, this is likely to change. I should also note that research questions like what is "the effect of x on y" are likely to work much better than questions like "what is the level of x?." Consider that in my Kahneman et al. (1986) example, I can create AI "socialists" who are not too keen on the price system generally. If I polled them about who they want for president, there is no reason to think it would generalize to the population at large. But if my research question was "what is the effect of the size of the price increase on moral judgments" I might get be able to make progress. That being said, it might be possible to create agents with the correct "weights" to get not just qualitative results but also quantitatively accurate results. I did not try, but one could imagine choosing population shares for the Charness and Rabin (2002) "types" to match moments with reality, then using that population for other scenarios.

In terms of reproducible research, the data from these experiments can always be released. Or even if not released, anyone with the source code can try to replicate the results. One issue is that OpenAI—or any other LLM provider—is under no obligation to continue offering access to any particular model. However, these experiments could be redone with new AIs as they become available. Consider that no lab experiment comes with the guarantee that the same humans will later be available for your replication.

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."

In traditional experimental work, there is a concern about researchers data-mining on experimental results to find "significant" findings. A partial solution is pre-registration. This is not likely to work for AI experiments. It is the fixed cost of doing an experiment that makes

a registry work, incentive-wise, but when it costs $1 and 30 seconds to run an experiment, it is hard to see what the benefit would be. The more realistic option would just be a norm of these experiments being "push button" reproducible where one can simply fork a repository and replace API keys and re-run. This would allow people to verify themselves if results are sensitive to slight differences in framing.

References

Aher, Gati, Rosa I Arriaga, and Adam Tauman Kalai, "Using Large Language Models to Simulate Multiple Humans," *arXiv preprint arXiv:2208.10264*, 2022.

Alberti, Chris, Kenton Lee, and Michael Collins, "A bert baseline for the natural questions," *arXiv preprint arXiv:1901.08634*, 2019.

Argyle, Lisa P, Ethan C Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate, "Out of One, Many: Using Language Models to Simulate Human Samples," *arXiv preprint arXiv:2209.06899*, 2022.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? (Parrot Emoji)," in "Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency" FAccT '21 Association for Computing Machinery New York, NY, USA 2021, p. 610–623.

Binz, Marcel and Eric Schulz, "Using cognitive psychology to understand GPT-3," 2022.

Charness, Gary and Matthew Rabin, "Understanding social preferences with simple tests," *The quarterly journal of economics*, 2002, *117* (3), 817–869.

Friedman, Milton, "The methodology of positive economics," 1953.

Horton, John J., "Price Floors and Employer Preferences: Evidence from a Minimum Wage Experiment," *Working paper*, 2023.

Horton, John J, David G Rand, and Richard J Zeckhauser, "The online laboratory: Conducting experiments in a real labor market," *Experimental economics*, 2011, *14* (3), 399–425.

Kahneman, Daniel, Jack L Knetsch, and Richard Thaler, "Fairness as a constraint on profit seeking: Entitlements in the market," *The American economic review*, 1986, pp. 728–741.

Kuhlman, Brian, Gautam Dantas, Gregory C Ireton, Gabriele Varani, Barry L Stoddard, and David Baker, "Design of a novel globular protein fold with atomic-level accuracy," *science*, 2003, *302* (5649), 1364–1368.

List, John A., "Does Market Experience Eliminate Market Anomalies? The Case of Exogenous Market Experience," *The American Economic Review*, 2011, *101* (3), 313–317.

Lucas, Robert E., "Methods and Problems in Business Cycle Theory," *Journal of Money, Credit and Banking*, 1980, *12* (4), 696–715.

MacKenzie, Donald, *Do Economists Make Markets?: On the Performativity of Economics*, Princeton University Press, 2007.

Samuelson, William and Richard Zeckhauser, "Status quo bias in decision making," *Journal of risk and uncertainty*, 1988, *1* (1), 7–59.

Schelling, Thomas C, "Dynamic models of segregation," *Journal of mathematical sociology*, 1971, *1* (2), 143–186.

Simon, Herbert A., *The Sciences of the Artificial, 3rd Edition*, Vol. 1 of *MIT Press Books*, The MIT Press, September 1996.

Westby, Samuel and Christoph Riedl, "Collective Intelligence in Human-AI Teams: A Bayesian Theory of Mind Approach," 2022.