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

Apostolos Filippas
Fordham

John J. Horton
MIT & NBER

Benjamin S. Manning
MIT

February 12, 2025

Abstract

Large language models (LLM)—because of how they are trained and designed—are implicit computational models of humans—a *homo silicus*. LLMs can be used like 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. Experiments using this approach, derived from Charness and Rabin (2002), Kahneman et al. (1986), and Samuelson and Zeckhauser (1988) show qualitatively similar results to the original, but it is also easy to try variations for fresh insights. LLMs could allow researchers to pilot studies via simulation, first improving their experimental design and searching for novel social science insights to test in the real world. In this paper, we offer a framework for when this approach is likely to prove useful.

*Thanks to the MIT Center for Collective Intelligence, Tyler Cowen, and the Mercatus Center for their generous funding. Thanks to Daron Acemoglu, David Autor, Mohammed Alsobay, Jimbo Brand, Elliot Lipnowski, Shakked Noy, Paul Röttger, Daniel Rock, and Hong-Yi TuYe, for their helpful conversations and comments. Special thanks to Yo Shavit, who has been extremely generous with his time and thinking. 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.; deducing its expected behavior, which can then be compared to the behavior of actual humans in (b)-type research.

In this paper, we argue that newly developed large language models (LLM)—because of how they are trained and what they are trained on—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 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 (Bowman, 2023). These responses can, in a growing number of cases, predict human behavior in never-before-seen experiments (Binz and Schulz, 2023; Li et al., 2024; Hewitt et al., 2024).

Building on these results, we consider 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—(i) can be thought of as computational models of humans and (ii) likely possess a great deal of latent social information. For (i), LLMs are designed to respond to specific instructions in ways determined and evaluated by humans. And when prompts are designed that effectively instruct an LLM to respond as if it were a human, they can be thought of as computational models of humans. For (ii), 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 many hours to work, what to do when prices increase, and so on.

Like all models, any particular *homo silicus* is 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 realistic, rational, or even sensical. But ultimately, what will matter in practice is whether these AI experiments are practically valuable for generating insights. As such, we demonstrate the method through a variety of experiments.

In this paper we do several things. (1) we demonstrate the basic approach with a variety of simple experiments. (2) we show how the robustness and generalizability of some simu-

[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."

lation result can be explored. (3) we offer a way to think about these kinds of simulations epistemologically (4) we offer a framework for when they are likely to prove useful

An earlier version of this paper reported several simulation experiments using the most capable models available at that time. Since then, several far more capable models have become available, including those with open weights.

For our simulations, we begin with the simple unilateral dictator games from Charness and Rabin (2002). We show that endowing an AI agent with various social preferences affects play. Instructing the AI agent that it only cares about equity causes it to choose the equitable outcomes; telling the agent it cares about efficiency causes the selection of the payoff maximizing outcomes; telling the agent is self-interested causes the selection of allocations that maximize narrow self-interest. We then use these results to generate new samples of agents that respond increasingly like human subjects in new games.

Next, we present the AI agents with a decision-making scenario introduced by Samuelson and Zeckhauser (1988). In the paper, the respondent must allocate a federal budget between highway and car safety. The original paper showed humans are subject to a status quo bias, preferring budget options when presented as the status quo. We replicate this study by putting AI agents through different versions of this same scenario. We find that some LLMs exhibit a status quo bias, preferring allocations closer to the status quo, while others do not.

We 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 a benefit of AI subjects, unlike Kahneman et al. (1986), we also vary the amount by which the store increases the price, and the political leanings of the respondent. We show that aggressive price gouging is viewed more negatively; the largest price increases earn approbation even from AI Conservatives. Endowed political views matter, with predictable effects—AI agents towards the right are generally more sanguine about gouging. The effects are robust to dozens of permutations of the original experiment.

Finally, we explore a more realistic decision-making scenario derived from two field experiments. Horton (2023) shows that employers facing a minimum wage will substitute for higher-wage workers. We create a scenario where an employer is trying to hire a worker as a dishwasher and faces pairs of applicants that differ in their experience and requested wage. We then randomly impose a minimum wage that forces applicants asking for wages below that minimum to bid up. To illustrate the flexibility of AI subjects, we simultaneously test a well-known finding from Bertrand and Mullainathan (2004), who found causal evidence of racial discrimination in labor markets. We randomly vary the names of the applicants to be more stereotypically African American or White-sounding. We find evidence of racial discrimination in the AI subjects' hiring decisions, but in the opposite direction of the original

study; AI subjects are a little more likely to hire the applicants with stereotypically African American names. Similarly to Horton (2023) imposing a minimum wage does causes a shift in the AI's hiring of more experienced applicants. Both of these results hold for dozens of LLMs, often trained by different developers.

Ultimately, we care about the behavior of actual humans and so for now, results from AI experiments will still require empirical confirmation. As such, what is the value of these experiments? The most obvious use is to pilot experiments in silico first to gain insights. Researchers could cheaply and easily explore the parameter space, test whether behaviors seem sensitive to the precise wording of various questions, evaluate their assumptions, and generate data that will "look like" the actual data. The advantages in terms of speed and cost are enormous. The experiments in this paper were run in minutes for a trivial amount of money with software that allows for seamless reproducibility and sharing of results.[2] 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.

As with most exciting new methods and technologies, AI experiments have come under increased scrutiny as they have proliferated in academic work. Some argue that opaque training processes and unrepresentative data make inference on AI experiments problematic. Others worry that the LLMs are simply regurgitating text from the training corpus or that they are answering in perfect accordance with the social science theories they have digested during training. A third flavor of critique questions where the boundaries of AI experiments lie, concerning their ability to generate useful versus misleading results.

We address each of these critiques and more in turn, although a common theme runs through our responses: despite appearing novel, many of the criticisms concern human subjects as much as they concern experiments with LLMs. Such parallels can help economists better understand the assumptions in their simulated and real-world work. Moreover, effective prompt engineering can often help address many of purported problems with the simulations (Schulhoff et al., 2024; Jahani et al., 2024; Raman et al., 2024).

Of course, some critiques are impossible to address with just a few experiments and rebuttals. No single paper can establish all the boundaries to when AI experiments are informative proxies for humans. Particularly when it comes to massive machine-learning models, whose developers often struggle to fully understand their behavior (Bowman, 2023). External validity is an empirical question—many empirical questions—that will require collective exploration at scale. We will need tools and benchmarks that can help guide research even as LLMs advance at a breakneck pace As such, we present a set of fully documented and

[2]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 et al., 2003)

open-source software tools to help facilitate this process.

In terms of contribution, the most closely related paper is Aher et al. (2022), which 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, their 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 twofold: (i) laying the foundation for a framework to guide how and when we should use these experiments in economic work, along with developing the necessary software and technical infrastructure. And (ii) drawing the connection to the common research paradigm of economics and the role a foundational assumption like rationality plays in research. LLM experimentation is more akin to the practice of economic theory, despite superficially looking like empirical research.

2 Large language models

Large language models (LLMs) are neural networks with complex architectures and large numbers of parameters, estimated (trained) using massive amounts of text. Given an input prompt, an LLM generates a response by first predicting the most likely next word, and then using the prompt and the words predicted so far to predict subsequent words.[3]

Training an LLM consists of a "pre-training" phase and an "reinforcement learning with human feedback" (RLHF) phase. In the pre-training phase, the LLM is trained on unlabeled text data, the training data, with the objective of predicting correctly words in text sequences. A pre-trained LLM estimates a conditional probability distribution over its training data; the exact estimated distribution depends on design choices such as the architecture and hyperparameters of the LLM. In the RLHF phase, the pre-trained LLM is trained further using human feedback. First, human evaluators prompt the LLM and evaluate its responses to these prompts.[4] The human evaluation data is then used to estimate a separate "reward model," which predicts human feedback given a prompt and an LLM response. The LLM's parameters are updated through a form of reinforcement learning so that its responses receive favorable feedback from the reward model. The fully trained LLM's responses are hence optimized toward both accurately predicting the likely next words in sequences of text and receiving high feedback scores from the reward model.

An LLM can always undergo further training called "fine-tuning."[5] We can fine-tune an

[3] Technically, LLMs use tokens, which can be sentences, words, parts of words, or characters.

[4] For example, the human evaluators may rate the responses on a scale of 1 to 10, depending on how well a response answers a question or matches a particular style of writing.

[5] Some refer to any training beyond the pre-training phase as fine-tuning, including RLHF, which is sometimes called instruction fine-tuning.

LLM by repeating either phase of the original training process: the pre-training phase with new data or the RLHF phase with additional human feedback. Fine-tuning is often employed to improve the LLM's performance on a specific task or to make its responses more consistent with a particular persona or writing style.

Users can adjust how the LLM responds to input prompts. For example, users can often change a "temperature" parameter to control how the LLM samples from its response distribution. At temperature zero, the LLM is deterministic: it always outputs the highest-probability response. Increasing the temperature makes the output stochastic and the response distribution more uniform. As such, prompting an LLM multiple times with the same prompt at a non-zero temperature generates a fuller picture of its response distribution.

3 Experiments

We conduct four experiments with LLMs: we replicate and extend three laboratory experiments, and then combine two field experiments from the economics literature. In the first experiment, we demonstrate that LLMs respond reasonably to prompts and follow instructions effectively. Then, we show how to leverage these responses to construct samples of AI subjects that respond increasingly like human subjects. The second experiment reveals that LLMs often exhibit human decision-making biases. In the third experiment, we show how we can use LLMs to easily extend experiments and search for new insights. The fourth experiment shows that researchers can use LLMs in complex, realistic decision-making scenarios.

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

We use games from Charness and Rabin (2002) to show that AI subjects can map social and economic concepts to understandable choices. For example, agents told they are efficiency-minded choose the efficient allocations; those told they are selfish choose selfish allocations, and so on. It is essential to note that this is a new possibility—the original version of this paper demonstrated that LLMs of slightly older vintage did not have such capabilities. We reproduce this result with a number of newer models, including open models.

3.1.1 Matching distributions

The actual human responses in Charness and Rabin (2002) are, of course, a distribution—what does the variation in human responses mean for agent-based simulations? With real humans, there there are no clear "types"—real human respondents do not uniformly choose the most efficient, most equitable, or most self-interested allocations. This is not surprising: human preferences are heterogeneous, or in more economic terms, samples of human subjects often consist of a mixture of types.

With AI agents, there are several ways to induce variation in responses: 1) the prompt used, 2) the model, 3) parameters like temperature. The same prompt, to the same model, with a 0 temperature will just return the same response each time. With so many free parameters, what can guide the creation of agent sets?

Suppose you were trying to model agents guessing the result of a die roll, and suppose the distribution was uniform. You could create agents 1 through 6 and give them instructions "you always report X"—this would 'work' on the die game, but fail badly on the *dice* game (guess the sum of two rolls).

However, showing that a "type" of agent plays the game the "right" way means to match the distribution of real humans would require the correct distribution of agent types. We do this emprically, using

Just as no economist would expect a single person to exhibit multiple conflicting preferences simultaneously, so, too, with AI subjects. However, we can construct AI samples that better mirror human behavior by identifying key types of preferences that cause the AI agents to respond differentially. We do this by creating a mixture of types of AI subjects whose responses match those of human subjects in the original experiment. Then, we use this sample of AI subjects plays a new set of games. This reweighted sample of agents behaves much more similarly to the human participants.

We first replicate the unilateral dictator games from Charness and Rabin. In one example

"Left": Person A gets 400 and Person B gets 600
"Right": Person A gets 700 and Person B gets 300,

the dictator (Person B) has to choose between two allocations of money, "Left" and "Right," between herself and the other player (Person A). We can write this game as:

$$\text{Person B Chooses: } (\underbrace{400}_{\text{to A}}, \underbrace{600}_{\text{to B}}) \text{ ``Left''} \quad \text{vs.} \quad (\underbrace{700}_{\text{to A}}, \underbrace{300}_{\text{to B}}) \text{ ``Right''}.$$

To replicate this experiment using AI subjects in lieu of human subjects, we prompt LLMs with descriptions for each game, and ask them to respond with their preferred allocation. We also create AI subjects endowed with "personas" or "types," by prepending a description of the persona in the prompt. The personas are:

Inequity-averse: "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."

We have each AI subject play each game 100 times with the temperature parameter set to one. We use GPT-4o, Claude-Sonnet-3.5, and Llama-3-70B—the most capable LLMs

to date developed by industry leaders OpenAI, Anthropic, and Meta. See Appendix A.1 for the precise wording of the prompt.

Figure 1 plots the results of our experiment. Each column corresponds to a different persona, and each row to a different game from Charness and Rabin (2002). Within each pane, the y-axis shows the model, and the x-axis shows the fraction of subjects choosing the option "Left." The white bars depict the AI agents' choices, and the gold bars depict either the choice of the human subjects in Charness and Rabin (leftmost column) or the choice that a perfectly aligned AI subject would make with the given persona (all other columns). For example, in the Barc2 game, 52% of the human subjects chose "Left" (400,400), and the efficient choice would be to always select "Right."

Starting from the leftmost column, we can see that the choices of the persona-less AI subjects are quite different from those of the human participants in the original experiment. The responses of AI subjects are consistent within each game: Llama-3-70B is seemingly more selfish, and GPT-4o and Claude-Sonnet-3.5 are more efficiency-minded. In contrast, the human subjects were often fairly split between "Left" and "Right," except for the extremely spiteful Berk23 scenario in which Person B forgoes 200 to ensure Person A does not get 800. Notably, the AI subjects' responses fail to replicate those of the human participants in Charness and Rabin, even though this study is almost certainly included in the training data for all three models.[6]

AI subjects endowed with personas respond consistently in line with those personas. All efficiency-minded AI subjects are perfectly compliant. With the exception of Claude-Sonnet-3.5 in a small proportion of the spiteful Berk23 games, and GPT-4o in one out of the 100 Barc8 games, all inequity-averse AI subjects reliably choose the option with the smallest difference between the allocations. Finally, self-interested AI subjects act selfishly with perfect precision.

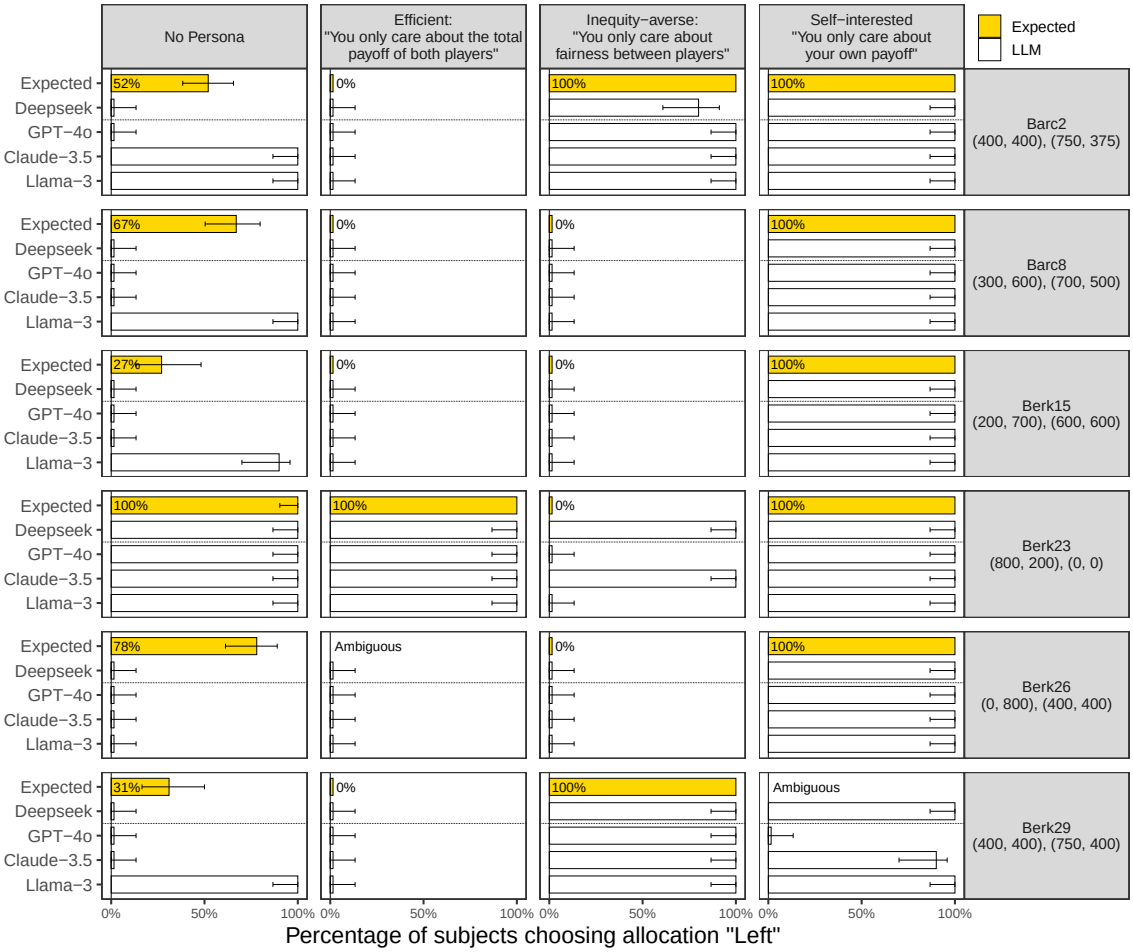
3.1.2 Calibrating AI agents

Although the first moment of the persona-less AI subjects' responses did not match that of the human subjects' responses, AI subjects endowed with personas exhibited substantial variation in their answers and generally responded as expected. We use this variation to construct samples of AI subjects who respond "Left" at rates as similar as possible to the human subjects in the games from Figure 1. Then, we have this "calibrated" sample of AI subjects play a new set of out-of-sample games and compare their responses to the human subjects'.[7] We also draw these new games and the corresponding human subjects from Charness and

[6] All LLMs respond accurately to highly specific questions about the types of games played, the payoffs of the players, and the proportions of people making each choice in Charness and Rabin.

[7] This approach is similar to traditional machine learning, where one uses training data to train a model—adjusting hyperparameters to minimize in-sample error—and then testing the model on out-of-sample data

Figure 1: Replication of single-stage dictator games from Charness and Rabin (2002).



Notes: This figure reports the results of replications of single-player dictator games from Charness and Rabin (2002). We construct AI subjects using GPT-4o, Claude-Sonnet-3.5, and Llama-3-70B. The columns correspond to personas with which we endow AI subjects, and the rows correspond to different games. Each AI subject plays each game 100 times. The x-axis shows the percentage of subjects choosing "Left." The y-axis shows both the AI subjects' responses and the "Expected" proportion for each column's given persona. The white bars depict the AI agents' choices, and the gold bars depict either the choice of the human subjects in Charness and Rabin (leftmost column) or the choice that a perfectly aligned AI subject would make with the given persona (all other columns). The error bars report 95% Wilson confidence intervals. See Appendix A.1 for more details on the implementation of the experiment.

Rabin

The new games are sequential: the monetary allocations depend on both players' decisions. In the first stage, Person A is asked to choose a given allocation or let Person B choose one of two other known allocations. Person B is asked to choose an allocation but is not informed of Person A's choice—until the payoffs are realized. The game is structured as follows:

Stage 1 (Person A chooses): $\underbrace{(\underbrace{500}_{\text{To A}}, \underbrace{500}_{\text{To B}})}_{\text{“Left”}}$ vs. $\underbrace{\underbrace{(400, 600) \text{ vs. } (700, 300)}_{\text{Let Person B choose}}}_{\text{“Right”}}$.

Stage 2 (Person B chooses): $\underbrace{(\underbrace{400}_{\text{To A}}, \underbrace{600}_{\text{To B}})}_{\text{“Left”}}$ vs. $\underbrace{(\underbrace{700}_{\text{To A}}, \underbrace{300}_{\text{To B}})}_{\text{“Right”}}$.

As a shorthand, we write this game as:

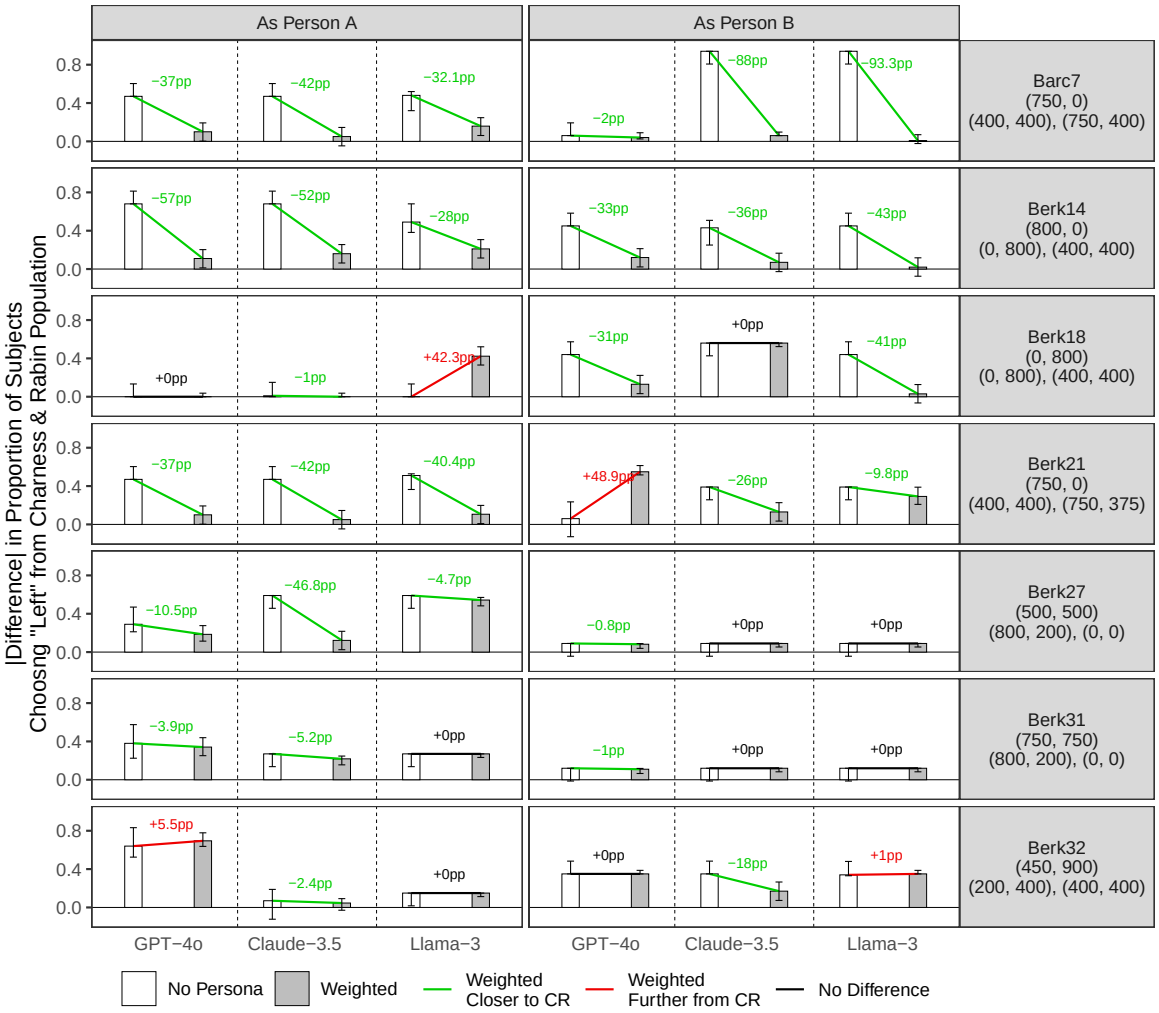
$$(500, 500)$$
$$(400, 600), (700, 300)$$

We next describe how we construct our AI agent sample. We represent each persona by a vector $v = (p_{Barc2}, p_{Barc8}, p_{Berk15}, p_{Berk23}, p_{Berk26}, p_{Berk29})$, where p_g is the proportion of AI subjects with this persona that chose "Left" in game g from Figure 1. For example, the "efficient" GPT-4o AI subjects are represented by $v_{\text{E}} = (0, 0, 0, 1, 0, 0)$, as they only choose "Left" in the Berk23 scenario; the population of Charness and Rabin is $v_{CR} = (.52, .67, .27, 1, .78, .68)$. For each LLM, we then compute the weights $w = (w_E, w_I, w_S) \in [0, 1]^3$ that minimize the mean squared error (MSE), that is, the weights that minimize $\frac{1}{6}||w_E v_E + w_I v_I + w_S v_S - v_{CR}||^2$, subject to the constraint that $w_E + w_I + w_S = 1$. These weights are $(0.417, 0.020, 0.562)$ for GPT-4o, $(0.490, 0, 0.510)$ for Llama-3, and $(0.458, 0, 0.541)$ for Claude-3.5.

We conduct our experiments as follows. For each model, we randomly sample 100 AI subjects according to the computed weights. Each AI agent plays several two-stage dictator game scenarios, both as Person A and Person B 10 times, with the underlying LLM's temperature set to 1. Additionally, a "control" persona-less AI subject plays each game as each Person 100 times. Figure 2 plots the results of our experiment with 21,300 AI subject responses. Each row corresponds to a different game. The left column shows the results for AI subjects as Person A, and the right for Person B. The x-axis corresponds to the different models. For each model, the grey bar is the sample of AI agents constructed with the weights determined by the mixture model, and the white is the persona-less controls. The y-axis shows the absolute difference between the fraction of AI agents choosing "Left" and the fraction of human subjects choosing "Left" in Charness and Rabin. The lines show the difference in accuracy between the weighted and persona-less samples. Green lines indicate the weighted sample is closer to the human subjects than the control, red lines indicate that the weighted sample is farther, and black lines mean they are equidistant.

The results are striking: across all models, the responses of the weighted AI subject samples are much closer to those of the human subjects. The MSE of the persona-less control (0.176) is more than three times larger than the MSE of the weighted samples (0.053).

Figure 2: Replication of two-stage dictator games from Charness and Rabin (2002).



Notes: This figure reports replications of the sequential two-player dictator game Charness and Rabin (2002) with three LLMs: GPT-4o, Claude-Sonnet-3.5, and Llama-3-70B. The y-axis provides the absolute difference between the fraction of AI agents choosing "Left" and the fraction of human subjects choosing "Left" from Charness and Rabin. The x-axis corresponds to the different LLMs. The rows each correspond to a different game labeled on the right, and the columns to when the agents are playing as Person A (Left column) or Person B (right column). The color of the bars corresponds to whether the AI agents are constructed with the weights determined by the mixture model (grey) or if they are the agents without personas (white). The lines connecting weighted and persona-less samples show their difference in distance from the human subjects and whether the weighted sample was closer (green), further (red), or equidistant (black) as compared to the persona-less sample. The error bars report 95% Wilson confidence intervals for the proportion of AI agents choosing "Left" centered at the absolute difference from the reported means from Charness and Rabin. The manner in which Charness and Rabin report two-stage game results does not allow us to compute confidence intervals for the human subjects. See Appendix A.1 for more details on prompt construction and the data-generating process.

Furthermore, the Llamma-3 and Gpt-4o weighted samples weakly dominate their respective persona-less samples in 11 out of the 12 panes; and Claude-3.5 weighted samples do so in 10 out of 12.

We could use this method of calibrating agents to construct AI subjects that respond like human subjects in other scenarios or with other types of personas or preferences. Such personas need not be as simple as the ones we used here. Types could be complex natural language profiles, including personality traits, political beliefs, or demographic profiles. The only requirement is that personas exhibit differential variation in their responses such that we can calibrate the proportion of personas to match the response distribution of an analogous human sample. Then, we can use these calibrated AI subjects to explore new scenarios with more confidence that their responses will be informative. More generally, this method shows we are not limited to whether AI subjects respond like humans, but rather if we can construct AI subjects that respond like humans in particular settings.

3.2 Prospect theory, risk, and complexity (Oprea, 2024)

Until recently, the literature on decision-making under risk was definitive: humans tend to be risk-averse when there is a small probability of gains and risk-seeking when there is a small probability of losses. Furthermore, these preferences often reverse as the probabilities of gains and losses increase. This fourfold pattern of risk and loss aversion—predicted by prospect theory—is one of the most well-documented findings in behavioral economics (Kahneman and Tversky, 1979, 1992). However, using simple expected value calculations (Oprea, 2024) demonstrated that the fourfold pattern is not just a product of non-standard utility functions or reference-dependent preferences. Rather, it partially reflects a more general limitation of human capacity to process information in complex environments.

In their main experiment, Oprea had two conditions. First, they asked subjects to provide certainty equivalents for 5 lotteries with probability $p \in \{0.1, 0.25, 0.75, 0.9\}$ of winning \$25 and 5 lotteries with the same probabilities of losing \$25. These lotteries were structured such that participants were shown a series of 100 boxes, each with $p \times 100$ boxes containing \$25 (or -\$25) and $(1 - p) \times 100$ boxes containing \$0. For each set of boxes, participants were asked to provide a certainty equivalent to opening one randomly chosen box and receiving (or losing) the amount of money inside. In the second "mirror" condition, participants were shown identical sets of boxes. But instead of providing a certainty equivalent for one random box, they were asked to provide certainty equivalents for opening all the boxes and receiving (or losing) the average amount of money in the boxes combined. In this condition, there is no uncertainty; therefore, there should be no fourfold pattern in participant preferences. Yet, the results between the two conditions were virtually indistinguishable. Participants responded in ways perfectly consistent with well-documented risk preferences to calculations without any risk.

We replicate this experiment with AI subjects—both the lottery and the mirror conditions. We create 10 subject profiles, each with a different hobby (e.g., hiking, reading, cooking).

We tell each AI subject that they are endowed with \$50 and ask them to provide certainty equivalents for sets of boxes with $p \in \{0.05, 0.10, \ldots, 0.95\}$ of winning or losing \$25. That is 13 additional probabilities for each condition-gain and loss combination beyond the original experiment. We use GPT-3.5-TURBO, GPT-4O, CLAUDE-HAIKU-3, and CLAUDE-SONNET-3.5 and set the temperature parameter to 1. Using these LLMs allows us to compare the AI subjects' responses as the capabilities of LLMs progress, as well as between different LLM developers. We repeated the experiment twice, once without additional information and once with the AI subjects instructed that they were "bad" at math. See Appendix A.2 for the precise wording of the prompt and additional details on the experiment. It is worth noting that this within-subject experimental design is possible because AI subjects have no memory unless we explicitly provide them with information about their previous decisions. This would be impossible with human subjects because they would likely become aware of the experimental manipulation after exposure to each status quo framing.

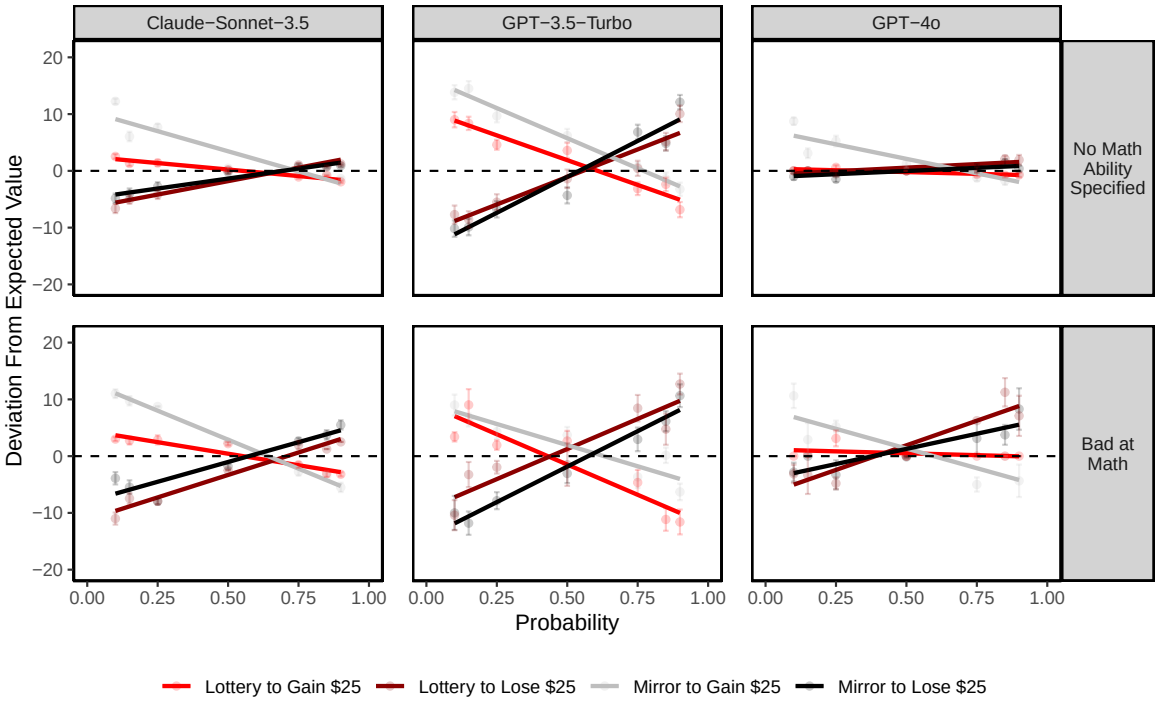
Figure 3 shows that AI subjects almost universally exhibit the fourfold pattern of risk and loss in both the lottery and mirror conditions. The x-axis shows the probability of winning (losing) \$25, and the y-axis shows the deviation from the expected value of the lottery or mirror. Each column corresponds to a different model, and the rows correspond to the different endowed mathematical abilities of the AI subjects. The red and grey dots represent the average responses of the AI subjects to the gain lotteries and mirrors, respectively. Maroon and black dots represent the same for the loss of lotteries and mirrors. Each point is given with a 95% confidence interval, and the colored lines show the OLS fits for each condition by loss and gain.

As a baseline, the less advanced models—GPT-3.5-TURBO and CLAUDE-HAIKU-3—strongly display the fourfold pattern in both conditions. They are "risk-averse" when the probability of gain is low and "risk-seeking" when the probability of winning is high, and the opposite is true for loss lotteries. This holds for both the lottery and mirror conditions, although the point estimates and OLS fits vary somewhat. For the more capable models, CLAUDE-SONNET-3.5 exhibits the pattern to a much lesser extent, and GPT-4O almost exclusively selects the expected value for every lottery and mirror. However, when we inform the AI subjects that they are "bad" at math, the pattern is much more pronounced for both the more capable models in both conditions.

3.3 *Status quo* bias in decision making (Samuelson and Zeckhauser, 1988)

Samuelson and Zeckhauser (1988) demonstrated in several decision-making scenarios that people are more likely to make a choice when it is presented as the *status quo*. In one of their scenarios, they asked subjects to allocate a safety budget between automobiles and highways. The instructions were:

Figure 3: TOO add



Notes: asdf

> "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 between four funding allocations: (70% auto, 30% highways), (60%, 40%), (50%, 50%), or (30%, 70%). The main experimental manipulation was to present the options relative to different status quo allocations.[8]

We replicate Samuelson and Zeckhauser's exp. First, we endow each subject with one of twelve beliefs about the importance of car and highway safety (see Appendix A.3 for more details on how we construct the belief prompts). For each belief, we create a "car owner" and a "non car owner" AI subject. We then have each of the 24 AI subjects choose one of four funding allocations in four scenarios: each one phrased to indicate a different option as the *status quo*. We ask each subject to respond with their preferred allocation 25 times with the LLM temperature set to 1. We use GPT-4, GPT-4O, CLAUDE-SONNET-3, and

[8]For example, in the (60, 40) status quo framing, the options were: "Decrease the highway program by 10% of budget and raise the auto program by like amount (70, 30), maintain present budget amounts for the programs (60, 40), ..."

Claude-Sonnet-3.5. Using these LLMs allows us to compare the AI subjects' responses as the capabilities of LLMs progress, as well as between different LLM providers. It is worth noting that this within-subject experimental design is possible because AI subjects have no memory unless we explicitly provide them with information about their previous decisions. This would be impossible with human subjects because they would likely become aware of the experimental manipulation after exposure to each status quo framing.

We model the AI subjects' responses as discrete choices over the four allocations using a Logit specification and report the results in Table 1. In Column (1), we see that AI subjects are less likely to choose an alternative as more funds are allocated to highway safety (AutoShare). In Column (2), we add a variable indicating whether an allocation was framed as the status quo. All AI subjects exhibit considerable status quo bias, although GPT-4o to a lesser extent. In Column (3), we replace the indicator variable with a variable measuring the absolute distance between each allocation's auto share and the status quo allocation's auto share. We see that allocations "farther" from the status quo are less likely to be chosen for all models. We also conduct likelihood ratio tests, which show that the specification of Column (3) explains the data better than the specification of Column (2) ($\chi^2 = 23.69$). Overall, our results suggest that some AI subjects shift their preferences toward the status quo allocation—similar to human subjects.

3.4 Fairness as a constraint on profit-seeking (Kahneman et al., 1986)

Kahneman et al. (1986) assess subjects' views about fairness in markets by asking them to evaluate several scenarios. In a price gouging example, subjects were asked to respond to the following vignette:

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

In the original paper, 82% of subjects responded either "Unfair" or "Very Unfair."

Two natural follow-up questions not explored in the original paper are (i) whether the subjects' political preferences and associated attitudes toward markets affect their views, and (ii) whether there exists a dose-response relationship, with more aggressive price hikes seen as more egregious. To explore these questions, we endow AI subjects with six political views ranging from "socialist" to "libertarian," and ask each AI subject to assess the fairness of price increases to $16, $20, $40, and $100. We use Gpt-4 with the model temperature at 1, and collect 100 responses to each combination of price hikes and political views, which generates 2,400 observations. More details on how we construct the prompts can be found in Appendix A.4.

Table 1: zeckhauser - regression

	Dependent variable:		
	Choice of Allocation		
	(1)	(2)	(3)
Claude-3: AutoShare	−0.004* (0.002)	−0.004* (0.002)	−0.008* (0.002)
Claude-3.5: AutoShare	−0.012* (0.002)	−0.013* (0.002)	−0.017* (0.002)
GPT-4: AutoShare	−0.009* (0.002)	−0.010* (0.002)	−0.016* (0.002)
GPT-4o: AutoShare	−0.005* (0.002)	−0.005* (0.002)	−0.008* (0.002)
Claude-3: I(StatusQuo)		0.739* (0.054)	
Claude-3.5: I(StatusQuo)		0.882* (0.054)	
GPT-4: I(StatusQuo)		1.054* (0.054)	
GPT-4o: I(StatusQuo)		0.341* (0.057)	
Claude-3: |DistStatusQuo|			−0.027* (0.002)
Claude-3.5: |DistStatusQuo|			−0.028* (0.002)
GPT-4: |DistStatusQuo|			−0.047* (0.002)
GPT-4o: |DistStatusQuo|			−0.019* (0.002)
Observations	5,596	5,596	5,596
Log Likelihood	-7,717.299	-7,307.459	-7,295.614

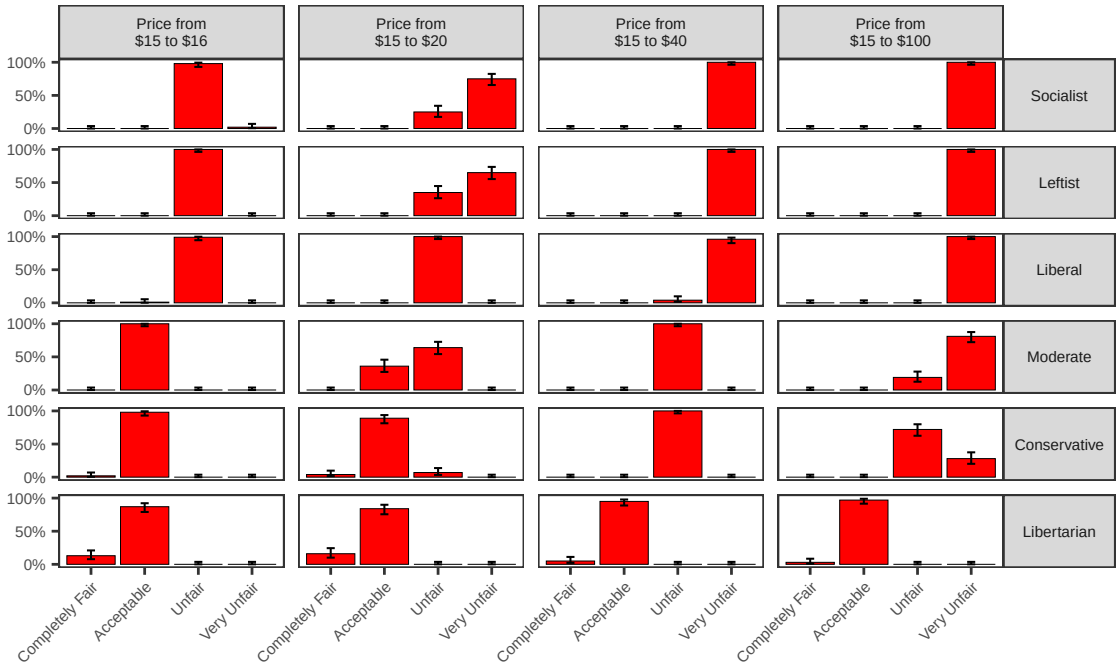
Notes: This table reports the results from our experiment replicating Samuelson and Zeckhauser (1988) with four LLMs: GPT-4, GPT-4O, CLAUDE-SONNET-3, *and* CLAUDE-SONNET-3.5. *Each column reports a discrete choice model using a logit to estimate the probability of choosing each of the four allocation alternatives for the National Highway Safety Commission. Observations are at the decision-maker level. The independent variables are the proportion of funds allocated to automobile safety for each choice alternative, an indicator of whether or not the choice alternative was framed as the status quo, and the absolute distance between each choice alternative and the option framed as the status quo. See Appendix A.3 for more details on prompt construction. Significance Indicator: *$p<0.05$.*

Figure 4 reports the results of our experiment. Columns correspond to different price increase scenarios and rows to different political bends of the AI subjects. The x-axis shows the possible responses, and the y-axis shows the proportion of AI experimental subjects who choose each option.

The AI subjects' political beliefs affect their views about the price hikes: socialist and leftist subjects judge the hikes to be "Unfair" or "Very unfair," and subjects with more conservative or libertarian views find price hikes more morally permissible. There is also a clear dose-response relationship, with larger price hikes generally viewed as less fair than smaller ones—even the "conservative" AI subjects judge the $100 price hike as "Unfair" or "Very Unfair" every time. The answers of the AI subjects are also fairly consistent: in about half of the panes, we get the same response in all 100 observations.

In terms of replicating the original study, Kahneman et al. report that their experimental subjects were a random sample of Canadians from Toronto and Vancouver, two major urban areas, and an about even split between male and female. This suggests that the original sample was relatively left-leaning. Directionally consistent with this result, AI subjects with

Figure 4: Replication and extension of price gouging experiment from Kahneman et al. (1986).



Notes: This figure reports the results of our experiment replicating and extending the price gouging experiment from Kahneman et al. (1986). Each column corresponds to a different price hike scenario, and each row to the AI subject's political leanings. The x-axis shows the subjects' responses, and the y-axis shows the proportion of AI subjects choosing each option for a given framing. The error bars represent 95% multinomial Wilson confidence intervals and include uncertainty estimates for responses with zero counts. AI subjects were constructed using GPT-4 with the temperature parameter set to 1. We collect 100 responses for each AI subject persona and scenario combination. See Appendix A.4 for more details on prompt construction.

left-leaning views always judge the original price hike to be "Unfair" or "Very Unfair."

3.4.1 Testing generalizability and evaluating memorization

One concern with our fairness experiment, is that the AI subjects may have memorized the results of the original study during training and are simply repeating said results back when prompted. Another concern is that the AI subjects responses are an artifact unique to the snow shovel example. We next show how to use AI subjects to directly address these threats to the experiment's credibility.

We conduct three follow-up experiments to the price-gouging study. First, we translate the original prompt into 10 different languages and run the experiment in each language.[9] Second, we ask GPT-4 to generate ten alternative versions of the original phrasing, varying the item being sold, the location, the event, or a combination of all three. For the full list of alternatives, see Appendix A.4, but one example is:

[9]These are French, German, Spanish, Italian, Portuguese, Greek, Japanese, Mandarin, Korean, and Arabic.

> A bakery has been selling artisan bread loaves for $15. The morning before a major holiday, the bakery raises the price to $20. Rate this action as: (1) Completely Fair, (2) Acceptable, (3) Unfair, or (4) Very Unfair.

Third, we ask GPT-4 to generate ten "adversarial" versions of the original phrasing, that is, to generate a vignette which would result in AI subjects responding differently than in our original experiment. The adversarial examples are also provided in Appendix A.4, but they are the same format as the alternative examples. We conduct all three of these additional experiments 10 times with the temperature parameter set to 1. With four price gouges, six political leanings, and 10 variations for each permutation, that's 2,400 observations for each of the additional experiments.

We regress the AI subjects' choices, measured as a continuous variable ranging from 1 ("Completely Fair") to 4 ("Very Unfair"), on the price change and the AI subjects' political endowments. Table 2 reports the results of this regression. Each column represents a different permutation of the experiment. Columns (1-3) report estimates for the baseline experiment, and Columns (4-6) report estimates for the follow-up experiments. We use "Libertarian" as the reference class.

The responses of AI subjects remain qualitatively similar in the follow-up experiments: higher price hikes are viewed as increasingly unfair, and all estimates are directionally consistent and statistically significant. We see that conservative and libertarian AI subjects are more tolerant across the board—the coefficients on the political preferences are generally increasing as we move down each column. Interestingly, the adversarial and scenario-change permutations do not affect the AI subjects' responses substantially, but the translated permutation does to some extent: the estimate for "Moderate" is far closer to zero. This is likely due to the different beliefs or definitions associated with so-called "moderate" political views outside the English-speaking world.

3.5 The minimum wage and discrimination in the labor market (Bertrand and Mullainathan, 2004; Horton, 2023)

Horton (2023) reports the results of a minimum wage experiment where employers were randomly assigned minimum wages in an online labor market: when applying for a job, applicants had to bid up to meet the employer's randomly assigned minimum wage. A key finding of this experiment was that there was little reduction in hiring but a substantial shift towards more productive workers, as proxied by past worker earnings and experience. The possibility of this labor-labor substitution margin had been noted in the literature, but had been difficult to pin down empirically in less controlled settings.

We explore the labor-labor substitution margin with AI subjects. We create scenarios where we ask AI subjects (employers) to select from pairs of applicants that vary in their

Table 2: Estimates of the effects of political beliefs and price hike levels on AI subjects' fairness assessments in four permutations of the Kahneman et al. (1986) experiment.

	Dependent variable:					
	Choice as Numeric from 1 (Very Unfair) to 4 (Completely Fair))					
	(1)	(2)	(3)	(4)	(5)	(6)
Δ Price	−0.010*		−0.010*	−0.009*	−0.014*	−0.013*
	(0.0005)		(0.0002)	(0.0005)	(0.0002)	(0.0002)
Socialist		−1.785*	−1.785*	−1.672*	−1.933*	−1.827*
		(0.038)	(0.029)	(0.056)	(0.027)	(0.028)
Leftist		−1.755*	−1.755*	−1.667*	−1.952*	−1.832*
		(0.038)	(0.029)	(0.056)	(0.027)	(0.028)
Liberal		−1.580*	−1.580*	−0.657*	−1.722*	−1.592*
		(0.038)	(0.029)	(0.056)	(0.027)	(0.028)
Moderate		−0.955*	−0.955*	−0.907*	−1.122*	−1.082*
		(0.038)	(0.029)	(0.056)	(0.027)	(0.028)
Conservative		−0.665*	−0.665*	−1.111*	−0.950*	−0.802*
		(0.038)	(0.029)	(0.057)	(0.027)	(0.028)
Constant	2.267*	3.092*	3.390*	3.479*	3.834*	3.669*
	(0.021)	(0.027)	(0.022)	(0.042)	(0.020)	(0.021)
Experiment	Baseline	Baseline	Baseline	Translated	Variations	Adversarial
Observations	2,400	2,400	2,400	1,198	2,400	2,400
R^2	0.167	0.598	0.765	0.577	0.825	0.797

Notes: This table reports regressions where the independent variables are the snow shovel price change relative to its original $15 price, and the AI subjects' political beliefs. We use "Libertarian" as the reference class. The dependent variable is the fairness assessment of the subjects, which we measure as a continuous variable ranging from 1 ("Completely Fair") to 4 ("Very Unfair"). Columns (1-3) show the results for the baseline experiment; Column (4) for the translated prompts experiment; Column (5) for the prompt variations experiment; and Column (6) for the adversarial prompt experiment. See Appendix A.4 for more details. Significance Indicator: *p<0.05.

years of work experience and requested wages. In our scenario, employers are hiring for the role of dishwasher, and we inform them of the typical wage for this role. The scenario in the prompt, fully detailed detailed in Appendix A.5, is:

> You are hiring for the role of "Dishwasher." You have 2 candidates.
>
> Person 1: Has 1 year(s) of experience in this role. Their name is `[name_1]`. Requests $[`wage_ask_1`]/hour.
>
> Person 2: Has 0 year(s) of experience in this role. Their name is `[name_2]`.

Requests $[wage_ask_2]/hour.

Who are you hiring? You must fill this role.

In our setup, Person 1 is the "experienced" worker and Person 2 the "inexperienced" worker. The inexperienced worker always asks for $13/hour, and the experienced worker asks for a wage between $12 to $17/hour. The employer is unknowingly and randomly assigned to either no minimum wage or a minimum wage of $15/hour. With a minimum wage, wage asks below the minimum wage threshold are forced up to $15/hour for both workers.

We also vary the applicants' names following the design of Bertrand and Mullainathan (2004).[10] In this paper, the authors studied racial discrimination in the labor market by sending out resumés with randomly assigned African-American or White-sounding names. They found that white-sounding names had 50% more interview callbacks. We can use the current experimental setup to explore whether the AI subjects demonstrate similar hiring biases by varying the candidates' names in addition to the employer-level minimum wage.

We repeat our experiment twice. In the first variation, we tell the employer that the typical wage ask for the position is $12/hour, and in the second, we do not provide any information about the typical wage.

We elicit the AI subjects' hiring decisions in 288 scenarios once with 17 different models at a temperature of 0.5.[11] The list of models used and the full prompt can be found in Appendix A.5. We ask the employer to always hire a candidate. Our empirical approach is to regress (i) the hired worker's wage and experience on an indicator for the minimum wage treatment status and (ii) the perceived race of the hired candidate being white on an indicator for the perceived races of the two candidates being different.

Table 3 reports the estimates of our regressions. All specifications include model fixed effects and standard errors are clustered at the model level. In Column (1), we see that without a reference wage, the minimum wage caused an increase of about $1.2/hour in the wage of the hired worker. This was expected because hiring was mandatory for the AI employer. However, in Column (2), we see that when the reference wage is provided, this effect grows by nearly 50%. When we regress the hired worker's experience on the presence of the minimum wage, the contrast between the two experiments is even more pronounced. In Column (3), without a reference wage, the minimum wage increased the years of experience of the hired worker by a little more than two weeks. But with a reference wage, Column (4) shows that, on average, the introduction of the minimum wage caused the employer to hire workers with roughly two and a half more months of experience. These results, increases in wage and experience, are consistent with Horton's findings. Finally, in Column (5), we

[10] The names are from the title of their paper: Emily and Greg (white) and Lakisha and Jamal (African-American).

[11] There are 6 possible wage asks for Person 1, 12 name combinations, 2 minimum wage conditions, and 2 reference wage conditions. $12 \times 6 \times 2 \times 2 = 288$.

see that, if anything, the AI employer is slightly more likely to hire the African-American candidate when the candidates' names are different. This is in contrast to Bertrand and Mullainathan (2004), albeit the effect is small and the experimental design is different.[12]

Table 3: Combined extension of Horton (2023) and Bertrand and Mullainathan (2004).

	Dependent variable:				
	Hire wage		Hire experience		Hire name race
	(1)	(2)	(3)	(4)	(5)
I(Minimum Wage)	1.234*	1.775*	0.056*	0.179*	
	(0.081)	(0.112)	(0.022)	(0.042)	
I(Diff. in Race of Name)					−0.017*
					(0.006)
Reference Wage	No	Yes	No	Yes	All Data
Model FE	Yes	Yes	Yes	Yes	Yes
Observations	2,448	2,448	2,448	2,448	4,896
R^2	0.221	0.492	0.199	0.200	0.001

*Notes: This table reports regressions where the independent variables are an indicator for whether a minimum wage was imposed on the AI employer, an indicator for whether a reference wage was given to the employer, and an indicator for when the racial association of the candidates' names is different. Observations are at the employer level and were sampled from 17 different models at a temperature of 0.5. The dependent variables are (1) and (2) the hired worker's hourly wage. (3) and (4) the hired worker's year(s) of experience, and (5) the racial association of the hired candidate's name. Columns (1) and (3) report the results of regressions without a reference wage, and Columns (2) and (4) report the results of regressions with a reference wage. Column (5) uses data across all reference wage conditions. The dependent variable in column (5) is one when the hired candidate's name is perceived as white and zero when African-American. See Appendix A.5 for more details on prompt construction. Significance Indicator: *p<0.05.*

4 Conceptual critiques

We develop a simple model and use it to examine conceptual critiques of AI experiments. We argue that criticisms of AI experiments often apply to human subjects research.

Consider a population of humans $\mathcal{H}$, a set of stimuli $\mathcal{X}$, and a set of responses $\mathcal{Y}$. The response y of a human h to a stimulus x is defined by the function $f_h : \mathcal{X} \to \mathcal{Y}$. The population's responses to a given stimulus x are a random variable $Y|x$, distributed as $P(f_h(x) = y)$ for all $h \in \mathcal{H}$ and $y \in \mathcal{Y}$.

Developers train an LLM with a dataset $(\mathcal{X}_T, \mathcal{Y}_T)$, comprised of stimulus-response pairs generated by a population $\mathcal{H}_T$. We define the LLM as a random variable $\hat{Y}|x$ whose distribution can be adjusted by conditioning on various stimuli (i.e., prompts).[13]

[12] We provide more specifications and robustness tests in the Appendix. In some, there is no advantage for applicants with African-American-sounding names.

[13] We refer to the elements of $\mathcal{X}$ as stimuli for humans and prompts to an LLM interchangeably.

4.1 Private training data and RLHF objectives

Arguably the strongest criticism of AI experiments is that for the most capable models to date, we lack precise details on the contents of the training data $(\mathcal{X}_T, \mathcal{Y}_T)$ and those humans $\mathcal{H}_T$ who generate it. Furthermore, the metrics and objectives that influence the distribution of $\hat{Y}|x$ during the RLHF phase of training are largely secret. Some argue that this opacity makes it challenging to understand exactly who or what we are studying with AI subjects.

One response to this critique 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 by helping us efficiently expand the frontiers of human knowledge. Suppose the primary use of AI experiments is to simulate experiments before trying them in the real world, and this method is adopted. Then, the debate is somewhat moot. Indeed, mounting evidence shows that this is already the case. Consider Hewitt et al. (2024), who successfully front-ran dozens of human subject experiments, demonstrating that the nearly 500 treatment effects estimated from LLMs and human populations were highly correlated ($r = 0.85$). Other similar examples abound (Li et al., 2024; Binz and Schulz, 2023; Brand et al., 2023).

But in direct response to the critique, we do have some insight into the contents of $(\mathcal{X}_T, \mathcal{Y}_T)$ and $\mathcal{H}_T$. For instance, OpenAI's training data includes millions of books, articles from major news organizations, the entirety of Wikipedia, Reddit and its outbound links, and the Common Crawl dataset, which spans a continually growing archive of over 250 billion web pages (Brown et al., 2020; OpenAI et al., 2024). These are just the publicly available resources—a lower bound for the true training corpus—all of which surely contain interesting examples of human behavior in countless contexts, dozens of languages (MetaAI, 2024), and from millions, if not billions, of people. To put in perspective how all-encompassing the training data sets are, some predict that we might soon exhaust the supply of human-generated data for training these models (Villalobos et al., 2024).

The objectives used during the RLHF process, however, largely remain largely unknown. Developers closely guard their human feedback evaluation metrics. What we do know is that the most capable LLMs, like GPT-4o and Claude-Sonnet-3.5, are optimized towards being "helpful," "harmless," and "honest" (Heikkilä, 2023; Bai et al., 2022; Ouyang et al., 2022). While people are often observed displaying such traits, they are certainly not a panacea of human behavior.

As such, the future of AI experiments may well lie with open-source models, where both the training data and the RLHF objectives are known. When possible, we fully endorse their use. While open models are not yet as advanced as the most capable models and not viable for many AI experiments, they are improving rapidly. They will be a better option in the future and will sidestep the challenges posed by closed, proprietary systems.

Until then, existing models remain valuable—immensely so. For one, they often generate reliable and intriguing results, results that can be used to assess assumptions, explore parameter spaces, and ideate at scale. Second, we can take some solace in the fact that companies keep their RLHF processes secret not only from researchers but also from one another. This mutual secrecy creates a form of independence between models. While each company may employ slightly idiosyncratic training processes on top of human-generated data, their approaches remain unknown and unused by their competitors. Consequently, when we run an experiment with multiple independent models, we can expect these individual quirks to average out in aggregate and that results are not simply an artifact of some unknown training process. That is, $\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\hat{Y}_m|x = E[Y|x]$ for a large enough set of indpendently optimized models $\mathcal{M}$ trained on similar distributions of underlying human behavior. We saw an example of this approach in the Horton (2023) experiment where our results were robust across many models: the presence of a minimum wage caused an increase in the wage and experience of the hired worker. It is easy to use the same approach with any other AI experiment to alleviate concerns of secret RLHF processes. Especially with the software tools we introduce in the next section.

4.2 Unrepresentative data

A more specific criticism of AI experiments is that their training data consists solely of text generated by the humans who choose to create public writing $\mathcal{H}_T$. If the researcher's population of interest $\mathcal{H}_I$ differs from $\mathcal{H}_T$, then the response distribution of $\hat{Y}|x$ could also differ substantially from $Y_{\mathcal{H}_I}|x$. Of course, this would be a problem for the fidelity of AI experiments. In the language of our model, a researcher may be interested in the distribution of the average effect τ of changing x to x' on the response distribution of $Y_{\mathcal{H}_I}|x$, where $\tau = Y_{\mathcal{H}_I}|x - Y_{\mathcal{H}_I}|x'$. The critique is that if a researcher runs an AI experiment with these stimuli to estimate $\hat{\tau} = \hat{Y}|x - \hat{Y}|x'$, then $E[\hat{\tau}]$ may not equal $E[\tau]$.

One advantage economists have in using LLMs is they tend to pose questions that place few demands on the sample. Most sets of people for $\mathcal{H}_T$ will do. We do not think of demand curves sloping downward as a "Western, Educated, Industrialized, Rich, and Democratic" phenomenon but rather as a result of rational goal-seeking that nearly all humans engage in. Similarly, 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 that the distribution of $Y|x$ is generally unaffected by the choice of $\mathcal{H}$ for most of their research questions.[14] More generally, much of social science is concerned not with the precise measure of some level but with the direction and magnitude of causal effects (Horton

[14] Snowberg and Yariv (2021) show that the correlations between lab and non-lab populations' behaviors are similar when participating in the same experiments.

et al., 2011).

The representativeness problem also rests partly on the notion that LLMs' responses are a weighted average, that every human in $\mathcal{H}_T$ impacts every response. This is not entirely correct. At a high temperature, LLMs are more like random number generators than estimators. If an LLM were pre-trained on data from millions of people (without RLHF), each selecting one random number from the unit uniform distribution, such that $\mathcal{X}_T = \{\text{"Generate a number"}\}$ and $Y_{\mathcal{H}_T}|x \sim [0,1]$, the model would not always respond with a number near the average. Instead, it would be almost equally likely to return any number between zero and one, i.e., $\hat{Y}|x \sim U[0,1]$.

When pre-training LLMs on more complicated distributions, the conditional nature of these models makes this stochastic property useful. Consider the case where the "true" human behavior when generating numbers is $Y_{\mathcal{H}}|x \sim U[0,1]$, but a researcher uses an LLM pre-trained on response data $Y_{\mathcal{H}_T}|x \sim U[0,1] + N(0,1)$. Suppose also that the model has been fine-tuned so that if prompted with the suffix "from true human behavior," it returns a number drawn from $U[0,1]$. For that model, $\hat{Y}|$ *"Generate a number"* $\sim U[0,1]+N(0,1)$, but $\hat{Y}|$ *"Generate a number from true human behavior"* $\sim U[0,1]$—i.e., true human behavior. Of course, no "from true human behavior" suffix exists, but prompting to generate particular conditional distributions of responses might be sufficient for a given research question. Argyle et al. (2022) make this point in their perfectly titled paper "Out of one, many." There is not a single LLM response type but rather a model capable of being conditioned to take on different personas that respond realistically, even for those not perfectly represented in $\mathcal{H}_T$.

Fortunately, many people are at least somewhat represented in the training data. More than 66% of the world's population—5.3 billion people—have access to the internet, the main source of text in the training corpora for these models. Such text is generated by a far more diverse set of people than those who participate in lab experiments.

And even if imperfect, the demands of representativeness in the social sciences have always depended on the research question. If the research question is "How do U.S. Presidents incorporate CIA intelligence estimates into decision-making?" one needs an extraordinary sample; if the research question is "Do humans have physical mass?" anyone will do. Most social science questions are somewhere in between. Ultimately, it is incumbent upon the researcher to make the case that the sample is appropriate for the question at hand.

4.3 Stated preferences

Another criticism related to the training data is that it consists of the text people choose to produce rather than the actions they actually take. That is, $\mathcal{Y}_T$ is a set of stated, not revealed preferences, and economists have historically taken a dim view of the economic content of mere statements rather than behaviors.

However, the data used to train LLMs is not millions of lines of people lying about their reservation values in bargaining scenarios. Much of the text publically available on the internet 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.[15] The training corpus likely even contains information about collective actions, such as aggregate purchasing decisions, which are readily available from most online retailers in the form of average ratings and the number of ratings from verified customers for specific products.

Whatever the exact content of the responses in $\mathcal{Y}_T$ may be, they surely include human decisions important for answering economic questions. And, with time, $\mathcal{Y}_T$ will become even more interesting. If they have not already, developers of these models will likely train them on individual consumer purchasing decisions across a vast set of products. The largest and most advanced developers have already announced formal partnerships with companies that generate vast amounts of such consumer data.[16]

4.4 Are these just simulations?

A common objection to AI-based experiments is that these are just agent-based models. Such simulation-based approaches have had a limited impact on economics. However, there is a critical difference between agent-based models and *homo silicus*: the amount of control the researcher has in developing the model.

With traditional agent-based models, the researcher is both judge and jury. The agents are programmed from scratch, and then their behavior is observed. They are not programmed to answer the question, "What would *homo economicus* do?" But rather "What would [*this model that does what we tell it to do*] do?" Understandably, the former is more scientifically interesting than the latter—it is much harder to control the result. 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 agent-based models, the process of training LLMs is not under the researcher's direct control. We do not get to select all of $(\mathcal{X}_T, \mathcal{Y}_T)$, nor do we have input on the functional form when estimating the distribution of $\hat{Y}|x$. Even the developers of these models struggle to control their behavior (Bowman, 2023). However, we can—as we have

[15] One of the author's fathers (Horton) 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 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.

[16] OpenAI and Anthropic, the developers of the two most advanced LLMs at the time of writing, have formal partnerships with Microsoft and Amazon, respectively. They have also received billions of dollars in funding from these companies.

shown—influence their responses with endowments of beliefs, political commitments, experiences, and so on.[17] But we are still constrained by the underlying model that determines their behavior, not our direct programming.

4.5 What counts as an observation

Foundational LLMs, like GPT-4o, Claude-Sonnet-3.5, and Llama-3-70B, are not fine-tuned to any particular language application. They have no fixed persona. Instead, as we have shown, they can be conditioned to play different agents via prompts. With a slight abuse of notation, we can consider persona-specific details as additions to any prompt x. For example, in the Kahneman et al. (1986) experiment, we had the agent answer "as" a libertarian, a socialist, a moderate, and so on, where the base prompt x asked for a judgment of fairness. 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 with LLMs, we are not limited to such simple lab-based practices. Hewitt et al. (2024) endow GPT-4 agents with demographic characteristics—both physical and ideological—and get responses that match the results from dozens of novel human subjects experiments.

Furthermore, single agents often respond stochastically.[18] $\hat{Y}|x$ has a distribution for any x and we can approximate this distribution by querying the model many times.[19] Comparing conditional distributions between a prompt x and another similar prompt x' can be informative. If we find that $Var(\hat{Y}|x) > Var(\hat{Y}|x')$, this can help uncover relationships between features that we might not expect. Like in the Kahneman et al. (1986) experiment, one might think that a Leftist and Liberal would have similar fiscally left-wing views on price gouging. Yet for the \$5 price hikes, the Liberals are uniform in their response, while the Leftist is far more evenly split between options. Maybe this is because of the varying definitions of the word conservative. Either way, it is a difference that might be worth exploring further.

Fine-tuning is another viable option for training existing models to take on particular personas. Many such models already exist. Researchers have trained some models to represent specific populations (Valicenti et al., 2023) and to provide better cognitive models of humans (Binz and Schulz, 2023). Some services offer personal LLMs trained on a user's text footprint (e.g., emails, tweets, texts, etc.) to help with certain tasks or respond on their behalf.[20] One could imagine a very near future in which someone could have a personally fine-tuned

[17] Park et al. (2023) endow a community of 25 AI agents with identities and complex memory systems, allowing them to simulate a small town. Without instruction, these agents go on dates, make new friends, discuss previously shared experiences, and gossip about each other.

[18] Unless the temperature is set to zero.

[19] Depending on the API interface and the model, it is sometimes possible to elicit the probability distribution of the responses directly.

[20] https://www.rewind.ai/

LLMs, a $\hat{Y}_h|x$ that approximates $f_h(x)$ to respond on their behalf when convenient. These models could be used as subjects in experiments, appropriately addressing privacy concerns, compensation, and so on.

4.6 Prompt p-hacking

The technical requirements are minimal to run an AI experiment. This low barrier to entry, however, comes with the risk of researchers iterating through various prompts to find one that generates a desired response. To illustrate, suppose there are a set of prompts $\mathcal{X}_S$, all of which are similar in meaning to x—a prompt of interest. And a researcher views all the prompts in $\mathcal{X}_S$ as relevant to her research question. If she is determined to demonstrate a particular result, say y, the researcher could iterate through many $x' \in \mathcal{X}_S$, querying the model with each prompt, until she gets the desired response.[21] Then, she could report only the prompt x' that generated said result.

Similar concerns exist in traditional experimental work, where researchers data-mine results to find significant relationships (i.e., p-hacking). A partial solution is pre-registration, This is an inadquate option for AI experiments for two reasons. First, AI experiments are often exploratory, a trial run to see if a particular line of inquiry is worth pursuing or an assumption is reasonable. Forcing researchers to pre-register their ideas when they are first trying to ideate is counterproductive. Second, pre-registration is cumbersome and slow. It stands in direct contrast to the key benefit of using AI subjects, namely, the ability to run thousands of experiments in a matter of minutes.

A better solution is a robustness check. One way to do this is to require that researchers present their results under many different partially controlled permutations of the prompt. For example, in the Kahneman et al. (1986) experiment, we increased the model's temperature and repeated the experiment many times, translated the prompt into different languages and then back to English, and varied the context of the experiment. When we compared the results across the permutations, we found the patterns of interest held. The results were robust, although we did not know this would happen beforehand.

All these manipulations varied the data-generating process in a way that is partially, but not entirely, in our control. They forced us to test and report many prompts in $x' \in \mathcal{X}_S$, but we did not get to choose exactly which. Such partially controlled variations can prevent researchers from easily cherry-picking results while ensuring the experiment maintains its focus.

[21] Similarly, she could use a single prompt but query the model many times at a high temperature until the desired response is produced.

4.7 Memorization

With billions of parameters, one might think LLMs are simply repeating back to us something they have already "read" somewhere in their massive training corpus. Put more simply, the estimation of $\hat{Y}|x$ has overfit on $(\mathcal{X}_T, \mathcal{Y}_T)$ to the extreme. First and foremost, for some research questions, this critique is irrelevant. The data used to train models like GPT-4o or Claude-Sonnet-3.5 often cuts off months, if not years, before the present day. Kozlowski et al. (2024) exploit such time lags to forecast polarization in during COVID-19 using an LLM trained on data up to January 2020. They demonstrate that much of the political divide in response to the pandemic was fairly predictable across party lines.

But even if the data were up-to-date, this view of pure memorization is not correct. It is inconsistent with their tendency to hallucinate or make up new "facts" (Li et al., 2023). Moreover, this is at odds with several findings presented in Section 3. For instance, in the Charness and Rabin (2002) experiment, the distributions of the responses from the persona-less AIs were not even close to those of the human subjects.

Another more striking example is the Bertrand and Mullainathan (2004) experiment. Despite the paper being widely discussed in the literature, we found very little evidence of discrimination. If any, the discrimination was slightly advantageous for candidates with African-American-sounding names. Although inconsistent with the original paper, this result is not surprising. The distribution of $\hat{Y}|x$ is not solely determined by the training data.[22] The model's potential responses are also influenced by the fine-tuning phase, during which many of the leading developers of LLMs explicitly aim to mitigate racial bias (Heikkilä, 2023; Bai et al., 2024, 2022).[23] More simply put, the presence of specific text in the training data does not guarantee that it will dictate the model's response to a given prompt. And even if it does condition on memorized information, this can be useful, if not preferable.[24]

In the Kahneman et al. (1986) experiment, the LLM's responses closely replicated the original results, assuming a left-leaning sample. While the result for this specific scenario may have been memorized, it is unlikely that the model memorized responses for all possible combinations of political views, price hikes, and framings, as these permutations were not present in the original paper.

One explanation for the "reasonable" relationships between changes in political views, prices, and fairness is that the model has appropriately learned correlations between these variables from the training data in other contexts. Suppose the original experiment is in the

[22] We can always try to check whether references to any piece of work are in the training corpus by querying the model for information about said piece of work, which we did extensively for Bertrand and Mullainathan (2004). Although an imperfect method, it can be informative.

[23] Although these debiasing techniques are far from perfect (Bai et al., 2024) and exactly how they work is proprietary.

[24] As a reminder, the human feedback phase changes the model's conditional probability distribution of the pre-trained model.

training corpus, and we consider it a noisy version of the truth for more-left-leaning humans evaluating a $5 price gouge. Such data may serve as a useful prior for the LLM. By combining this information with other references to politics and price gouging in the training data, the LLM generates a reasonable response for a never-before-seen prompt. Within the framework of our model, if there is some stimulus $x' \notin \mathcal{X}_T$, then the distribution $\hat{F}|x'$ may be a convex combination of $\hat{F}|x$ for many stimuli in $x \in \mathcal{X}_T$ that hold information relevant to x'. In this case, the x's are the questions we ask the LLM with different political leanings and price hikes that were not in the original experiment, whereas the xs may be the stimuli in the training data that discuss the original paper, politics, price gouging, and so on. Indeed, it could even be seen as a negative if the LLM ignored the original Kahneman et al. (1986) experiment, assuming that these results are informative in describing human behavior.

4.8 Performativitiy

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 because they have read about them (MacKenzie, 2007). The critique being that $\hat{F}|$*Prompt instructing model to do what humans do* $\not\sim$ *What humans do,* but instead $\hat{F}|$*Prompt instructing model to do what humans do* $\sim$ *What human theories say humans do.*

However, performativity is not a new problem for social science. Researchers have long been dealing with experimenter demand effects where perceived cues about what human subjects "should" do in experiments can affect their behavior (Zizzo, 2010). Often, reframing the experiment or using a subtler intervention can help, both of which are easy to implement with LLMs. For example, Argyle et al. (2022) reduces performativity by providing demographic attributes in the first person to LLMs, improving the fidelity of their responses. Brand et al. (2023) compare direct and indirect strategies for eliciting willingness to pay from LLMs for various goods—another promising approach.

The fact that an LLM does not "know" these theories is also useful to us because it will not always try to apply them. These models often behave like students who have memorized information for an exam but struggle to apply that knowledge in new contexts. Consider auctions, one of the most well-studied phenomena in economics. Its rich theoretical literature makes precise predictions in various settings. When asked to describe different auctions and these predictions, GPT-4 will recite them in textbook language. Yet when asked to predict the results of simple simulated auctions using LLM agents, it fails to do so with any accuracy—even when the results of the simulations almost perfectly match the theory (Manning et al., 2024). This failure to apply relevant information when responding to a prompt—a common human oversight (Handel and Schwartzstein, 2018)—makes the performativity critique less important.

More generally, the impact of performativity on the realism of *homo silicus* is not inherently positive or negative. LLMs are fine-tuned to "perform" with respect to the criteria used to build the reward model and the distribution of responses in the training corpus. If the appropriate criteria and training data are selected to align the model's performativity with some desired type of behavior, economists can better use the LLM to answer their research question of interest.

4.9 Causal inference

A final critique of AI experiments is that even with perfect randomization, all else is not equal when comparing treatment groups. The issue is that changing one factor often alters others, which may drive any observed effects. An experiment may aim to isolate a single cause, but in reality, it is capturing the impact of multiple changes. However, such a problem is a question of external validity—whether a given randomized manipulation generalizes to environments with other downstream variables, variables that may influence outcomes in ways the original controlled experiment did not capture. Of course, external validity is very much a problem for any experiment, not just those with AI subjects. Both for human and AI subjects, the problem is most clearly illustrated by example.

Let us define the prompt, $x_{Hire}(w_1, w_2, r)$ such that $x_{Hire} : \mathbb{R}^2 \times \{\mathbb{R}, \emptyset\} \to \mathcal{X}$, which was parameterized by the wage asks of the applicants w_1, w_2 and the presence of a typical wage for reference r.[25] The experimental design iterated over combinations of w_1 and w_2, with an externally imposed minimum wage forcing up the wage asks when they were below \$15/hour. We first ran the experiment without a reference wage $r = \emptyset$. When we repeated the experiment with a reference wage $r = \$12$, the effect of the minimum wage on both the wage and experience of the hired worker increased.

The critique is an explanation for this difference. In the experiment without a reference wage, the LLM may impute values for the reference wage as a variable function of the wage asks $\hat{r} : \mathbb{R}^2 \to \mathbb{R}$ learned during training. And if $\hat{F}|x_{Hire}(w_1, w_2, r)$ changes with r, then treatment assignment could influence the potential outcomes of the hired worker's wage and experience. Instead of the experiment generating the LLM's response distribution being $\hat{F}|x_{Hire}(w_1, w_2, \emptyset)$, the experiment is actually $\hat{F}|x_{Hire}(w_1, w_2, \hat{r}(w_1, w_2))$. Conversely, when $r = \$12$, the reference wage is specified, so there is no value to impute. Indeed, Figures A.15 and A.16 show that the AI typically hires the more experienced worker when the reference wage is specified but often chooses the less experienced worker when it is not.

Now, suppose we repeated the Horton (2023) experiment with people—with and without a reference wage. It seems plausible, even likely, that if unspecified, the hiring manager would presume some sort of reference wage for applicants based on their wage asks. As with

[25] For simplicity, we ignore the names of the applicants.

the LLM experiment, human subjects may impute values for unspecified information that depend on the treatment and their past experiences. This imputation may affect the outcome of interest. This is actually a common critique of Bertrand and Mullainathan (2004). Some have argued that the names of the applicants were as much associated with class as with race, and that the perceived socioeconomic status of the applicants—based on their names—was the driving factor in the differences in callback rates (Simonsohn et al., 2024).

Stepping back, both the AI and the hypothetical human subjects' experiments are perfectly randomized. We can always estimate the downstream causal effects of such exogenous manipulations. But if we hope to generalize these results to other settings, to claim that the results are externally valid, we must be careful; unspecified information may be relevant to the causal relationships under investigation.

In human subjects experiments, we deal with questions of external validity in a few ways. If the experiment is executed in the field (e.g., in a labor market with real employers), it operates in the exact environment of interest, and all possible information is, by definition, specified. This is not the norm. Many experiments are much less naturalistic—they are conducted in a lab or a lab-in-the-field and surely only specify some of the information that might be relevant to the outcome of interest. In these cases, it is incumbent upon the researcher to justify the external validity of the results. They can do this through theory, robustness checks, additional experiments, etc. We can use AI experiments to help here, to think deeply about the potential sparsity in our experiments and how it might affect the results—just as we do with human subjects experiments. We can run robustness checks and, most importantly, run additional simulations quickly at almost no cost AI experiments can provide a powerful check on the assumptions experimental economists may often take for granted and help us identify possible threats to external validity.

5 Using *homo silicus*

In this final section, we provide high-level guidance on how to best incorporate AI experiments into economic research. But before doing so, we briefly explain the form and content of our recommendations. Specifically, we avoid making any guarantees when AI experiments will or will not be informative proxies for human behavior. We do this for two reasons. First, LLMs are advancing at a breakneck pace. An easy way to seem foolish is to claim that version N of some LLM cannot perform a task, but then, three months later, version $N+1$ performs that task better than any human ever. They will continue to improve in ways we cannot possibly predict.

Second, by all theoretical accounts, LLMs and deep neural networks more generally, should not be so capable (Ananthaswamy, 2024). These models are overparameterized, often completely memorizing their training data. Yet, when trained for long periods, they perform

shockingly well on unseen data. Computer scientists cannot explain this phenomenon; deep neural networks should not be able to improve their performance on unseen data simply by overfitting their training data to the extreme. This is not to say we cannot probe an LLM's inner workings. Recent literature is making great strides in parsing the relationships learned by sub-networks of LLM parameters (Templeton, 2024). We are just unable to make guarantees.

Some might find it problematic for our purposes that no reliable theory explains how neural networks behave from the bottom-up. On the contrary, it motivates our work. To do economics—even behavioral economics—we do not have to study neurons and parts of the brain. Indeed, we do economics, at least in part, because we cannot yet understand human behavior from firing synapses alone. Economics often informs neuroscience, as observations of human behavior have yielded valuable insights into the inner workings of our minds (Camerer et al., 2005).

LLMs are similar to human subjects in this respect. We must study them empirically to know for sure how they will behave. Fortunately, LLMs are far less complex and opaque than the human brain.[26] With an LLM, we know the contents of large portions of its training data and have a sense of its architecture, and with open models, we could know everything about how it was constructed. All of these design choices are geared towards generating useful approximations of a massive corpus of human-generated text, text that we know reflects human thoughts and behaviors by definition. In contrast, human brains are the product of millions of years of evolutionary pressures, of which we can often only speculate, and the interactions of complex cultural processes.

With this in mind, our guidance is designed to avoid rigid absolutes or guarantees. Instead, we focus on reproducibility and recommend practices that maximize the external validity of results. We aim to build on what we do know about LLMs, while fully acknowledging that these models, like human subjects, are complex and their behavior may evolve in unpredictable ways.

5.1 When AI Subjects can be useful

Evaluating assumptions and Exploration. First and foremost, these experiments help us evaluate our assumptions and explore possible behavioral phenomena. Take the Horton (2023) experiment, where variation spilled down in an unexpected manner. It was only when we ran many AI experiments that we discovered providing a reference wage so drastically affected the results. Some labor economists may presume they would never make such a

[26] Many believe the human brain is "the most complex thing we have yet discovered in our universe" (Ackerman, 1992). Parsing a deep neural network is simple in comparison.

mistake, but many others are not so savvy.[27] And while this exact problem might not affect human subjects, it is certainly plausible. Either way, if we were to run a similar experiment with people, as in Horton (2023), we could benefit from this improved interrogation of our assumptions.

In short, AI experiments allow us to consider what "could" matter more thoroughly. This is not to say it will always be right, but given how difficult it can be to do exploratory work with human subjects and the general "reasonableness" of the AI agents' responses, they offer a valuable resource for exploring parameter spaces and piloting studies. This could be as simple as testing different wordings of experiments, improving instructions, fixing errors, or controlling for a diverse set of attributes in the agents to evaluate potential confounders. At the minimum, AI experiments used in this way will force researchers to lay out their experiments in full, not unlike a pre-analysis plan, which will help them interrogate their design.

Power calculations and piloting Often, power calculations are arbitrary. This is especially true when designing experiments without pre-treatment data (i.e., most experiments). In these cases, we must make many subjective predictions about effect sizes and covariates, and researchers are notoriously bad at such speculation (Gandhi et al., 2024). An obvious and simple use case for AI experiments is to simulate experiments to improve power calculations. Of course, there will still be plenty of researcher discretion—AI experiments will not be perfect stand-ins for pre-treatment data, but they can offer a substantial improvement over no data.

No human equivalent. As noted in Aher et al. (2022), physically impossible or ethically unacceptable experiments also offer an exciting opportunity for exploration. For example, in the Charness and Rabin (2002) social preferences experiment, we exogenously manipulate agents' preferences towards equity, efficiency, or self-interest. It is not possible to perfectly manipulate a human's preferences in this way, and yet, we can generate valuable insights by doing so with AI agents.[28]

Another context where it is often difficult, if not impossible, to run human subjects experiments is a more macro-environment with many agents. For example, Törnberg et al. (2023) simulates a Facebook-like social media platform using LLMs instead of humans. On different versions of the platform, they vary the news-feed algorithm. They find that when an algorithm highlights content liked by AI agents endowed with opposing political views, the agents have more constructive conversations. While social media companies have the

[27] Indeed, several authors on this paper identify as labor economists and did not predict the importance of a reference wage.

[28] Technically, we can instruct human experimental subjects to "act" according to some set of traits or preferences they do not intrinsically hold, but it is unclear if such practices are effective.

technical capabilities to conduct similar experiments on their platforms, they are generally reluctant to do so. When they do, they are even less likely to share the results unless they paint the company in a positive light. Even if large social media companies were willing to design and implement such experiments publicly, there are still many research questions that would be unethical study despite being extremely important.[29]

More generally, AI experiments offer an exciting avenue to explore large-scale simulations. The AI agents do not need to know the specific details about the expected macro-level behavior of the environment. All the agents need to do is respond appropriately given their endowed traits and the structure of the environment, and then we can observe the equilibrium that follows. Even when economists can study the dynamics of group behaviors, say, in negotiations, public goods games, and the like, the samples are extremely limited in size. Often, a single observation requires multiple humans and a significant time investment. AI experiments have no such limitations.

As replacements for human subjects. Researchers are increasingly questioning whether AI agents can replace human subjects in social science experiments (Cui et al., 2024; Dillion et al., 2023). However, some argue there is much to lose from such replacement (Messeri and Crockett, 2024). That we may limit the scope of scientific inquiry by doing away with human subjects entirely.

We do not fundamentally object to the idea of using AI experiments as replacements for human subjects. If AI agents can consistently respond in ways analogous to human subjects, both in direction and magnitude, at a fraction of the cost, it is hard to argue against their utility. It is not as if the status quo is perfect. Empirical work in the social sciences is often flawed. Theory often fails to produce the results of experiments (Watts, 2017), and problems with replicability, p-hacking, and straight-up fraud abound. With AI experiments, social scientists can finally study at-scale phenomena that have previously been accessible only from small lab experiments and observational data. From a cost-benefit perspective, or even the practical perspective of reliably generating useful and replicable results, a future where AI experiments are capable replacements for human subjects in many domains of social scientific inquiry is appealing. And given the rapid pace at which these models are advancing, such a future may not be so far off.

For now, however, it is not yet clear enough when AI experiments fail to justify replacing human subjects in general. Although, there are certain situations where we are more confident in their responses, as we will discuss. As such, unless a researcher has a compelling reason (or is solely studying the behavior of the LLM), we recommend AI experiments be empirically validated when making strong inferences about human behavior. Of course, use

[29] Take, for example, the question of whether social media use causes an increase in youth suicide (Haidt, 2024). Imagining an experiment that could ethically test such a hypothesis with real people is difficult.

and validation are subject to the research question at hand. And certain research questions are far more amenable to exploration with AI experiments. Indeed, as the capabilities of these models continue to improve, we expect the number of research questions for which AI experiment are a suitable replacement for human subjects will increase. Simply because AI subjects are poor proxies for one phenomenon does not mean they are poor proxies for all. As we continue to explore the boundaries of their behavior, it will become increasingly clear when they are reliable.

Training data and experiment format. We expect AI experiments to perform better when drawing on behaviors effectively captured in text, such as negotiation tactics, consumer preferences, or survey responses. They are also likely to be more reliable in experiments where an individual's stated preferences align with their revealed preferences. For instance, if individuals post a resumé on an online labor market, they may withhold their most unattractive qualities. In contrast, in anonymous forums like Glassdoor reviews or Reddit comments, people are more likely to provide honest, unfiltered opinions, leading to more accurate reflections of their true preferences.

LLMs may perform poorly in tasks requiring physical interaction or non-verbal cues, such as experiments involving facial expressions, body language, or sensory reactions. It is only possible to capture descriptions of these behaviors in the training data. However, as multimodal models that process images and video in addition to text become more common, this limitation may attenuate.

5.2 Best-practices

Many models. As we showed in the AI version of the Samuelson and Zeckhauser (1988) experiment, resutls can vary significantly across models. In other instances, like the Horton (2023) experiment, the results held across dozens of models. Running experiments with many different models helps address concerns about model-specific quirks or biases. There is an analogy, albeit imperfect, to the law of large numbers. As we discussed in Section 4, even if the idiosyncracies of a single model are not representative of human behavior, the average responses from many independently trained models might be. By averaging results across models, we can reveal insights that are more likely to be reliable. As such, we generally recommend researchers run experiments with many different models developed by independent organizations.

The one caveat to this recommendation is open-source models. If a researcher has access to a capable open-source model, running experiments on many models may not be necessary. That depends on whether the objectives used to train the reward model during RLHF are

interesting in their own right or explicitly designed to be representative of some set of human preferences.

Permutations. A key concern when running AI experiments is their sensitivity to small variations in prompts, which can lead to markedly different responses (Mohammadi, 2024). This is problematic for two reasons.[30] For one, the results of any given AI experiment may not be robust. Second, as we have already discussed, a researcher could, knowinlgy or unknowlingly, adjust the prompt for an AI experiment until it generates the desired result.

To address this, we encourage economists to test many different partially controlled variations of any given AI experiment. Permutations can include the examples from Kahneman et al. (1986), like translating the experiment to other languages, running the experiment many times at higher temperatures, generating alternatives in a different context, or simply rewording the experiment in many different ways. However, there is no singular "correct" way to permute an experiment, and there are surely many more permutation strategies than those we have outlined in this paper. Ultimately, it is incumbent upon the researcher to convince both themselves and others that their findings are robust for whatever claim they are trying to make. When a result is robust to many permutations, external validity for any purpose is more plausible.

Calibration. When appropriate and possible, we recommend that researchers "calibrate" their AI agent samples as we did for the two-stage dictator game in the Charness and Rabin (2002) experiment. By generating samples of AI agents that matched moments with a known human sample in one experiment, we were better able to approximate the results of a separate human subject experiment. To effectively calibrate across multiple traits, we only need to make sure that there is sufficient variation in the agents' responses. Otherwise, if their responses are colinear, an infinite number of weights would solve the optimization problem.

Importantly, we need not be limited to simple preference descriptions to construct our AI agent samples. We can use any traits: complex preferences, demographic information, personality types, etc. We suggest selecting traits based on the research goal at hand. If we are simply interested in matching the results of some human subjects' experiments as closely as possible, we may want to use salient demographic features. However, if we are trying to explore a theoretical framing of some phenomena, like types of preferences in Charness and Rabin (2002), we might not care as much about perfectly matching results.

More generally, calibrating agents is similar to the traditional use of synthetic controls in quasi-experimental work (Abadie and Gardeazabal, 2003). However, with AI agents, we can be creative with the traits with which we generate the control agents. We are not limited to

[30] Inconsistent responding is not an intrinsic problem of AI experiments per se. Humans often do the same, but to a lesser extent.

weighted combinations of existing data.

5.2.1 Expected Parrot Domain-Specific Language

References

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132.

Ackerman, S. (1992). *Discovering the Brain*. National Academies Press (US), Washington, DC. Foreword.

Aher, G., Arriaga, R. I., and Kalai, A. T. (2022). Using large language models to simulate multiple humans. *arXiv preprint arXiv:2208.10264*.

Ananthaswamy, A. (2024). *Why Machines Learn: The Elegant Math Behind Modern AI*. Penguin Publishing Group.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., and Wingate, D. (2022). Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2209.06899*.

Bai, X., Wang, A., Sucholutsky, I., and Griffiths, T. L. (2024). Measuring implicit bias in explicitly unbiased large language models.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback.

Bertrand, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013.

Binz, M. and Schulz, E. (2023). Turning large language models into cognitive models.

Bowman, S. R. (2023). Eight things to know about large language models.

Brand, J., Israeli, A., and Ngwe, D. (2023). Using gpt for market research. *Harvard Business School Marketing Unit Working Paper*, 23-062.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A.,

Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33.

Camerer, C., Loewenstein, G., and Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, 43(1):9–64.

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

Cui, Z., Li, N., and Zhou, H. (2024). Can ai replace human subjects? a large-scale replication of psychological experiments with llms.

Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Friedman, M. (1953). *The Methodology of Positive Economics*. University of Chicago Press, Chicago.

Gandhi, L., Duckworth, A. L., and Manning, B. S. (2024). Effect size magnification: No variable is as important as the one you're thinking about—while you're thinking about it. *Current Directions in Psychological Science*.

Haidt, J. (2024). *The anxious generation: How the great rewiring of childhood is causing an epidemic of mental illness*. Random House.

Handel, B. and Schwartzstein, J. (2018). Frictions or Mental Gaps: What's Behind the Information We (Don't) Use and When Do We Care? *Journal of Economic Perspectives*, 32(1):155–178.

Heikkilä, M. (2023). How openai is trying to make chatgpt safer and less biased. *MIT Technology Review*. Accessed: 2024-02-11.

Hewitt, L., Ashokkumar, A., Ghezae, I., and Willer, R. (2024). Predicting results of social science experiments using large language models. *Preprint*. *Equal contribution, order randomized.

Horton, J. J. (2023). Price floors and employer preferences: Evidence from a minimum wage experiment. *Working paper*.

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

Jahani, E., Manning, B. S., Zhang, J., TuYe, H.-Y., Alsobay, M., Nicolaides, C., Suri, S., and Holtz, D. (2024). As generative models improve, people adapt their prompts.

Kahneman, D., Knetsch, J. L., and Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *The American economic review*, pages 728–741.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291. Accessed: 18 Oct. 2024.

Kahneman, D. and Tversky, A. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323. Accessed: 18 Oct. 2024.

Kozlowski, A. C., Kwon, H., and Evans, J. (2024). In silico sociology: Forecasting covid-19 polarization with large language models.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *science*, 302(5649):1364–1368.

Li, J., Cheng, X., Zhao, X., Nie, J.-Y., and Wen, J.-R. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Li, P., Castelo, N., Katona, Z., and Sarvary, M. (2024). Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science*, 0(0):null.

Lucas, R. E. (1980). Methods and problems in business cycle theory. *Journal of Money, Credit and Banking*, 12(4):696–715.

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

Manning, B. S., Zhu, K., and Horton, J. J. (2024). Automated social science: Language models as scientist and subjects. Technical report, NBER. Accessed: 2024-03-12.

Messeri, L. and Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58.

MetaAI (2024). Introducing llama 3. Accessed: 2024-08-29.

Mohammadi, B. (2024). Explaining large language models decisions with shapley values. Available at SSRN: `https://ssrn.com/abstract=4759713` or `http://dx.doi.org/10.2139/ssrn.4759713`.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T.,

Zhuang, J., Zhuk, W., and Zoph, B. (2024). Gpt-4 technical report.

Oprea, R. (2024). Decisions under risk are decisions under complexity. *American Economic Review*. Available online at `https://www.aeaweb.org/journals/aer/forthcoming`.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Raman, N., Lundy, T., Amouyal, S., Levine, Y., Leyton-Brown, K., and Tennenholtz, M. (2024). Steer: Assessing the economic rationality of large language models.

Samuelson, W. and Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of risk and uncertainty*, 1(1):7–59.

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

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., et al. (2024). The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Simonsohn, U., Nelson, L., and Simmons, J. (2024). Data colada: The funnel plot is invalid. `https://datacolada.org/51`. Accessed: August 28, 2024.

Snowberg, E. and Yariv, L. (2021). Testing the waters: Behavior across participant pools. *American Economic Review*, 111(2):687–719.

Templeton, A. (2024). *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic.

Törnberg, P., Valeeva, D., Uitermark, J., and Bail, C. (2023). Simulating social media using large language models to evaluate alternative news feed algorithms.

Valicenti, T., Vidal, J., and Patnaik, R. (2023). Mini-gpts: Efficient large language models through contextual pruning.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., and Hobbhahn, M. (2024). Position: Will we run out of data? limits of LLM scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*.

Watts, D. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, 1:0015.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13:75–98.

A Code and prompts for experiments

A.1 Code for Charness and Rabin (2002)

Figure A.5: Code for Charness and Rabin (2002) experiment to generate Figure 1.
Notes: **TO ADD!**

Figure A.6: Code for Charness and Rabin (2002) experiment to generate Figure 2.
Notes: **TO ADD!**

A.2 Code for Oprea (2024)

Figure A.7: Code for Oprea (2024) experiment to generate Figure 3.
Notes: **TO ADD!**

A.3 Code for Samuelson and Zeckhauser (1988)

Figure A.8: Code for Samuelson and Zeckhauser (1988) experiment to generate Table 1.
Notes: **TO ADD!**

A.4 Code for Kahneman et al. (1986)

Figure A.9: Code for Kahneman et al. (1986) experiment repeated 25 times at temperature 0.
Notes: **TO ADD!**

Figure A.10: Code for Kahneman et al. (1986) experiment with 10 alternative versions
Notes: **TO ADD!**

The ten alternative versions of the experiment are as follows:

Figure A.11: Code for Kahneman et al. (1986) experiment with 10 adversarial versions

Notes: **TO ADD!**

The ten adversarial versions of the experiment are as follows:

Notes: **TO ADD!**

Figure A.12: Code for Kahneman et al. (1986) repeated once for each model

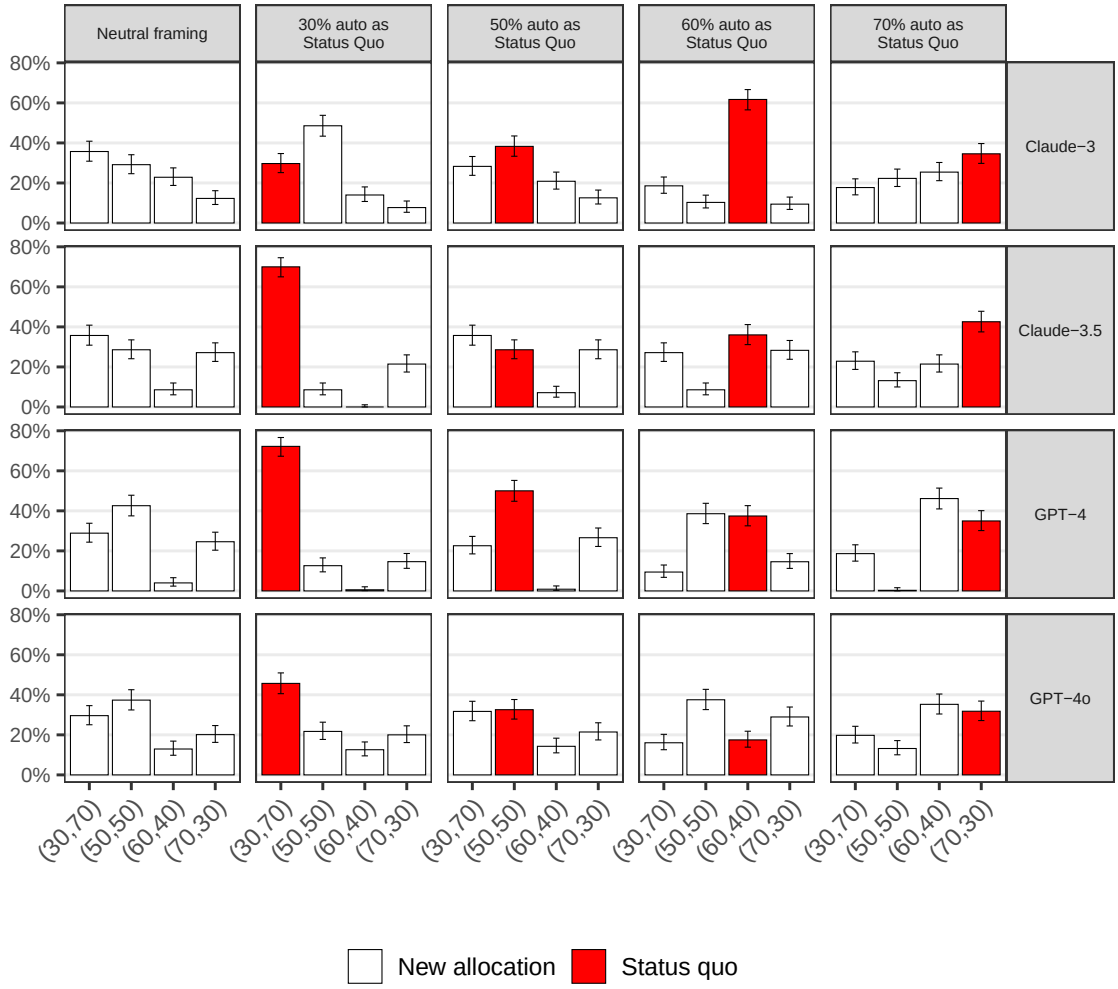
A.5 Code for Horton (2023)

Figure A.13: Code for Horton (2023) experiment used to generate Table 3

Notes: **TO ADD!**

B Additional Figures and Tables

Figure A.14: Replication of status quo experiments from Samuelson and Zeckhauser (1988).



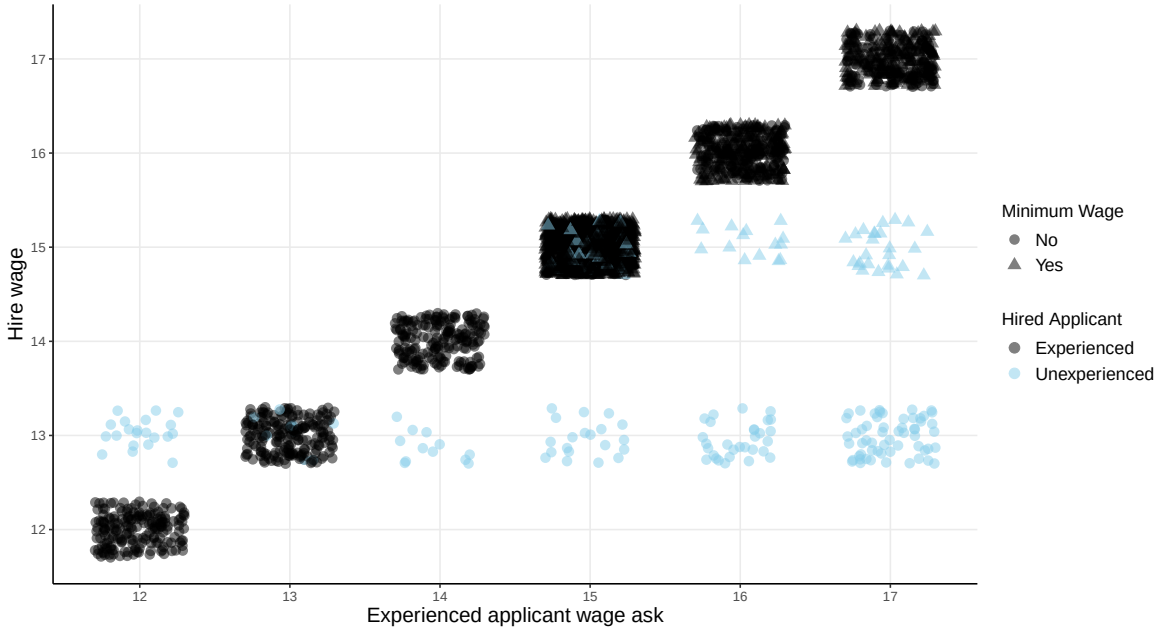
Notes: This figure reports the results of our experiment replicating the status quo experiments from Samuelson and Zeckhauser (1988). The x-axis indicates the possible allocations (auto%, highway%), and the y-axis is the proportion of AI experimental subjects choosing each option. Error bars represent 95% Multinomial Wilson confidence intervals and include uncertainty estimates for responses with zero counts. The leftmost column reports the answers in the neutral framing, and the other columns report the answers for each labeled status quo framing. See Appendix **??** *for more details on prompt construction.*

Table 4: kkt with interactions

	Dependent variable:			
	Choice as Numeric from 1 (Completely Fair) to 4 (Very Unfair)			
	(1)	(2)	(3)	(4)
Δ Price (ΔP)	−0.001*	−0.002	−0.009*	−0.007*
	(0.001)	(0.001)	(0.001)	(0.001)
Socialist	−1.609*	−1.458*	−1.871*	−1.708*
	(0.034)	(0.073)	(0.034)	(0.034)
Leftist	−1.560*	−1.459*	−1.898*	−1.716*
	(0.034)	(0.073)	(0.034)	(0.034)
Liberal	−1.282*	−0.461*	−1.577*	−1.374*
	(0.034)	(0.073)	(0.034)	(0.034)
Moderate	−0.467*	−0.581*	−0.795*	−0.638*
	(0.034)	(0.073)	(0.034)	(0.034)
Conservative	−0.270*	−0.828*	−0.793*	−0.599*
	(0.034)	(0.073)	(0.034)	(0.034)
ΔP × Socialist	−0.006*	−0.007*	−0.002*	−0.004*
	(0.001)	(0.002)	(0.001)	(0.001)
ΔP × Leftist	−0.007*	−0.007*	−0.002*	−0.004*
	(0.001)	(0.002)	(0.001)	(0.001)
ΔP × Liberal	−0.010*	−0.007*	−0.005*	−0.008*
	(0.001)	(0.002)	(0.001)	(0.001)
ΔP × Moderate	−0.017*	−0.011*	−0.011*	−0.015*
	(0.001)	(0.002)	(0.001)	(0.001)
ΔP × Conservative	−0.014*	−0.010*	−0.005*	−0.007*
	(0.001)	(0.002)	(0.001)	(0.001)
Constant	3.131*	3.274*	3.709*	3.485*
	(0.024)	(0.052)	(0.024)	(0.024)
DGP	Temp.	Translate	Vars.	Adv.
Observations	2,400	1,198	2,400	2,400
R^2	0.812	0.595	0.843	0.829

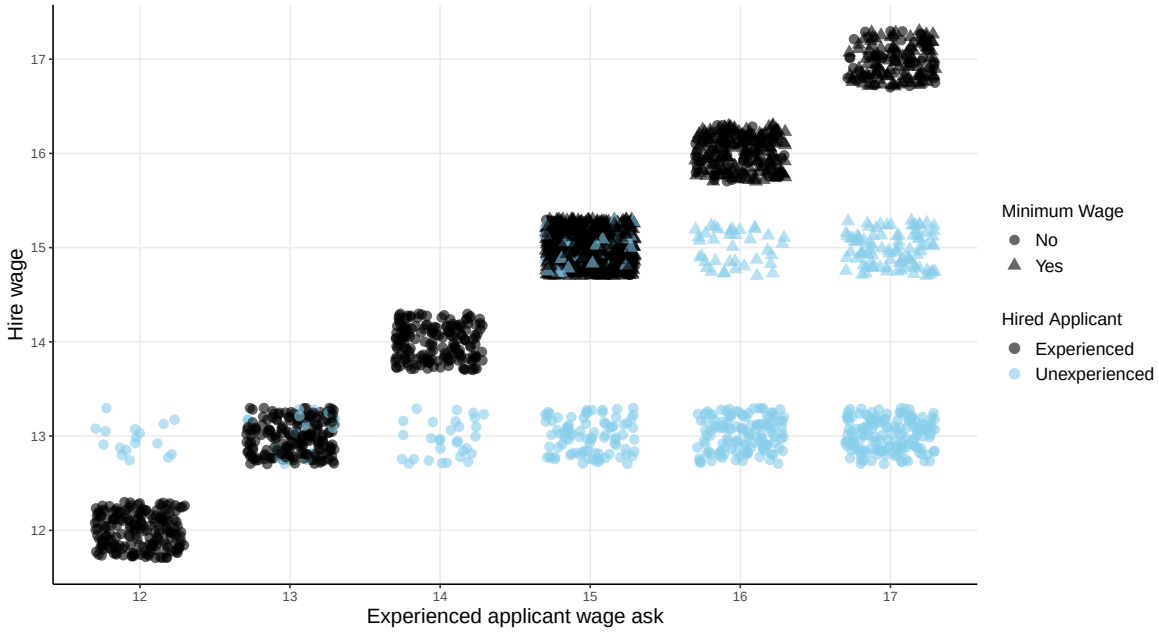
Notes: asdf

Figure A.15: No Reference Wage



Notes:

Figure A.16: With a Reference Wage



Notes: temperature 1, only 5x (instead of 10 because kept breaking)