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

John J. Horton
MIT & NBER

Apostolos Filippas
Fordham

Benjamin S. Manning
MIT

December 8, 2025

Abstract

We argue that newly-developed large language models (LLMs), 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), Samuelson and Zeckhauser (1988), Oprea (2024b), and Horton (2025), show qualitatively similar results to the original, and when they differ, it is often generative for future research. We discuss potential applications, conceptual issues, and why this approach can inform the study of humans.

*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, Charlotte Siegmann, 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. All authors contributed equally to this work. Author contact information, code, and data are currently or will be available at https://benjaminmanning.io. Horton is an economic advisor to Anthropic; all authors have a financial interest in www.expectedparrot.com. In preparing this paper, the authors utilized generative AI models extensively as tools to assist with editing and evaluation. The authors retain full responsibility for all content and conclusions presented herein.

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*, endows it with different resources, preferences, information, and so on, subjects it to various economic scenarios, and then deduces its behavior. This behavior can then be compared to that of actual humans in (b)-type research.

In this paper, we argue that large language models (LLMs)—because of how they are trained and designed—can be thought of as implicit computational models of humans—a *Homo silicus*. These models can be used the same way economists use *Homo economicus*: they can be given endowments, put in scenarios, and then their behavior can be explored—though in the case of *Homo silicus*, through computational simulation, not a mathematical deduction.[1] This is possible because LLMs can now respond realistically to a wide range of textual inputs, giving responses similar to what we might expect from a human (Aher et al., 2022; Bowman, 2023).

Like all models, any particular *Homo silicus* is wrong, but that judgment is separate from a decision about usefulness. Of course, each *Homo silicus* is a flawed model and can often give responses far away from what is rational or even sensical. Yet unlike mathematical economic models with fixed input parameters, AI agents can be flexibly applied to simulate any scenario expressed in natural language. Such flexibility means that agent fidelity to human responses is not fixed: the underlying instructions or "prompts," which control their behavior can always be adjusted—even in ways that incorporate conventional economic theory. Indeed, recently developed methods can dramatically improve agent fidelity, and as we will show, the process of better matching human responses can itself be generative.

We consider the reasons why AI simulations might be helpful in understanding actual humans. The core of the argument is that LLMs—by nature of their training and design—are (i) computational models of humans and (ii) possess a great deal of social information. For (i), the creators of LLMs have designed them to respond in ways similar to how a human would react to prompts (Bai et al., 2022; Ouyang et al., 2022)—including prompts that are economic scenarios. For (ii), these models capture two rich sources of social information. The first is latent in the form of economic activities, decision-making heuristics, and

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

common social preferences that the LLM implicitly learns from its training data. These massive corpora contain 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. Second, LLMs have also been trained on the majority, if not all, of published social scientific research. This comprises the facts, theory, and empirical results codified by researchers; welfare theorems, prospect theory's account of loss aversion, the income returns to education, and so on.

It is essential to note that the sources of social information in the LLM's training data are not perfect ingredients for simulating human behavior. For instance, directly applying insights from published research may enhance the realism of simulations. But it can also make brittle simulations appear robust because models may simply regurgitate problematic and memorized results (Sarkar and Vafa, 2024; Ludwig et al., 2025). We address such tensions along with other conceptual issues related to LLMs' training corpora, opacity, and the generalizability of simulations.

While conceptual issues are important, what ultimately matters is whether these AI simulations are practically valuable for generating insights. As such, the majority of this paper is devoted to presenting and discussing five simulated experiments. We selected these experiments because they are simple to describe and implement. They also have clear qualitative results that can be compared to the AI outcomes.

We begin with 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, and are asked to state whether doing this was fair or unfair. Illustrating a benefit of AI agents, 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 large gouging is viewed more negatively and that endowed political views matter, with predictable effects—right-leaning AI agents are more sanguine about gouging. The effects are robust to dozens of permutations of the original experiment.

Next, we simulate the simple unilateral dictator games from Charness and Rabin (2002). We show that endowing AI agents 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 it is self-interested causes it to select allocations that maximize narrow self-interest. Interestingly, without any endowment, some models choose efficiently while others tend to be selfish. We use these results to generate new samples of agents that respond increasingly like human subjects in more complicated two-stage games.

In a study of framing, we next present the AI agents with a decision-making scenario introduced by Samuelson and Zeckhauser (1988). In the paper, the respondent has to allocate a federal budget between highway safety and car safety. The original results show that humans are subject to a status quo bias, preferring budget options that are presented as the status quo. We recapitulate this result by endowing AI agents with baseline views about the relative importance of car or highway safety. We then put those agents through various scenarios, with each possible allocation taking a turn as the status quo. Some LLMs exhibit a strong bias towards any option presented as the status quo, while others barely do.

In the fourth set of simulations, we explore the experiments in Oprea (2024b), where participants were asked to supply certainty equivalents for a series of binary choice lottery-type questions. This research challenges conventional interpretations of prospect theory by showing that the classic "fourfold pattern" of risk attitudes emerges even when participants face no risk. Rather, the paper posits that the pattern emerges because cognitively constrained participants are heuristically responding to complexity. We simulate the experiment across an expanded set of questions with five different models serving as the AI agents. Even though the paper is recent and contrasts decades of research, our simulated results provide suggestive evidence in support of Oprea's complexity-driven interpretation.

Finally, we explore a hiring scenario motivated by Horton (2025), which shows in a field experiment that employers facing a minimum wage substitute towards higher-wage workers. We create a scenario where an employer is trying to hire a worker as a dishwasher and faces a collection of applicants that differ in their experience and wage asks, which are chosen randomly. The AI agent makes an experience-wage tradeoff. We then impose a minimum wage that forces applicants bidding below that minimum to bid up. We run these scenarios across LLMs with varying capabilities. On average, the results are directionally consistent with Horton: the minimum wage increase causes a shift in the AI's hiring of more experienced applicants and higher hourly wages.

The first draft of this paper featured versions of the now-deprecated GPT-3 as the underlying models in our experiments. Results from these now woefully out-of-date models appear in Appendix A. We now employ dozens of contemporary LLMs—both open-source and proprietary—and introduce methodological improvements that can increase realism and test robustness of the simulations. However, as newer models are inevitably released, even our updated experiments will become dated. Because this pattern will continue to repeat, and access to proprietary systems is uncertain, we emphasize the importance of simulations that are easy to design, replicate, and scale.

This paper is now accompanied by an open-source Python package that enables researchers to replicate our experiments on new models or explore their own ideas with minimal technical overhead, often using only a fraction of the code required for the simulations in the initial draft.[2] Using this package, all experiments in this paper are now button-push reproducible with almost any LLM, and the software is actively maintained as methods, models, and prompts improve. We include links to code, prompts, and tutorials for all simulations throughout the manuscript. Whether or not researchers choose to employ this software, or something else, it is essential—and now feasible—for AI simulations to have the features we have highlighted in the package: to be nearly costless to replicate, share, and update. These are precisely the features social scientists should desire from any experimental method.

Ultimately, we care about the behavior of actual humans, and so results from AI experiments will still require empirical confirmation (Ludwig et al., 2025).[3] As such, what is the value of these experiments? The most obvious use is to pilot experiments in silico first to gain insights. They could cheaply and easily explore a parameter space, test whether behaviors seem sensitive to the precise wording of various questions, generate data that will "look like" the actual data, and inform power calculations.[4] 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. As relationships are identified, they could guide actual empirical work—or interesting effects could be captured in more traditional theory models. Indeed, by endowing AI agents with prompts grounded in social science theory, their predictive capabilities can improve dramatically (Manning and Horton, 2025)—precisely what researchers seek from a good theory. 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 and identify promising patterns.

Since the first draft of this paper in 2023, an extensive and rapidly growing literature testing AI agents has emerged. This work, including thousands of experiments, now spans every social scientific discipline (Brand et al., 2023; Chang et al., 2024; Shah et al., 2025; Wang et al., 2025; Fish et al., 2025; Qian et al., 2025), with some studies building on these simulations to study humans in ways previously infeasible (Manning et al., 2024; Xie et al., 2025; Bybee, 2025). In a growing number of cases, AI agents put in simulations of the relevant settings can even predict human behavior in never-before-seen surveys and experiments (Li et al.,

[2] See `https://pypi.org/project/edsl/` to download and `https://docs.expectedparrot.com/` for documentation.

[3] Our focus is not on using LLMs as experimental subjects or studying their behavior directly, though both are valuable research directions (Alberti et al., 2019; Westby and Riedl, 2022; Binz and Schulz, 2023), but on using them as tools to understand humans.

[4] There is an analogy to protein-folding, where simulations identify proteins later found in the real world (Kuhlman et al., 2003).

2024; Hewitt et al., 2024; Tranchero et al., 2024; Park et al., 2024; Binz et al., 2025; Suh et al., 2025).

Given this rapid expansion, the contributions of this paper are now twofold. First, we aim to help economists effectively incorporate AI simulations into their methodological toolkit, utilizing the most up-to-date practices, and ensuring these tools remain useful as models and methods continue to evolve. Second, the relative—and original—contribution of this paper is drawing the connection to the common research paradigm of economics and the role a foundational assumption (or model) like rationality plays in research. LLM experimentation is more akin to the practice of economic theory, despite superficially looking like empirical research.

The remainder of this paper is structured as follows. In Section 2, we describe and analyze five simulated experiments. Section 3 discusses why LLMs can often predict human behavior and when they are likely to be useful. In Section 4, we address conceptual critiques related to the use of LLMs for social science. The paper concludes in Section 5.

2 Experiments

We conduct five experiments with LLMs. We recapitulate and extend four laboratory experiments and then one field experiment from the economics literature. Our use of the term "recapitulate" in lieu of "replicate" is intentional: we do not view LLM experiments as direct copies of the originals. The goal is not to just reproduce results. Rather, we aim to see how simulations behave in empirical scenarios where we already understand the analogous human behavior. Then, by contrasting with the originals, we demonstrate key features of simulations that researchers can manipulate to better explore new ideas.

The first experiment demonstrates how we can use LLMs to easily extend experiments and search for new insights. In the second, we show that LLMs respond reasonably to prompts and follow instructions effectively. We also show how to leverage these responses and conventional economic theory to construct samples of AI agents that respond increasingly like human subjects. The third experiment shows that LLMs exhibit well-established human decision-making biases. The fourth shows that LLMs can be used to study subtle elements of human decision-making and induce experimental variation in new ways. The fifth experiment demonstrates that researchers can use AI agents to study realistic field-like decision-making scenarios.

For readers unfamiliar with LLMs, we provide a brief primer in Appendix B. Other introductory resources include Bowman (2023) and Ouyang et al. (2022), while Zhao et al. (2025) provide a comprehensive and

technical survey of the field. In addition, Korinek (2023) and Charness et al. (2025) examine how LLMs can be used across the full spectrum of experimental economics research.

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

Kahneman et al. (1986) conducted experiments designed to study people's fairness perceptions in market transactions. In their price gouging scenario, they asked subjects 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) Very Unfair, (2) Unfair, (3) Acceptable, (4) Completely Fair.

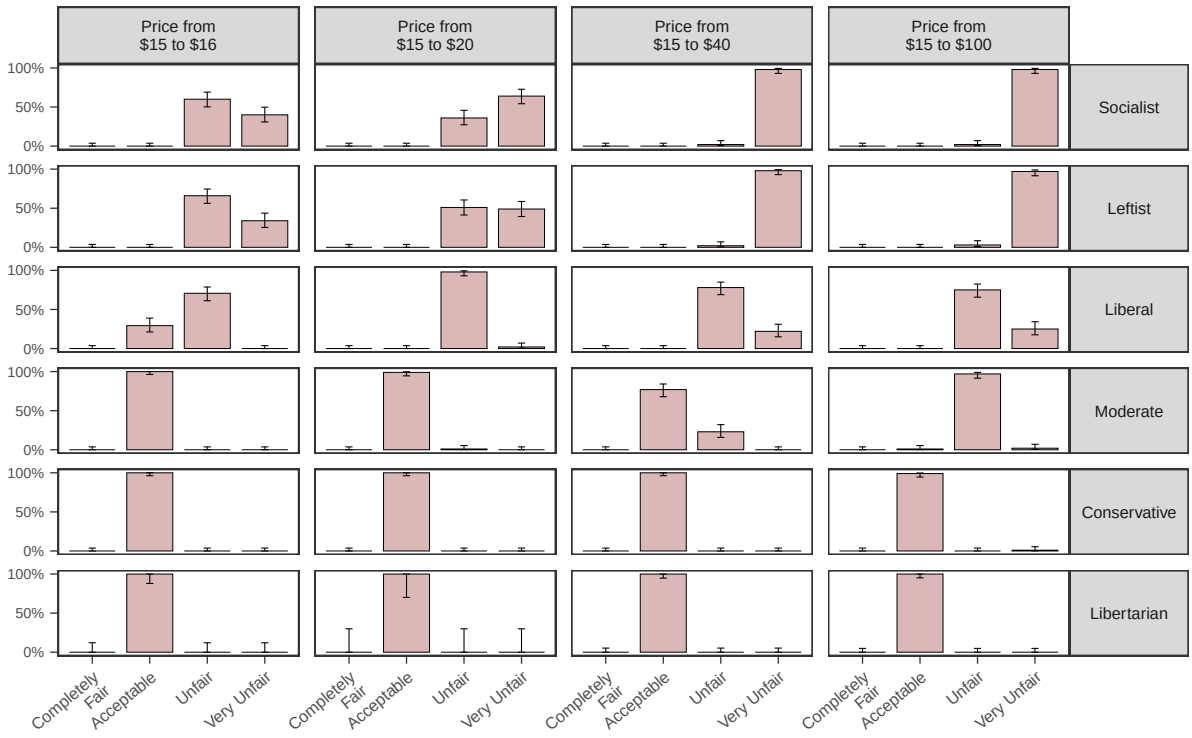
About 82% of subjects responded either "Unfair" or "Very Unfair."

We use AI agents to explore two natural follow-up questions: (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 do so, we create AI agents endowed with "personas" or "types," by prepending a description of the persona in the prompt. We endow AI agents with political views ranging from "You are a socialist" to "You are a libertarian," and ask each to independently 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.

Figure 1 reports the results of our experiment and details on how we constructed the prompts can be found in the notes via the linked Jupyter notebook. To get a sense of the structure and content of the code, see Figure A4, which provides a screenshot of the notebook in the appendix. We provide such notebooks, which are accompanied by extensive documentation and instructions, for all experiments throughout this paper in their respective figure or table notes. Columns of this figure correspond to different price increase scenarios and rows to different political bents of the AI agents. The x-axis shows the possible responses, and the y-axis shows the proportion of agents who choose each option.

The AI agents' political beliefs affect their views about the price hikes: Socialist and Leftist agents 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 a few of the Conservative AI agents judge the

Figure 1: Recapitulation and extension of Kahneman et al. (1986)'s price gouging experiment.



Notes: This figure reports the results of our experiment recapitulating 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 agents' political leanings. The x-axis shows the possible responses, and the y-axis shows the proportion of AI agents choosing each option. Error bars represent 95% multinomial Wilson confidence intervals. AI agents use GPT-4 with the temperature set to 1. We collect 100 responses for each AI agent persona and scenario combination. See `www.expectedparrot.com/content/0e39f555-fe34-4911-b516-46638d1640ca` for more details on prompt construction.

$100 price hike as "Unfair" or "Very Unfair". The answers of the AI agents are also fairly consistent within political view-price gouge combinations: in about half of the panels, we get the same response in all 100 observations.

Kahneman et al. report little about the composition of their original study, making it difficult to know which combination of political views is representative. They do note that the sample consisted of Canadians from Toronto and Vancouver, with an approximately even split between males and females. While we have no way to know the political leanings of the participants, it is plausible that the original sample was skewed left-leaning, given their urban origins. Directionally consistent with this speculative view and the results in Kahneman et al., AI agents endowed with left-leaning views always judge the original price hike to be "Unfair" or "Very Unfair."

2.1.1 Testing generalizability and possible memorization

One concern with our fairness simulations is that the AI agents may have memorized both the original study and other texts describing the relationships between political labels and fiscal views. If true, our results might be brittle—an artifact unique to the snow shovel example that does not generalize to other related scenarios. We next show how to better evaluate the robustness of AI simulations.

We conduct three follow-up sets of simulations of the price-gouging study. First, we translate the original prompt into 10 different languages and run the experiment in each language.[5] 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 the accompanying Jupyter notebook in the Figure 2 notes. One example is:

> 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) Very Unfair, (2) Unfair, (3) Acceptable, or (4) Completely Fair.

Others include a bookstore selling novels after an author announces a signing event, a convenience store selling sunscreen right before a heatwave, etc. 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 agents responding differently than in our original experiment. Specifically, we prompted the LLM to make the 10 adversarial versions "as diverse as possible such that the same combination of variables will get people to answer differently." The adversarial examples are the same format as the alternative examples and are also provided in the same notebook. We conduct all three of these additional experiments 10 times for each of the 10 prompt variations within every political view-price gouge combination, with the temperature parameter set to 1. With four price gouges, six political leanings, and 10 variations for each permutation, we collect 2,400 observations for each of the additional experiments.

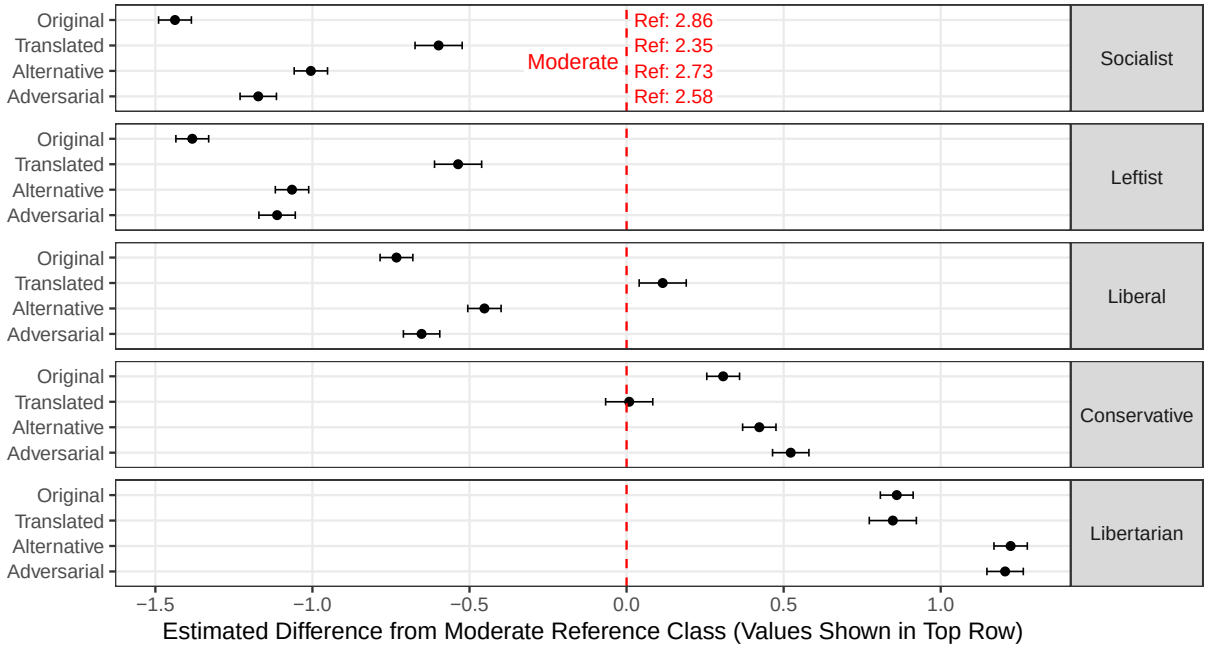
To measure responsiveness to price gouging, we regress the AI agents' choices, measured as a continuous variable ranging from 1 ("Very Unfair") to 4 ("Completely Fair"), on the price change and the AI agents' political endowments. Figure 2 reports dummy regression coefficients for different political endowments across the four experimental permutations, using "Moderate" as the excluded reference category. Each row represents a different political belief, with the x-axis showing estimated differences from the reference class.

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

Values for the reference OLS constant are shown in red in the top panel for each experimental permutation.

For interpretation, the translated permutation's Moderate intercept of 2.35 indicates that, on average, AI agents endowed with moderate political views in this permutation rate price increases as roughly one third of the way from "Unfair" to "Acceptable." The Socialist coefficient of -0.60 for the translated permutation in the top panel means Socialist AI agents, on average, rate price increases as one quarter of the way from "Unfair" to "Very Unfair" (2.35 + (-0.60) = 1.75).

Figure 2: Coefficients of the effects of political beliefs on AI agents' fairness assessments in four permutations of the Kahneman et al. (1986) experiment.



Notes: This figure reports regression coefficients for political belief indicators, with Moderate as the omitted reference group, controlling for price gouging. The dependent variable is the fairness rating (1 = Very Unfair, 4 = Completely Fair). Full specifications appear in Table A2. The x-axis shows estimated differences from the Moderates; the y-axis shows the experimental permutation. Each row corresponds to a different political belief. The center of the top panel reports the reference intercepts for each permutation—the mean fairness rating for Moderates. All other estimates are the difference from these values. See `www.expectedparrot.com/content/6b9ca070-bf21-4887-b68d-060c9177650f` for simulation code and all experimental variations.

As in the original simulations, agents with more right-leaning political orientations generally rate price gouging as increasingly fair across all permutations. This pattern is evident in the diagonal trend of coefficients from the upper left to the lower right. Notably, the adversarial and alternative examples yield strikingly similar results. We were unable to invert the relationship—agents with more left-leaning orientations never rated price gouging as fairer than those on the right. At a high level, the original pattern appears quite robust.

The translated permutation, however, exhibits a few small differences. Liberal AI agents show positive

coefficients, suggesting they are more right-leaning than the Moderate reference group. Conversely, the Conservative agents in this permutation appear nearly indistinguishable from the Moderates. This variation may reflect different definitions of "Conservative" and "Liberal" political views outside English-speaking contexts.[6]

Although not visualized here, higher price increases are viewed as increasingly unfair, consistent with results in Figure 1. Table A2 in the appendix reports the full specifications with these values. Across the four permutations, including the original, a $1 price hike causes between a 0.004 and 0.009 point decrease in the Likert fairness assessment.

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

Charness and Rabin (2002) ask experimental subjects to choose between allocations that involve trade-offs between efficiency and equity. We focus on their unilateral dictator games, where the dictator (Person B) chooses between two allocations of money between herself and the other player (Person A). For example, in the following game:

"Left": Person A gets 400 and Person B gets 600

"Right": Person A gets 700 and Person B gets 300,

the dictator chooses between the "Left" and "Right" allocations. We can write this game as:

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

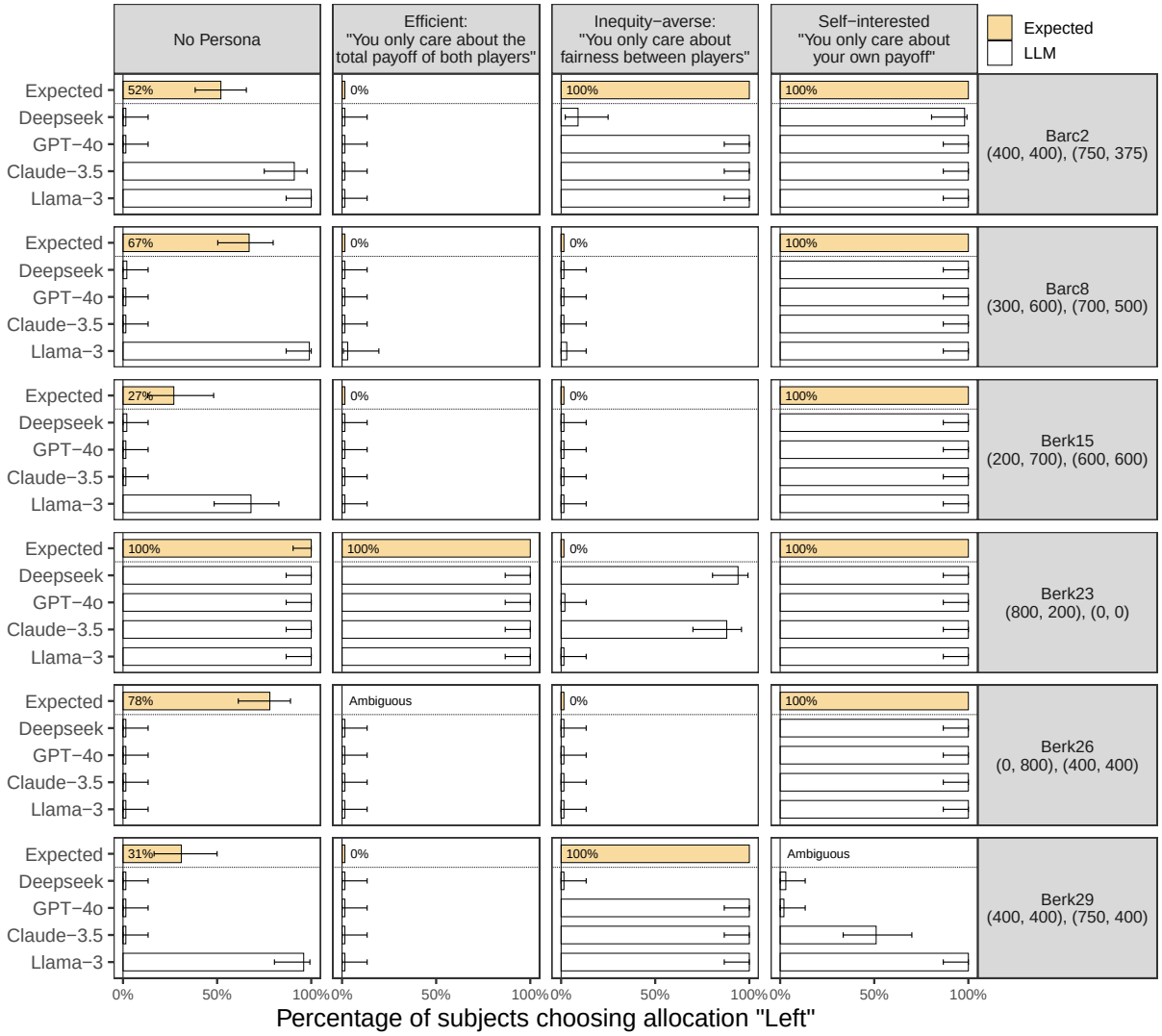
We recapitulate this experiment using AI agents. First, we prompt LLMs with descriptions for each game, and ask them to respond with their preferred allocation without any additional information. Second, we construct three AI agents with the following personas: (i) inequity-averse (*"You only care about fairness between players"*), (ii) efficient (*"You only care about the total payoff of both players"*), and (iii) self-interested (*"You only care about your own payoff"*). We have each AI agent play each game 100 times with the temperature set to one. We use GPT-4O, CLAUDE-SONNET-3.5, LLAMA-3-70B, and DEEPSEEK as the underlying models.

Figure 3 plots the results of our experiment with the notebook to run the simulations linked in the notes. Each column corresponds to a different persona, and each row to a different game. The x-axis shows the

[6]For example, in China: en.wikipedia.org/wiki/Conservatism_in_China.

fraction of agents choosing the "Left" option. The white bars show the AI agents' choices, and the gold bars show either the choice of the human subjects in Charness and Rabin (leftmost column) or the choice that a perfectly adherent AI agent 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, with the maximum total payout, would be to always select "Right."

Figure 3: Recapitulation of single-stage dictator games from Charness and Rabin (2002).



Notes: This figure reports the recapitulation of unilateral dictator games from Charness and Rabin (2002) with AI agents. The columns correspond to personas with which we endow AI agents, and the rows to different games. Each AI agent plays each game 100 times. The x-axis shows the percentage of agents choosing "Left." The y-axis shows both the AI agents' responses and the "Expected" proportion for each column's given persona. White bars depict the AI agents' choices, and gold bars depict either the choice of the human subjects in Charness and Rabin (leftmost column) or the choice that a perfectly aligned AI agent would make with the given persona (all other columns). Error bars report 95% Wilson confidence intervals. See `www.expectedparrot.com/content/6bc88957-c346-46a3-910a-e63ec2d77035` for simulation walkthroughs.

The choices of the persona-less AI agents are quite different from those of the human participants in

the original experiment, despite the fact that the original experiment is almost certainly included in the training data for all models (as is true for most well-known economics papers). These responses are also different across models, although consistent within: Llama-3-70B is seemingly more selfish, and Gpt-4o, Claude-Sonnet-3.5, and Deepseek are more efficiency-minded. Despite this baseline heterogeneity, AI agents endowed with personas respond consistently in line with those personas. Except for a few cases—such as Deepseek and Llama-3-70B failing to choose the inequity-averse option in the spiteful Berk23 scenario—efficiency-endowed agents generally chose the option with the highest combined payoff, inequity-averse agents selected the option with the smallest difference between payoffs, and self-interested agents maximized their own payoff.

2.2.1 Constructing a calibrated sample of AI agents

If we wanted to use these off-the-shelf LLMs without personas to predict behavior in new allocation games, the evidence above suggests their accuracy would be poor. Such agents might suffice for low-stakes applications—like stress-testing experimental designs or checking data pipelines—where fidelity to human distributions is not the primary concern. But if the goal is to employ AI agents as credible guides for empirical work—a far more valuable use case and promising application of AI simulations—then this lack of predictive reliability becomes a serious limitation.

By contrast, the AI agents endowed with personas exhibited high fidelity: they followed the instructions tied to their assigned type almost perfectly. This disciplined behavior provides a foundation for building calibrated samples of agents that can better predict human responses. We adopt the approach from Manning and Horton (2025), who show that by optimizing mixtures of theory-grounded agents to match human responses in one setting, one can improve their predictive power in other related settings. Here, "theory-grounded" means the personas are motivated by some theory that we generally understand, can clearly evaluate, and reasonably expect to predict the relevant human behavior—like how fairness, self-interest, and efficiency agents reflect the original economic model laid out in Charness and Rabin.

We implement the approach and construct an optimized mixture of AI agents as follows. Intuitively, we choose the "recipe" of agent types—self-interested, inequity-averse, and efficient—whose combined choices best reproduce the aggregate choices of human subjects. Each persona is represented by a vector $v = (p_{Barc2}, \ldots, p_{Berk29})$, where p_g is the proportion of AI agents with this persona that chose "Left" in game g from Figure 3. For example, the "efficient" Gpt-4o AI agents only chose "Left" in the Berk23 scenario,

so they are represented by $v_{\mathrm{E}} = (0, 0, 0, 1, 0, 0)$; the population of Charness and Rabin is represented by $v_{CR} = (.52, .67, .27, 1, .78, .68)$. For each LLM, we then compute the weights $w = (w_E, w_I, w_S)$ that minimize the mean-squared error (MSE) $||w_E v_E + w_I v_I + w_S v_S - v_{CR}||^2$, subject to $w_E + w_I + w_S = 1$ and $w_E, w_I, w_S \geq 0$.[7] The optimal weights for each LLM are $(0.49, 0.00, 0.51)$ for Llama-3-70B, $(0.44, 0.00, 0.56)$ for Claude-Sonnet-3.5, $(0.37, 0.10, 0.53)$ for Gpt-4o, and $(0.44, 0.00, 0.56)$ for Deepseek. We use these weights to randomly sample personas for each LLM.

Having calibrated these mixture weights on the unilateral games, we now evaluate whether they generalize to new settings. We use the two-stage games from Charness and Rabin for this out-of-sample evaluation. These games have a different structure, but are theoretically similar to the unilateral dictator games in the sense that we might reasonably expect some combination of efficiency, fairness, and self-interest to influence human responses in both. If our calibrated agents can better predict responses in these new games than the persona-less LLMs, this provides evidence that they can generalize across structurally distinct data-generating processes.

In the sequential games, monetary allocations depend on both players' decisions. In the first stage, Person A chooses either a given allocation or to 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. For example,

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

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

We write this game as:

$$(500, 500)$$

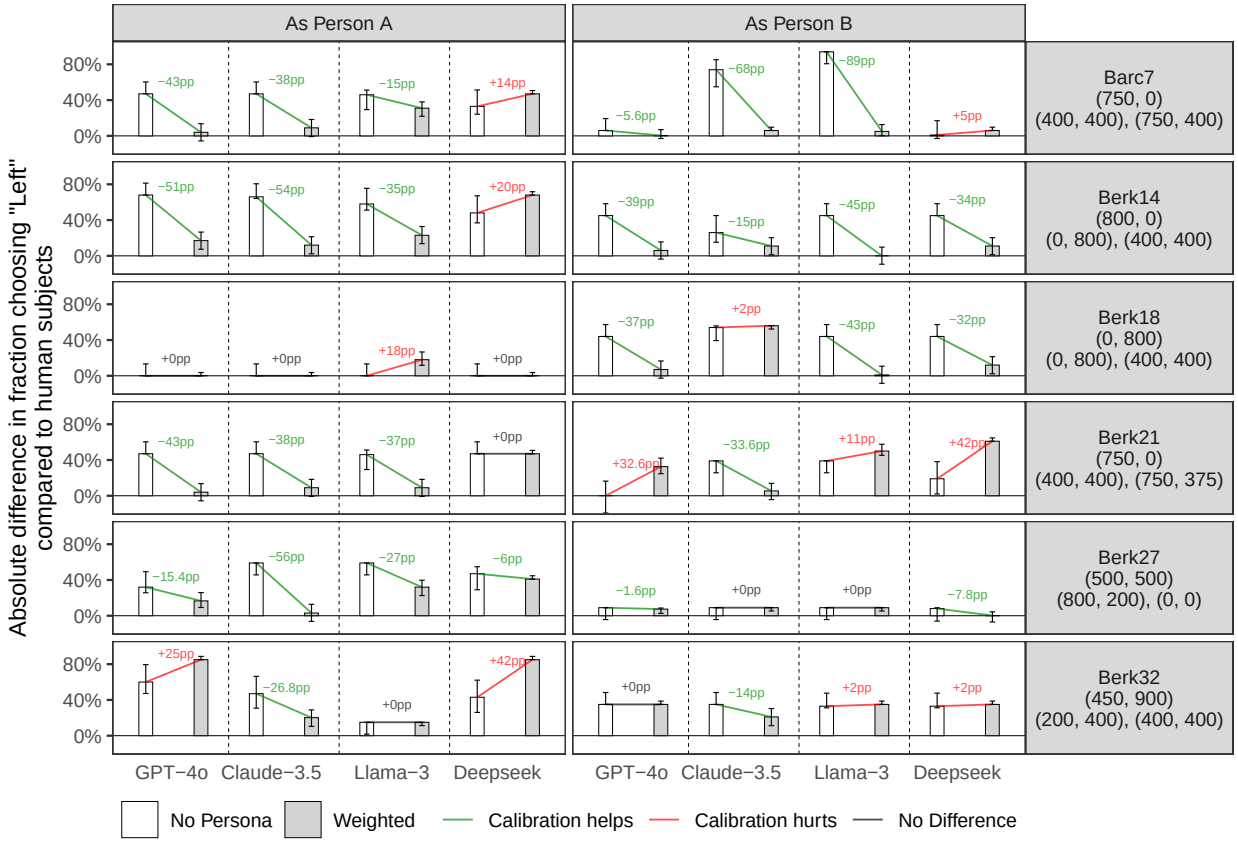
$$(400, 600), (700, 300)$$

We conduct our experiment as follows. For each LLM, we sample randomly 100 AI agents according to the computed weights w. Each AI agent plays each game scenario 5 times, both as Person A and Person B. We also have "control" persona-less AI agents play each game and role 100 times.

[7] Others have explored this mixture method (Leng et al., 2024; Xie et al., 2025; Bui et al., 2025).

Figure 4 plots the results of the experiment, and its notes contain a linked notebook with the code for the simulations. The y-axis is the absolute difference between the fraction of AI agents and humans choosing "Left." For each model, the white bar shows the responses of the persona-less AI agents, and the grey bar shows the responses of the calibrated sample. The differences are substantial: across all models, the responses of the weighted AI agent samples are much closer to those of the human subjects. The MSE of the calibrated samples (0.094) is about half of the persona-less control MSE (0.182). This improvement suggests that theory-grounded persona mixtures can indeed capture meaningful patterns in human decision-making that transfer across game formats.

Figure 4: Out-of-sample evaluations of calibrated samples of AI agents.



Notes: This figure compares the performance of persona-less and calibrated samples of AI agents using sequential games from Charness and Rabin. Rows correspond to different game scenarios, and columns to different roles (Person A or Person B). The x-axis shows different LLMs, and the y-axis shows the absolute difference between the fraction of AI agents and the fraction of human subjects choosing "Left" in the original experiment. White bars correspond to the persona-less AI agents, and the grey bars correspond to the calibrated AI agents. Error bars show 95% Wilson confidence intervals for the AI proportion, symmetrically shifted to the absolute difference from the human proportion. See `www.expectedparrot.com/content/9ff43614-2726-440b-a745-f325555eab4b` for simulation code.

These results show that we are not limited to whether off-the-shelf AI agents respond like humans. Instead, we can construct AI agents that respond more like humans in particular settings using personas

based on some relevant theory. We can then use these calibrated agents to explore new scenarios plausibly governed by the same theory with more confidence that their responses will predict the relevant human responses. Indeed, Manning and Horton show that by constructing and then validating agents in this way (on some of these very games), we can substantially improve their ability to predict human responses in novel settings where no prior human data exists.

2.3 *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 select an option if it is presented as the status quo. In one scenario, they asked subjects to allocate a safety budget between automobiles and highways. The instructions were:

> "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 recapitulate Samuelson and Zeckhauser's experiment. First, we endow each AI agent with one of twelve beliefs about the importance of car and highway safety. For each belief, we create a "car owner" and a "non-car owner" AI agent. We then have each of the 24 agents choose one of four funding allocations in four scenarios: each one phrased to indicate a different option as the *status quo*. We run this experiment across all five LLMs with temperature set to 1, generating multiple choice occasions per agent-scenario combination. We use GPT-4, GPT-4O, CLAUDE-SONNET-3.5, CLAUDE-SONNET-3.7, and DEEPSEEK. Using this combination of models allows us to compare the AI agents' responses as the capabilities of LLMs progress, as well as between different LLM providers. We provide further simulation details inside the notebook linked in the Table 1 notes.

It is worth noting that this within-subject experimental design is possible because AI agents have no memory unless we explicitly provide them with information about their previous decisions. This would be

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

impossible with human subjects. They would likely become aware of the experimental manipulation after exposure to each status quo framing.

We model the AI agents' responses as discrete choices over the four allocations using a conditional logit specification and report the results in Table 1. We pool data across all five LLMs but allow coefficients to vary by model using model-specific interactions. Each observation is one choice occasion (an agent choosing among four alternatives in a given scenario), and the model estimates the probability of choosing each alternative conditional on the choice set.

Table 1: Recapitulation of Samuelson and Zeckhauser (1988) with AI agents

	Dependent variable:		
	Choice of Allocation		
	(1)	(2)	(3)
Claude-3.5: AutoShare	−0.010* (0.002)	−0.011* (0.002)	−0.015* (0.002)
Claude-3.7: AutoShare	−0.004* (0.002)	−0.005* (0.002)	−0.007* (0.002)
Deepseek: AutoShare	0.017* (0.002)	0.018* (0.002)	0.017* (0.002)
GPT-4: AutoShare	−0.008* (0.002)	−0.009* (0.002)	−0.014* (0.002)
GPT-4o: AutoShare	−0.004* (0.002)	−0.004* (0.002)	−0.008* (0.002)
Claude-3.5: I(StatusQuo)		0.911* (0.054)	
Claude-3.7: I(StatusQuo)		0.703* (0.055)	
Deepseek: I(StatusQuo)		0.762* (0.055)	
GPT-4: I(StatusQuo)		0.931* (0.054)	
GPT-4o: I(StatusQuo)		0.316* (0.058)	
Claude-3.5: |DistStatusQuo|			−0.032* (0.002)
Claude-3.7: |DistStatusQuo|			−0.020* (0.002)
Deepseek: |DistStatusQuo|			−0.029* (0.002)
GPT-4: |DistStatusQuo|			−0.040* (0.002)
GPT-4o: |DistStatusQuo|			−0.023* (0.002)
Observations	6,999	6,999	6,999
Log Likelihood	-9,630.112	-9,173.293	-9,188.728

Notes: This table reports the results from our replication of Samuelson and Zeckhauser (1988) with GPT-4, GPT-4O, CLAUDE-SONNET-3.5, CLAUDE-SONNET-3.7, and DEEPSEEK. Columns show logit discrete choice models estimating the probability of choosing each of the four allocations. Observations represent one funding allocation choice—excluding one observation with missing data. Each choice occasion is expanded to four rows (one per alternative), with the dependent variable indicating the selected option. We pool data across all five LLMs but estimate model-specific coefficients through interactions with model indicators. Independent variables are the proportion of funds allocated to automobile safety for each option, an indicator of whether or not the option was framed as the status quo, and the absolute distance between each option and the status quo. See www.expectedparrot.com/content/0bacba4d-91fa-4005-8f7d-abd34a4a1f52 for prompts and code. Significance Indicator: *p<0.05.

In Column (1), we see that AI agents, with the exception of DEEPSEEK, 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 agents exhibit considerable status quo bias, although GPT-4O does so 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, on average, less likely to be chosen for all models. Both specifications (Columns 2 and 3) show strong evidence of status quo bias, with Column (2) fitting the data slightly better. Overall, our results suggest that AI agents shift their preferences toward the status quo allocation—similar to human subjects. Even with substantial post-training fine-tuning, possibly suboptimal heuristic features of human choice remain relevant to LLM behavior.

2.4 A complexity view of prospect theory (Oprea, 2024b)

Prospect theory predicts that when making risky choices, people overweight small probabilities, underweight large probabilities, are risk-averse over gains, and risk-seeking over losses (Kahneman and Tversky, 1979). This gives rise to the classic "fourfold pattern" of risk attitudes. Oprea (2024b) examined if this fourfold pattern can emerge in the absence of uncertainty. In a novel design, participants were given endowments and faced sets of 100 closed boxes, each containing either a gain or a loss with probability $p \in \{0.1, 0.25, 0.75, 0.9\}$. In the "gain" framing, some boxes contained \$25 and the rest \$0; in the "loss" framing, some contained −\$25 and the rest \$0.

In the "lottery" condition, for each probability p and each framing (gain or loss), certainty equivalents were elicited by asking participants to choose between a sure amount and the outcome from one randomly selected box. This lottery condition is equivalent to a standard binary-choice lottery. As expected, participants' certainty equivalents collectively displayed the fourfold pattern.

A second, "mirror" condition used the same box setup, probabilities, framings, and dollar amounts, but altered the payoff rule. Instead of randomly choosing one box, all boxes were opened and participants received the average payoff across them. Unlike the lottery condition, this decision involves no risk: the payoff is deterministic and always equals the expected value ($\pm 25 \times p$). The rational certainty equivalent is therefore the expected value. The subjects' responses to risk should not matter, and the fourfold pattern should not appear.

Oprea finds that participant responses in the mirror condition still displayed the pattern in the aggregate. This suggests the pattern does not arise solely from probability weighting or reference dependence, and that prospect theory is not just a theory of uncertainty. Instead, Oprea argues it reflects systematic responses to complexity under cognitive constraints. These findings have sparked some debate. Banki et al. (2025) contend that the fourfold pattern appears in the mirror condition because subjects are confused about the

instructions, and interpret the condition as a risky lottery—likely because of prior exposure to conventional certainty equivalent tasks.

AI agents may offer a unique perspective to help us understand features of this debate. In simulations, we can systematically vary features of their personas that might indicate which perspective is correct. This would otherwise be difficult, if not impossible, to do with human subjects.

We extend Oprea's central experiment using AI agents. We endow each agent with \$50, and then ask it to provide certainty equivalents for 100 box sets with probabilities $p \in \{0.05, 0.10, \ldots, 0.95\}$, adding 14 probability conditions beyond those in the original experiment for each condition-framing combination. We construct three personas with the goal of manipulating the LLMs as if they had varying levels of available cognitive resources—a well-established method for identifying complexity effects (Oprea, 2024a). The personas are: (i) "you are very bad at math," (ii) "you are average at math," and (iii) "you are a mathematical genius." If the fourfold pattern reflects behavior under complexity rather than under risk, not only should the lottery and mirror conditions be similar, but we should also expect larger deviations from expected utility—toward prospect-theoretic predictions—among the lower-math ability personas. We run simulations for each agent, including a baseline agent without any persona, 10 times on five LLMs (GPT-3.5-TURBO, GPT-4O, CLAUDE-HAIKU-3, CLAUDE-SONNET-3.5, and DEEPSEEK), all at temperature 1. This generates 15,200 observations.[9]

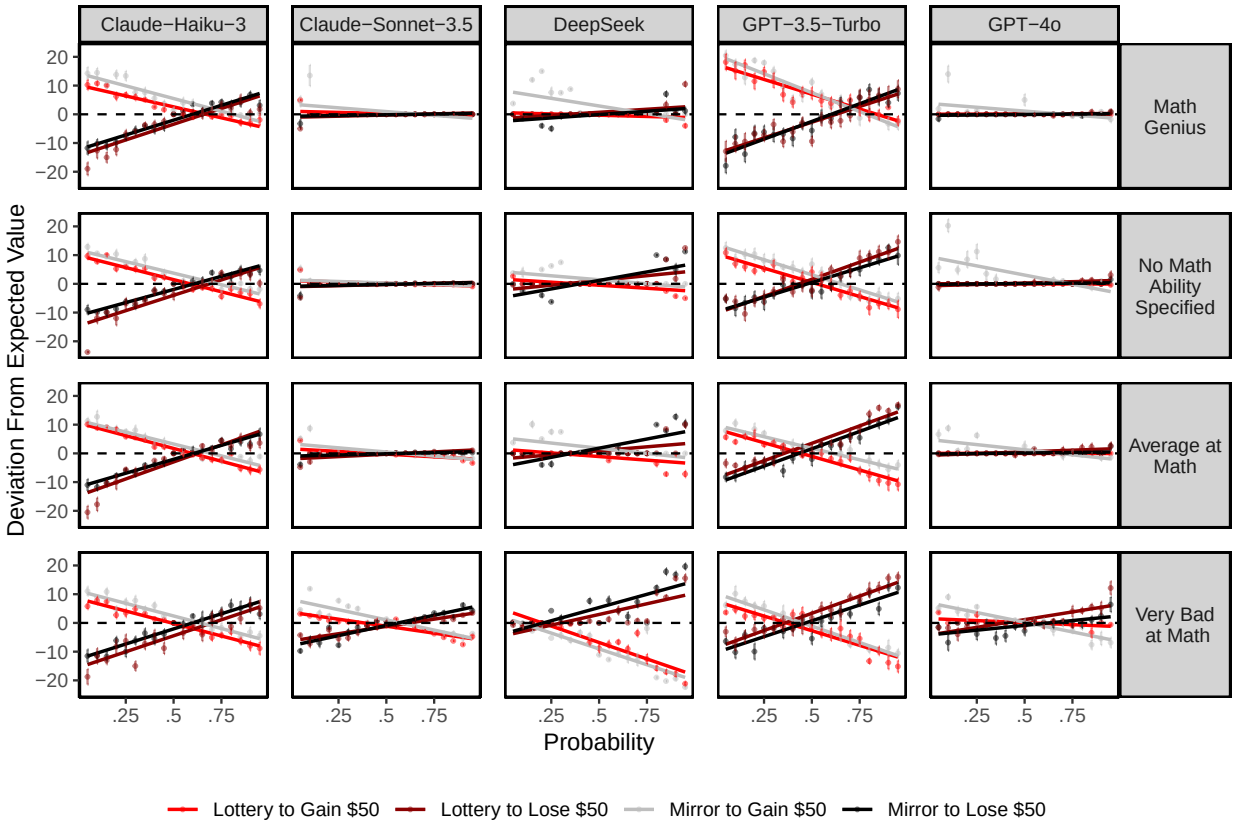
Figure 5 displays the simulation results. The x-axis shows the probability of winning (or losing) \$50, and the y-axis shows the deviation of each agent's certainty equivalent from the lottery's (or mirror's) expected value. Columns correspond to different models, and rows indicate the AI agent's assigned mathematical ability. Colors represent condition-framing combinations. Each point is the mean certainty equivalent across 10 runs of a persona at a given probability, with 95% confidence intervals. Lines show OLS fits regressing deviation from expected value on probability for each condition independently for losses and gains.

Across models, the lottery and mirror conditions generally move in lockstep: either both show the characteristic X-shape or both are flat at the expected value. There are no instances where the lottery displays the fourfold pattern but the mirror does not. Notably, there are also no panels where the slopes are reversed: under gains, all regression lines are weakly decreasing in p, and under losses, all regression lines are weakly increasing in p. Less advanced models—GPT-3.5-TURBO and CLAUDE-HAIKU-3—show pronounced deviations between small and large probabilities. By contrast, more capable models—GPT-4O,

[9] 4 agents × 2 conditions × 2 framings × 19 probabilities × 5 models × 10 repetitions = 15,200.

CLAUDE-SONNET-3.5, and DEEPSEEK—more often select expected values as a baseline, particularly when endowed with stronger mathematical abilities, but most clearly exhibit the fourfold pattern when prompted as "very bad at math."

Figure 5: Recapitulation and extension of Oprea (2024b)



Notes: This figure shows the results of our recapitulation and extension of Oprea (2024b) with AI agents. The x-axis is the probability of winning or losing $50, and the y-axis is the deviation from the lottery's (or mirror's) expected value. Columns correspond to different LLMs, and rows to whether different agents were endowed with different mathematical abilities. Each dot is the mean certainty equivalent across AI agents (with 95% confidence intervals), and the solid lines are OLS fits within each framing-condition combination. Colors represent condition-framing combinations. See `www.expectedparrot.com/content/05009e7a-84e8-4c14-bc5d-de0408df89a2` for simulation code.

On its face, our recapitulation supports Oprea's claim that the fourfold pattern is not a special phenomenon of risk. However, it is worth considering systematically the plausible explanations for these AI responses—both for exposition of *Homo silicus* and as commentary on the results.

Could our findings reflect mere noise or rote memorization by the LLMs? Pure noise seems implausible, given the consistency of the fourfold pattern across five separate models and thousands of simulations. We would expect to see at least one regression line with an inverted slope relative to the prospect-theoretic predictions if this were the case. Memorization of Oprea is also unlikely. Early drafts may be in training

data, but the paper only gained broad visibility in 2024, and its deterministic mirror condition is novel. We also vary probabilities and payoffs beyond the original, so a narrow script tailored to those exact experiments would not suffice. It would represent a remarkable degree of transfer learning.

Another possible explanation is that LLMs may have ingested decades of lottery literature, so they default to a standard prospect-theoretic response. This is similar to Banki et al.'s argument that Oprea's subjects may have reverted to familiar responses because they misunderstood the payoff structure. The variation in responses across personas suggests that the "confusion with lotteries" perspective is unlikely. The more capable models largely select the expected value in all conditions when assigned a high, medium, or unspecified mathematical ability, but generate responses corresponding to an increasingly pronounced fourfold pattern when they are "very bad at math." All three of the more capable LLMs—each independently developed with its own training data and fine-tuning processes—appear to have learned some relationship between stated mathematical ability and deviation away from expected utility theory in a way that aligns precisely with prospect theory. After an extensive search of the literature, we found no such robustly reported relationship for an LLM to have learned to apply by default.[10]

The most plausible explanation in our view is that LLMs, having ingested innumerable examples of humans solving problems and demonstrating their preferences over possible states of the world—one of the implicit sources of latent social information that we discussed in the introduction—learned systematic relationships about how humans trade off features of decision-making problems under complexity. This view is consistent with both the similarity between lottery and mirror conditions, and the increasingly prospect-theoretic responses as the agents get worse at math. Of course, this does not provide definitive evidence on whether Oprea or Banki et al. are correct. For now, real human data are necessary to shed light on such a question. But these findings highlight the power of using LLMs to study subtle behavioral patterns likely present in their training corpora.

2.5 Labor-labor substitution under a minimum wage (Horton, 2025)

Horton (2025) reports the results of an 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

[10] While there is some evidence of a weak negative correlation between cognitive ability and risk aversion in gains, it is not large, and we did not find published evidence for a general relationship relating to relative losses (Lilleholt, 2019; Dohmen et al., 2010).

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 agents. We create scenarios where we ask agents (employers) to select from pairs of applicants that vary in their 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 is:

> You are hiring for the role of "Dishwasher." You have 2 candidates.
>
> Person 1: Has 1 year(s) of experience in this role. Requests $[`wage_ask_1`]/hour.
>
> Person 2: Has 0 year(s) of experience in this role. 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 is the "inexperienced" worker. The inexperienced worker always asks for $13/hour, and the experienced worker asks for a wage between $12 and $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 repeat our experiment twice: once when we tell the employer that the typical wage ask for the position is $12/hour, and once when we provide no information about the typical wage. We elicit the AI agents' hiring decisions in 24 scenarios with 43 different models five times each at a temperature of 0.5.[11] We ask the employer to always hire a candidate. Our empirical approach is to regress the hired worker's wage and experience on an indicator for the minimum wage treatment status. Table 2 reports the regression estimates; the notes include a linked notebook with the list of models and all simulation details. All specifications include model fixed effects, and standard errors are clustered at the model level. Column (1) shows that without a reference wage, the minimum wage caused an increase of about $1.3/hour in the wage of the hired worker. Column (2) shows that when a reference wage is provided, this effect grows by nearly 50%. Positive coefficients are expected because hiring was mandatory for the AI employer.

When we regress the hired worker's experience on the presence of a minimum wage, we see similarly substantial effects. Column (3) shows that without a reference wage, the minimum wage increased the hired

[11] There are 6 possible wage asks, 2 minimum wage conditions, and 2 reference wage conditions. $6 \times 2 \times 2 = 24$.

Table 2: Effects of minimum wage on observed wage and hired worker attributes

	Dependent variable:			
	Hire wage		Hire experience	
	(1)	(2)	(3)	(4)
I(Minimum Wage)	1.314*	1.984*	0.081*	0.265*
	(0.067)	(0.053)	(0.018)	(0.023)
Reference Wage	No	Yes	No	Yes
Model FE	Yes	Yes	Yes	Yes
Observations	2,580	2,579	2,580	2,579
R^2	0.273	0.628	0.226	0.179

Notes: This table shows regressions where the independent variable indicates whether a minimum wage was imposed on the AI employer. Dependent variables are (1) and (2) the hired worker's hourly wage and (3) and (4) their years of experience. Columns (1) and (3) show results without a reference wage, and Columns (2) and (4) with a reference wage. Employer-level observations are sampled from 43 different models. See www.expectedparrot.com/content/15eb59ad-f7c8-4c9b-87d1-a58cfecc13ec for prompt details and the list of models. Significance Indicator: *p<0.05.

worker's years of experience by about a month. With a reference wage, Column (4) shows that the introduction of the minimum wage caused the employer to hire workers with roughly three more months of experience on average. We return to why the reference wage matters in Section 4, but even across simulations of dozens of independently developed models, these increases in wage and experience are directionally consistent with Horton's findings.

3 Why and when we can learn from *Homo silicus*

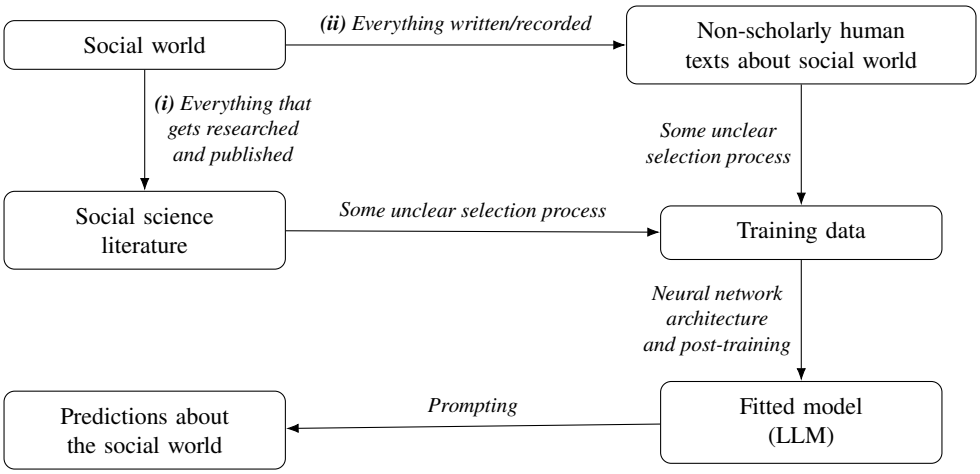
In the previous section, we demonstrated that AI agents can simulate features of human behavior in experiments along with various methods to facilitate such simulations. Next, we examine why LLMs possess these capabilities in the first place, which can inform when the approach is most likely to be useful. We begin by describing the data that enables highly general AI simulations, and then discuss how economists can conceptualize them in their methodological toolkit.

3.1 Why LLMs can simulate human behavior

Good predictions start with good training data. LLMs can flexibly predict human responses because their training data contain rich information about the social world, and their architectures and post-training processes allow them to learn and apply generalizable human behavioral patterns from that data.

Figure 6 depicts the data pipeline from the social world to LLM predictions about the social world. LLMs gather informative signals about the social world through two distinct paths: (i) explicitly, by memorizing findings and theories from the social science literature and (ii) implicitly, by ingesting text that captures aspects of the social world, such as news articles, social media content, books, and so on.

Figure 6: Why LLMs are good predictive models of the social world



Notes: This figure is a simplified depiction of the data pipeline from the social world to the LLM's training data, and then to the LLM's predictions about the social world. It highlights that LLM simulations can match human behavior in the real world through two distinct paths: (i) by acting in accordance with explicit theories and findings in the social science literature, or (ii) implicitly by ingesting elements of human behavior "in action" captured in non-scholarly human-generated text.

The social science literature path contains records of generalizable patterns, processes, relationships, and theories about human behavior—precisely the kinds of insights we would like LLMs to internalize and apply more broadly. Consider that at the core of economic analysis is the idea that humans are goal-oriented agents maximizing some objective under constraints (von Neumann and Morgenstern, 1944), or the empirical evidence that humans often derive utility with diminishing marginal returns to consumption (Samuelson, 1937) and exhibit strong social preferences by valuing fairness, reciprocity, and others' welfare alongside their own (Fehr and Schmidt, 1999). These concepts do not just magically exist in the ether. They have been codified through decades of research, dramatically improving economists' understanding of human behavior. These are the essential ideas that any "general" human model should take into account. And while learning from the social science literature comes with distinct challenges, which we discuss in more detail in Section 4, it is a necessary ingredient for LLMs to effectively simulate human responses in many settings.

In contrast to the social science literature path, the non-scholarly, human-generated text path provides LLMs with true aspects of the social world "in action:" how people communicate, make decisions, interact

with each other, and perceive the world. We might be skeptical that a language model could, say, represent physics we have yet to theorize, but it surely has internalized relevant knowledge about the social world that has never been written down in academic work. This is not to say that the data are the social world or a perfect representation of it. Rather, it contains countless examples that reiterate basic economic relationships across a wide range of contexts.

Consider the observation that people prefer lower prices and greater quantities, all else equal. Any economic textbook would contain this observation either in some written form or with the condition $\frac{\partial}{\partial p}\left(u(x^*) - px^*\right) < 0$. But the same observation also appears in widely-read, pre-social science sources. In the Book of Amos, the prophet denounces unfair traders: "*And the Sabbath, that we may offer wheat for sale, that we may make the ephah small* [a weight to put on a balance] *and the shekel great and deal deceitfully with false balances.*"

What about more subtle knowledge, such as the literature on cheap talk in bargaining games? There, we would not expect the LLM to learn every detail and figure out all implications of Crawford and Sobel (1982), but its data contain sources that capture some of the basic ideas and tensions. For example, "*'It's no good, it's no good!' says the buyer—then goes off and boasts about the purchase*" (Proverbs 20:14) captures the idea that in bargaining, the buyer and seller are not always forthright.

Modern non-scholarly sources echo these dynamics as well. Over the past three decades, an increasing share of economic activity has occurred online, creating an extensive record of transactions, reviews, negotiations, and market interactions. These digital traces provide ample raw material from which LLMs learn realistic representations of human decision-making in many settings.

Importantly, both the implicit and explicit paths are shaped by distinct and often opaque selection mechanisms. Peer review determines which studies enter the formal literature, and individual self-presentation—what people choose to write, post, or record—filters the broader corpus of human text. Model developers then apply a further layer of curation when assembling the final training data, excluding or weighting sources according to proprietary criteria.

Neither stream of information is complete or representative, which is why AI simulations should not be treated as unequivocal evidence about human behavior. Over time, however, we expect that the models most frequently used for simulation will converge toward greater transparency and openness, with clear data inclusion standards, explanations of post-training processes, and documentation. While open models are not yet as advanced as the most capable models and not viable for many AI experiments, they are improving

rapidly. When possible, we fully endorse the use of open models. For now, there remain trade-offs between model capability and openness, though these have narrowed considerably even since the first release of ChatGPT.

While training data are key to the ability of LLMs to simulate human behavior, these information sources alone do not create a coherent model. LLMs' architectures and post-training processes translate rich training data into useful predictions about human behavior. Modern transformer LLMs are flexible sequence-to-sequence estimators: they can approximate complex mappings from contextual variables such as prices, payoffs, and information sets to a distribution of actions in their textual responses (Yun et al., 2019). Their training objective rewards outputting empirically truthful conditional probabilities over their data (Gneiting and Raftery, 2007)—minimizing log loss is a strictly proper scoring rule.

Post-training methods (like instruction tuning or reinforcement learning with human feedback) steer models away from brittle responses driven by spurious patterns in their training data and toward cooperative, instruction-following behavior—we exploit this property when we endow agents with roles and preferences. The most advanced frontier models even have specific reasoning capabilities which allow them to better "think through" how a human would respond based on a given set of characteristics. Together, massive training corpora combined with these architectural and training innovations enable LLMs to capture generalizable patterns of human behavior that can be used to generate predictions in a wide range of settings.

3.2 *Homo silicus* as theory

In Section 2, we used elements of the research processes traditionally associated with empirical work to construct and analyze our AI simulations. We designed experiments, collected data, and used statistical analyses to summarize the results. Despite these ostensible similarities to applied work, many of the most valuable use cases and insights gained are more akin to the practice of economic theory.

To see why, consider an economist who is trying to understand the following puzzle: why do large firms pay higher wages than observationally equivalent small firms? The researcher begins by trying to create a model where such a gap exists in equilibrium. If she assumes perfectly competitive labor markets and costless mobility, her model would likely predict no firm-size wage differences—workers are simply paid their marginal product. Of course, this disappointing first step does not prove that theorizing is worthless. Instead, it highlights that the baseline model omits important features of real labor markets.

The next step in the researcher's journey is to create richer models. With each new theory direction, she

would attempt to formalize a different real-world mechanism. At each step, the researcher must draw on domain knowledge and the literature to specify a new model, work out the equilibrium (often painfully), and derive useful, testable predictions.

With *Homo silicus*, we can easily construct a large number of AI simulations—based on a world model informed by empirical data—that might bear on the firm-size premium. For example, suppose we simulate a job seeker weighing offers from a small startup and a large incumbent. If an AI agent spontaneously emphasizes differences in layoff risk and promotion timelines—and accepts a lower starting wage at the larger firm because it anticipates faster wage growth and insurance value—this suggests a mechanism and maybe even an empirical design: measure expected wage paths and mobility frictions by firm size.

Similar to modeling, simulations can lead us in the wrong direction. But they likely make expensive data collection more useful by helping us target efforts where the marginal value of information is highest. Economists have long been aware of how seemingly simple micro-decision making can lead to surprising and subtle equilibrium implications (Akerlof and Yellen, 1985). It is not hard to imagine how the ability to easily conduct rich, complex simulations might lead to surprising and generative insights. Indeed, this is one of the major differences between the physical and social sciences: the ability to iterate and run follow-up experiments. Given the differences in predictive power between physical and social scientific theories, there is reason to believe a low-cost and high-speed approach to "learning by doing" (Arrow, 1962) at speeds closer to the lab bench could be a powerful tool for progress.

While the many parallels with economic theory are clear, a key difference—and an advantage—of *Homo silicus* is its flexibility. We can use AI simulations to generate predictions in virtually any setting described in natural language. This is often not possible with conventional economic theory because it generally requires fixed input parameters and structural assumptions.

To make concrete what we mean by flexibility, and why economic models are relatively inflexible, consider the model in Charness and Rabin used to estimate the impact of social preferences on participant allocation decisions. For the unilateral dictator games we studied in Section 2, they model the utility player B derives from a given allocation $\boldsymbol{\pi} = (\pi_A, \pi_B)$ as:

$$U_B(\pi_A, \pi_B) = (\rho \cdot r + \sigma \cdot s) \cdot \pi_A + (1 - \rho \cdot r - \sigma \cdot s) \cdot \pi_B, \tag{1}$$

where $r = 1$ if $\pi_B \geq \pi_A$ and $r = 0$ otherwise; $s = 1$ if $\pi_B \leq \pi_A$ and $s = 0$ otherwise. Here π_A and π_B

are the payouts for players A and B in allocation $\boldsymbol{\pi}$, respectively. ρ and σ are free parameters representing distributional preferences which Charness and Rabin estimate from their human data.

With values for the free parameters, Equation 1 could be used to predict dictator preferences over a range of two-player allocations. However, this model is fundamentally limited in two ways. First, it cannot incorporate any additional information from a given setting that may be relevant to the dictator's decision. Suppose that after running their experiment, Charness and Rabin learn that about half of the participants are friends and believe this might strongly interact with baseline social preferences. To use this information to improve the model's predictions, say, as an extra variable, one must make additional assumptions and re-estimate all parameter values. Second, this model cannot make predictions in structurally distinct games—even if these games share similar underlying motivations. Equation 1 is useless for predicting the dictator's choice over allocations for any number of players greater than two.

Traditional machine learning models, which have become increasingly relevant in predicting human behavior in economics (Fudenberg and Liang, 2019; Peterson et al., 2021; Hirasawa et al., 2022; Zhu et al., 2025b; Andrews et al., 2025), have similar limitations. It is not possible to simply add new features to these models to make predictions in structurally distinct settings. For example, if a regression model is trained to predict earnings linearly from education and age, but a new observation comes along that also includes cognitive ability, the linear model's estimated coefficients offer no information about the relationship between cognitive ability and earnings. This extra column of data must either be ignored by the model or the model must be retrained from scratch using data from all three covariates. In short, most economic and machine learning models cannot function outside the domain for which they were built.

AI simulations, however, are not subject to the same structural limitations. In the parlance of the machine learning literature, we may think of LLMs as those models with a strong "inductive bias" for social science prediction problems. We can simply adjust the setting provided in a prompt (literally rewriting the natural language description of the setting), and the LLM will predict a response based on the relationships it has learned from its training data. While we cannot directly apply Equation 1 to three-player dictator games, we could supply the model with the parameters estimated in two-player games to an LLM and prompt it to extend the mapping in a reasonable, explicitly described way to games with any number of players.

This is possible, in part, because LLMs can reasonably impute missing information necessary to make a coherent prediction. Since they are next-token predictors, they can accommodate the functional form of any model written in a prompt. Such flexibility is well-embodied in the "P" in ChatGPT (Generative Pre-trained

Transformer)—the idea that by pretraining on a massive corpus of text, the model can generalize well.[12]

Flexibility is not without its drawbacks. LLMs are black boxes that can fail to generalize in unexpected ways (Vafa et al., 2024a). But viewing AI agents as vessels to apply interpretable economic theory to new settings is a powerful way to mitigate this risk. Because instruction following is the central post-training objective of modern LLMs (Ouyang et al., 2022; Bai et al., 2022), a more reliable way to steer them is to express our assumptions as explicit instructions. Economic theory maps naturally into such instructions: it declares state variables and constraints, specifies an objective, and implies decision rules. When we write those elements clearly in the prompt (e.g., "You only care about efficiency"), the model can apply the rule across new settings without retraining—and we can audit whether it did so. As we discuss in Section 4, explicit, unambiguous instructions are valuable because they require far less from the LLM's underlying world model. They also make behavior checkable in simple test cases and reveal where extrapolation fails before scaling simulations up in complexity. Put differently, economic theory supplies the instructions; LLMs supply the capacity to execute them flexibly.

We applied a version of this theory-as-flexible-instruction approach when calibrating our sample for the two-stage allocation games in Section 2. For the unilateral dictator games, we endowed "efficient" agents to maximize total payoff, "inequity-averse" agents to minimize payoff differences, and "self-interested" agents to maximize their own payoff. In each of the six allocation decisions, there is a clear mathematical mapping from the instruction to the preferred allocation; only Berk26 is equivocal for the efficiency persona because both allocations are equally efficient. After calibrating on the unilateral tasks, we applied the same agents to the two-stage games, where "efficient," "inequity-averse," and "self-interested" become more ambiguous. For example, in Berk27, should the self-interested Player A choose "Left" (certain 500) or "Right" (either 800 or 0 depending on B's move)? Although 800 seems more likely than the spiteful 0, there is no guarantee B cooperates. Still, the calibrated agents predicted human responses far better than persona-less LLMs, suggesting that even this imperfect theoretical mapping was relevant.

A natural question raised by these results is what they teach us about humans rather than models. Are we actually learning that human dictators behave like mixtures of efficient, inequity-averse, and self-interested types? More broadly, whenever an interpretable prompt improves a simulation's fit to human data, should we take that as evidence that the theoretical idea associated with that prompt is literally true of people and should

[12] To be sure, there are settings where parametric models are better—e.g., response data from 10,000 binary-choice lotteries (Peterson et al., 2021) when we are specifically studying interpolation in risky choice (Enke and Shubatt, 2023; Zhu et al., 2025a).

anchor our economic models? This echoes the broader idea of reverse-engineering predictive power from machine learning models as a tool for hypothesis generation (Ludwig and Mullainathan, 2024). For now, that inference is too strong. LLMs remain black boxes: we do not yet know why a given prompt improves performance or how the internal representations that drive these gains relate to human behavior. The nascent field of mechanistic interpretability is beginning to shed light on these questions (Olah et al., 2020; Elhage et al., 2021), but it does not yet let us map prompts cleanly onto interpretable features that explain human responses.

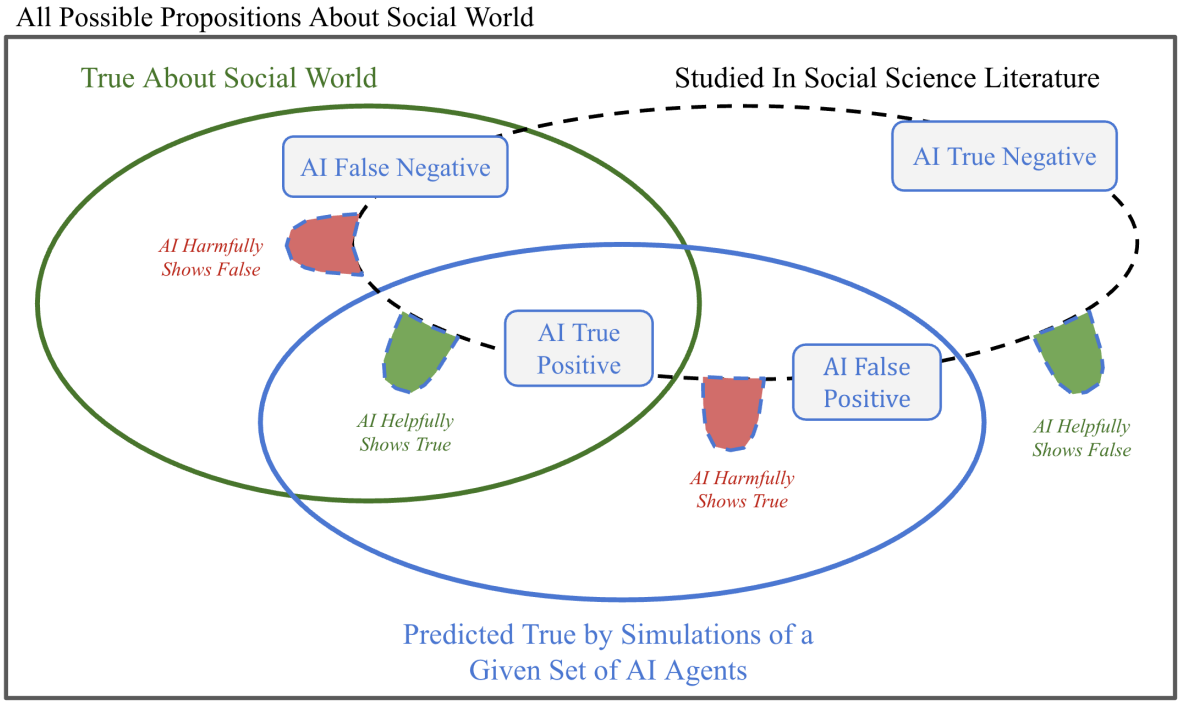
Even so, these patterns of improved predictive accuracy provide compelling—and increasingly abundant—evidence that LLMs have internalized meaningful relationships between interpretable theory and human behavior (Zhu et al., 2025a).[13] That conclusion is reinforced by the opposite pattern: prompts that endow AI agents with atheoretical or scientifically meaningless traits generally do not improve predictive accuracy. Manning and Horton (2025), for example, show that giving agents hobbies or favorite TV shows does little to improve fit on Charness and Rabin's games, either in-sample or in new settings, while theory-based social-preference personas do. Of course, there are surely settings where the correct theory does not improve fidelity, or meaningless prompts do. But the fact that we cannot yet make precise inferences about human behavior from which prompts improve predictive fidelity does not mean these patterns lack value. They may still provide structured signals about which theoretical ideas are promising for modeling and exploration, even if the mechanisms underlying those gains are not yet fully understood.

3.3 Using *Homo silicus* to navigate uncharted territory

Figure 7 shows how we can think about the relationship between the social world, the social science literature, and AI simulations. The green oval (upper left) depicts the set of true propositions about the social world. The black oval (upper right) depicts the set of propositions that have been studied and recorded in the social science literature—regardless of the findings. These may or may not have been labeled true, which is why the oval is dashed, but they have been evaluated by researchers. The blue oval (bottom center) depicts the propositions that simulations of a given set of AI agents powered by a particular LLM would predict to be true. The blue "feet" protruding from the black oval represent how new AI results from the given agent set would expand on the existing social science literature.

[13] Xie et al. (2025) offer a complementary illustration: when they optimize prompts to match human behavior across seven classic economic games, the high-fidelity prompts that emerge invoke precisely the constructs economists would anticipate—risk preferences in investment games, cooperation in public-goods games, fairness in allocation games, and so on.

Figure 7: How AI simulations may help or hurt social science research on new propositions



Notes: This figure displays the predictions made by simulations of AI agents (bottom center blue oval) relative to the true social world (upper left green oval). The black dashed oval in the upper right offers the propositions that have been evaluated by social scientists—it is agnostic as to whether such propositions are recorded as true or false. The blue "feet" show how AI results could expand the existing literature. The Venn diagram implicitly offers a confusion matrix for where AI simulations can either help or hinder researchers as they explore novel ideas.

Along with Figure 6, this diagram shows how AI simulations show a selective and noisy slice of the social world. They are not yet an adjudication of truth but a navigation aid: a disciplined way to mark the types of propositions that appear promising. Read as a confusion matrix, Figure 7 implies four possible cases: (i) *AI True Positives*: true propositions supported by AI simulations, (ii) *AI True Negatives*: false propositions rejected by AI simulations, (iii) *AI False Positives*: false propositions supported by AI simulations, and (iv) *AI False Negatives*: true propositions rejected by AI simulations. Researchers stand to gain when AI simulations fall in the true positive and true negative regions. On the other hand, AI simulations may amplify errors when they generate false positives or neglect well-established results. This underscores that simulations can be either helpful when they corroborate good social science or refuse to imitate bad social science, or harmful when they introduce spurious results or fail to predict legitimate findings.

Notably, humans are not necessarily strong benchmarks for predicting the results of social science research. A growing literature shows that even trained social scientists are surprisingly poor at forecasting the outcomes of interventions and experiments (DellaVigna et al., 2019; Milkman et al., 2021; Gandhi et al.,

2025, 2024; Duckworth et al., 2025). This implies that the "bar" for a useful predictive tool may be lower than we often assume. AI simulations need not be perfect to be informative—they may only need to improve on human judgment.

Still, how would we know in which of the four regions we are likely to find ourselves when using AI simulations? Without a correct causal model, no statistical procedure can, ex ante, guarantee performance in novel settings (Pearl, 2009). To get unbiased estimates of economic parameters or to make predictions with guaranteed statistical bounds, one needs at least some representative data or sufficient statistics. Only then, with the appropriate statistical frameworks, can researchers use AI agents' responses as data to draw unequivocal inferences about human behavior (Ludwig et al., 2025). In other words, researchers generally need some sort of ground truth or gold standard dataset to compare against. In general, ground truth is the analogous human subjects' data for a given set of simulations.

One such case where ground truth is available is when researchers have already committed to collecting a human sample and want to stretch it further. For these settings, prediction-powered inference is especially useful. The method starts with a ground truth sample, trains a model to predict the outcome from observed covariates, and then combines those predictions with inputs lacking outcome labels to form estimators that remain unbiased for the target parameter (Angelopoulos et al., 2023; Egami et al., 2023). The more accurate the predictive model, the more precise the resulting estimates and the lower the cost of achieving a desired level of precision. Intuitively, the machine learning model's predictions act like additional observations that are corrected using the ground truth sample. AI simulations can be particularly effective in this role (Broska et al., 2025; Ludwig et al., 2025), in no small part because of their flexibility.

Ground truth, however, is often unavailable for another one of the most useful applications of AI simulations: exploring new settings where no prior data, and therefore no guarantees, exist. In such cases, we are not generally interested in unbiased inference, but instead in exploration and hypothesis generation—hence our continued emphasis on theory comparison. The evidentiary burden for such research tasks is far lower than for unbiased inference. Still, researchers should not want to explore blindly.

The key is to construct a set of AI agents that are more likely to perform well in a given setting, even if perfect fidelity cannot be guaranteed. One way to do so, as we have shown, is to design theory-grounded agents and test them in settings where human ground truth is available and theoretically related to the target domain of interest. If these agents can accurately predict human responses in such benchmark settings, we gain confidence that they will generalize effectively to nearby, unlabeled settings. Conceptually, there are

many distinct blue ovals in Figure 7, each representing predictions of simulations from a different set of AI agents. These could be different models, prompts, or a combination of both. The goal is to identify those that can be shifted closest to the green oval—the set of true propositions—especially for those propositions where the black oval provides reliable labels. When the new settings lie close to these correctly labeled regions, we might expect similar predictive success—illustrated by the green "AI helpful" feet in the figure.

Our experiments with Charness and Rabin's unilateral and two-stage allocation games provide an example of this approach. We used agents grounded in social preferences to match human responses in the unilateral dictator games, then applied them to the related two-stage games where we presumed the same underlying theory (various social preferences) drove behavior. Manning and Horton (2025) provide a full elucidation of this approach and demonstrate that the method can substantially improve the predictive power of AI simulations in entirely novel settings where no prior human data exists.[14]

Designing theory-grounded agents helps us improve fidelity, but their effectiveness ultimately depends on what the underlying LLM already knows about the world. In practice, this means that some domains will be more conducive to reliable simulation than others. Certain types of settings may therefore inspire greater confidence that AI simulations will be effective—specifically, those we would expect to be well-represented in the LLM's training data. This follows a simple machine learning principle: models are most accurate in domains richly represented in their training data. Indeed, evidence shows that LLMs perform far better on problems and questions drawn from well-covered settings (McCoy et al., 2024).

Perhaps unsurprisingly, two of the largest experimental papers in terms of participant sample size and number of distinct experiments demonstrating that LLMs have broadly applicable predictive power in novel settings focus on domains that are likely to be richly represented in digital text: laboratory experiments in psychology (Binz et al., 2025) and digital marketing experiments (Hewitt et al., 2024). Although we cannot know the boundaries of what is in the training corpus without open models—and even then, the sheer volume of data may obscure clear boundaries—trusting results from simulations of settings well-captured by text on the internet is likely a sound heuristic.

[14] An alternative to effective prompting for improving the fidelity of simulations is to fine-tune the model on a particular dataset related to the target setting (Suh et al., 2025; Binz et al., 2025). However, if the goal is ideation, such fine-tuning sacrifices many of the lessons learned from identifying the appropriate theory-grounded agents in the first place. For now, it is also far more unwieldy and can lead to a phenomenon known as catastrophic forgetting—where the model "forgets" much of the original information from its original training data (Luo et al., 2023).

4 Conceptual issues

Critiques of *Homo silicus* question the suitability of using LLMs for social science. One potential response mirrors the argument of Friedman (1953) that the realism of *Homo economicus* and our assumptions more generally do not matter, and instead we should evaluate an approach on whether it can generate useful results. The proof of the pudding is in the eating, and "eating" for our purposes is whether AI simulations can help us to expand the frontiers of human knowledge more quickly. For example, suppose that the primary use of *Homo silicus* is to simulate experiments before we try them in the real world, and this method is adopted. In that case, these debates are somewhat moot. Nevertheless, it is helpful to consider non-Friedman responses to common critiques of AI simulations.

4.1 The "garbage in, garbage out" critique

One critique of using LLMs for social science is that their training data is too large to be carefully curated, and this may create a "garbage in, garbage out" problem. This critique rests partly on the incorrect notion that the responses of LLMs are a sort of weighted average. However, LLMs are more like random number generators than estimators. If an LLM is trained on millions of people reporting random draws from $U[0, 1]$, it would not always respond with approximately 0.5, but rather it would be more or less equally likely to return any number in $[0, 1]$. LLM developers also have strong incentives to curate the training data carefully (Brown et al., 2020; MetaAI, 2024), and post-training methods act as an additional layer of filtering.

Even great data may disproportionately represent the highly selected pool of "humans creating public writing." One advantage economists have in using LLMs is that they tend to pose questions that place few demands on the sample. 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 all humans engage in.[15] The willingness of economists to use undergraduates at elite four-year universities for laboratory experiments is partially a convenience—but also consistent with a disciplinary point of view we share with psychologists that it is not likely to matter much.

A more subtle problem is that LLMs are trained on what humans choose to say—not on what they choose to do. Economists have historically taken a dim view of the economic content of mere statements rather than behaviors. This would seem to be a damning critique of a model literally trained on statements. However,

[15] Although preliminary work shows how LLMs may effectively model non-WEIRD populations (Capra et al., 2024).

this "stated versus revealed preference" critique is only superficially persuasive. As we have discussed and illustrated in Figure 6, the internet contains both implicit and explicit sources of social information. The training data is not millions of lines of people lying about their reservation values in bargaining scenarios. Instead, much of it is text about people reasoning how to approach various economic questions, including "stage whispers" about their true intentions, explaining how they would deal with a situation, or recording economic decisions such as purchases and ratings.[16]

4.2 "Memorization" and "performativity" problems

With billions of parameters and a massive training corpus, one worry is that LLMs are parroting back to us what they have "read" somewhere in their training data. This view is not entirely correct. First, it is inconsistent with the fact that LLMs can "hallucinate" and make up new facts. Second, our findings in Section 2 contradict this view: the AI and human subjects' responses differ in scenarios that are in the training data. Moreover, memorization is not intrinsically bad.

Memorization would be problematic if the LLM's predictions were so "shallow" as to be brittle. To make an analogy, consider an economics student studying for a game theory exam. A bad student has memorized that {Defect, Defect} is the unique pure strategy Nash equilibrium for the prisoner's dilemma game, and then selects that prediction when it appears on the midterm. This shallow memorization is problematic because it does not generalize to games with different payoffs or action spaces. A better student would memorize that a Nash equilibrium is a strategy profile where every player best-responds to every other player, and would be able to apply this concept to identify equilibria in novel settings. The usefulness and flexibility of LLMs suggests that their predictions are often more like the good student's.

Another related concern is a potential "performativity" problem, in the sense that AI agents may behave following the theories and empirical results in their training data (MacKenzie, 2007). However, as with the "good" game theory student, performativity may be a desirable feature of AI simulations—assuming the underlying theory is correct. More generally, a critical view of memorization and performativity overshadows their important benefits. If our goal is to have a "realistic enough" model of human behavior, rather than to make unequivocal inference about human behavior, and we believe that the cumulative body of social science

[16]The parent of one of the authors runs a construction company and has to negotiate constantly. He claims that 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). 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.

is useful in predicting actual human behavior, then internalizing and applying these facts is beneficial.

4.3 Black boxes with incoherent world models

We do not fully "understand" the behavior of LLMs today, including how training data, model architectures, and prompts affect model outputs. A deeper understanding would certainly be useful, but it is not necessary to use LLMs for social science. To do economics, we do not have to study neurons and parts of the brain. Instead, we advance our body of knowledge by studying how humans respond in different scenarios and by developing theories, one study at a time. Similarly, what matters is not that we fully understand how LLMs work, but rather that they are systems created for some understandable purpose with an explicit goal of optimization. As Simon (1996) notes, "sciences of the artificial" can usefully abstract away from the micro-details of construction so long as the created object is viewed as trying to maximize something subject to the constraints of the environment.

A related concern is that LLMs may have incoherent world models, which can undermine their credibility. An interesting case in point is Vafa et al. (2024b). They train a language model to predict paths between points in New York City based on rides from taxi drivers. To do this, they label every single intersection in the city with a unique identifier. Each ride is then tokenized as an initial intersection, an ending intersection, and a series of actions—"left," "right," "straight," etc.—between them. When trained on thousands of these rides, the model can predict the correct sequence of actions to get between hold-out sets of starting and ending intersections with high accuracy. Yet, the authors show that the inferred map of New York City omits streets, invents others, and the predictions are far less accurate when street detours are introduced. The language model's world model is brittle. It is not explicitly designed to be a map of the city.[17]

While a perfectly coherent world model would be a boon to AI simulations, it is not strictly necessary to have credible predictions in many contexts. Effective simulations can rely on what the LLM was designed to do—follow instructions. Consider an LLM being used to predict how humans might navigate some city. Given a perfect GPS, this would be trivial. However, the LLM might still navigate well if given a broadly applicable set of instructions and updated information about the immediate environment. If it were endowed with *"Stop at stop signs. Stay on the right side of the road. Follow the speed limit."* and intermittent updates

[17] Vafa et al. (2025) offer another related context focusing on Newtonian mechanics. Some points largely ignored by both papers are the now-well-established scaling-and-breadth effects: LLMs trained on broad, heterogeneous tasks tend to have the best performance across the board (Kaplan et al., 2020; Brown et al., 2020; et al., 2022; OpenAI, 2024). It is unclear how the most capable LLMs suffer from the incoherencies in Vafa et al. (2024b).

about the immediate environment, it could move around reasonably. Navigating with those instructions is a far simpler task—and requires a far simpler world model—than reconstructing the entire street grid from scratch, and it is robust to many small perturbations.[18]

4.4 Are these "just simulations?"

An objection to AI-based experiments is that they are just simulations or agent-based models. Economists generally take a dim view of simulation-based approaches because the researcher is both judge and jury: you program the agents and then see what they do. Rather than "What would *Homo economicus* do?" simulation-based approaches often ask "What would this model [that does what I tell it to do] do?"—a far less scientifically interesting question. This is arguably why Schelling (1971) is the exception that proves the rule. His model was so simple and obvious that readers knew there was no card up his sleeve, no trick to ensure the surprising emergent phenomenon.

In contrast to conventional agent-based models, *Homo silicus* is not under direct researcher control. We can influence *Homo silicus* with endowments of beliefs, experiences, and so on (see Section 2). But we remain constrained by the fact that what determines behavior is the underlying model, not our direct programming.

Homo silicus also gives researchers a much greater degree of flexibility than agent-based models in the past. With LLMs, we can map preferences, beliefs, and so on, to purposeful actions in the simulated social world; we can "program" complex simulations only using natural language; and we can often parse the LLMs' responses to study the reasoning behind their choices. These factors expand greatly the scope of AI simulations.

The flexibility of *Homo silicus* makes it less susceptible to one of the core criticisms of both conventional agent-based models and economic theory itself: the Lucas critique (Lucas, 1976).[19] At its core, the Lucas critique emphasizes that behavioral rules guiding agents cannot be treated as fixed parameters but as endogenous responses shaped by the prevailing policy environment. LLM-based agents mitigate this concern because they can engage in flexible reasoning about environmental changes rather than following

[18] This perspective may also shed light on why many AI simulations in the literature have such poor fidelity. Much of this research focuses on the prompts as simple social and demographic traits (Park et al., 2024; Santurkar et al., 2023; Atari et al., 2023; Röttger et al., 2024). Such "instructions" like *"respond as a 30-year-old male"* or *"respond like a professor from MIT"* assume the LLM can coherently represent how those traits interact with the strategic setting. When its world model is shaky, as may often be the case for complex constructs like the interactions of human identities, the simulation may deteriorate or default to caricatures of the target persona (Cheng et al., 2023).

[19] This may help explain why agent-based models have had relatively little uptake in mainstream economics.

hard-coded behavioral rules.

4.5 It is easy to "prompt hack" AI experiments

The technical requirements of AI simulations are minimal. This low barrier to entry, however, comes with the risk of researchers iterating through many prompts until they get a desired response and report only favorable results. This is conceptually similar to the problem of p-hacking in traditional experimental work.

There are two easily implementable solutions to this problem. First is to require researchers to present their results under many different partially controlled permutations of the prompt as a robustness check. For example, in the Kahneman et al. (1986) experiment, we repeated the experiment with different temperatures, translated the prompt into other languages and then back to English, and varied the context of the experiment. These manipulations varied the data-generating process in a way that is partially but not entirely within our control. The second solution, which we have implemented in all experiments throughout this paper, is to make all simulation code public. Then anyone can inspect the exact data-generation process underlying any analysis. Indeed, this transparency and portability are major advantages of AI relative to human subjects. Replicating experiments often only requires a few dollars and the click of a button.

Both solutions require little of the researcher beyond a willingness to share their work—a willingness they should have anyway. Coding up alternative simulations is straightforward. It requires a few minutes of additional prompting and some minor code changes. Sharing results is also easy. Frontier LLMs can clean up code syntax and provide comprehensive documentation for an entire codebase with a single prompt.

4.6 Causal inference with AI simulations is challenging

Two features of LLM experiments ostensibly make causal analysis difficult. First, an LLM is a single model, and so it would seem that we can only get one "observation." Unlike the one *Homo economicus* that is rational, there are many *Homo silici*. We demonstrated in Section 2 that we can induce LLMs to play different agents via prompts, leading to varying responses.[20] This agent "programming" is not unlike the experimental economics practice of giving a subject a card that says their marginal cost of producing a widget is 15 tokens. Additional sources of variation may come from setting the temperature parameter and from using different LLMs.

[20] Argyle et al. (2023) make this point in their perfectly titled "Out of one, many" paper—there is not a single LLM but rather a model capable of being conditioned to take on different personas that respond realistically.

Which set of *Homo silici* to use depends on the researcher's question of interest. Researchers should not just use off-the-shelf LLMs to conduct simulations, but rather construct a set of agents relevant to their domain of study. This may require matching simulations to existing data or grounding the agents on existing theories.

A second challenge concerns identification. Even with perfect randomization, all else may not be equal when comparing treatment groups in AI experiments (Gui and Toubia, 2023). The core challenge is that editing a prompt to change one factor may inadvertently cause other factors to change; the experiment may then not manage to isolate a single cause. 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.

External validity is a problem for experiments both with human and AI agents. Consider the prompt parameterized by the wage asks of the two applicants in our dishwashing experiment. The experimental design iterated over combinations of the two wage asks, 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. When we repeated the experiment with a reference wage of \$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. Without a reference wage, the LLM may impute values for the reference wage as a function of the wage asks learned during training. And if the LLM's response distribution changes with the reference wage, 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 based on the hiring prompt with no reference wage, the experiment is actually based on the hiring prompt with an imputed reference wage that depends on the wage asks. Conversely, when the reference wage is \$12, the reference wage is specified, so there is no value to impute.

Now, suppose we repeated the Horton (2025) 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 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 potential outcomes.

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 (Levitt and List, 2007). In these cases, it is incumbent upon the researcher to justify the generalizability of the results (Camerer, 2015). They can do this through theory, robustness checks, additional experiments, etc. We can use AI simulations 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 negligible marginal cost (Shahidi et al., 2025). *Homo silicus* can provide a powerful check on the assumptions researchers may often take for granted and help us identify possible threats to generalizability.

5 Conclusion

This paper reports the results of several experiments using dozens of LLMs as simulated experimental subjects. We recover core qualitative patterns observed in the analogous human-subjects experiments and surface deviations that are themselves likely informative for future research. These demonstrations highlight our essential conceptual contribution: *Homo silicus* is best understood as theory in flexibly executable form. AI simulations can also be a fast, low-cost way to explore ideas, stress-test experimental designs, and calibrate power calculations before committing to costly human data collection.

We make a practical contribution by releasing open-source software that standardizes the design and implementation of AI experiments. It allows researchers to switch in any model to run an experiment, making it easy to update results even as models evolve. All experiments in this paper are now button-push reproducible, with extensive documentation. Our view is that open, reproducible research can be facilitated by pushing the costs of replication, transparency, and distribution toward zero.

In terms of future work, the most pressing task is to better understand when AI simulations are high-fidelity: to sketch the families of settings and assumptions where they track *Homo sapiens*, and to be candid

about where they do not. This missing map is the key limitation today. However, there is already progress on this front. For now, we can say that AI simulations are most informative when mechanisms can be written as clear instructions, the domain is well represented in digital text, and researchers can benchmark on related tasks with human data when possible. As we improve our mapping, future opportunities will emerge. One particularly promising direction is automated scientific loops: systems that propose hypotheses, instantiate agents, run and score simulations, and iterate (Manning et al., 2024). Such systems could dramatically accelerate research productivity.

References

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

Akerlof, George A. and Janet L. Yellen, "Can Small Deviations from Rationality Make Significant Differences to Economic Equilibria?," *The American Economic Review*, 1985, *75* (4), 708–720.

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

Andrews, Isaiah, Drew Fudenberg, Lihua Lei, Annie Liang, and Chaofeng Wu, "The Transfer Performance of Economic Models," 2025.

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic, "Prediction-powered inference," *Science*, 2023, *382* (6671), 669–674.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate, "Out of One, Many: Using Language Models to Simulate Human Samples," *Political Analysis*, 2023, *31* (3), 337–351.

Arrow, Kenneth J., "The Economic Implications of Learning by Doing," *The Review of Economic Studies*, 1962, *29* (3), 155–173.

Atari, M., M. J. Xue, P. S. Park, D. E. Blasi, and J. Henrich, "Which Humans?," Technical Report, Arxiv 09 2023. https://doi.org/10.31234/osf.io/5b26t.

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan, "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," 2022.

Banki, Daniel, Uri Simonsohn, Robert Walatka, and George Wu, "Decisions under Risk Are Decisions under Complexity: Comment," *SSRN Electronic Journal*, 02 2025.

Binz, Marcel and Eric Schulz, "Turning large language models into cognitive models," 2023.

_ , Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Natalia Scharfenberg, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz, "A foundation model to predict and capture human cognition," *Nature*, 2025, *644* (8078), 1002–1009.

Bowman, Samuel R., "Eight Things to Know about Large Language Models," 2023.

Brand, James, Ayelet Israeli, and Donald Ngwe, "Using GPT for Market Research," *Harvard Business School Marketing Unit Working Paper*, March 2023, *23-062*.

Broska, David, Michael Howes, and Austin van Loon, "The Mixed Subjects Design: Treating Large Language Models as Potentially Informative Observations," MIT Sloan Research Paper No. 7154-24, MIT Sloan School of Management January 31 2025. Available at SSRN: https://ssrn.com/abstract=5133034 or http://dx.doi.org/10.2139/ssrn.5133034.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, "Language Models are Few-Shot Learners," in "Advances in Neural Information Processing Systems," Vol. 33 2020.

Bui, Ngoc, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying, "Mixture-of-Personas Language Models for Population Simulation," 2025.

Bybee, J. Leland, "The Ghost in the Machine: Generating Beliefs with Large Language Models," Working Paper, University of Chicago Booth School of Business February 2025.

Camerer, Colin F., "The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List," in "Handbook of Experimental Economic Methodology," Oxford University Press, 02 2015.

Capra, C. Monica, Augusto Gonzalez-Bonorino, and Emilio Pantoja, "LLMs Model Non-WEIRD Populations: Experiments with Synthetic Cultural Agents," SSRN Electronic Journal December 2024.

Chang, Serina, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec, "LLMs generate structurally realistic social networks but overestimate political homophily," 2024.

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

_ , B. Jabarian, and John A. List, "The next generation of experimental research with LLMs," *Nature Human Behaviour*, 2025, *9*, 833–835.

Cheng, Myra, Tiziano Piccardi, and Diyi Yang, "CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations," *ArXiv*, 2023, *abs/2310.11501*.

Crawford, Vincent P. and Joel Sobel, "Strategic Information Transmission," *Econometrica*, 1982, *50* (6), 1431–1451.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt, "Predict science to improve science," *Science*, 2019, *366* (6464), 428–429.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde, "Are Risk Aversion and Impatience Related to Cognitive Ability?," *American Economic Review*, June 2010, *100* (3), 123–60.

Duckworth, Angela L., Ahra Ko, Katherine L. Milkman, Joseph S. Kay, Eugen Dimant, Dena M. Gromet, Aden Halpern, Youngwoo Jung, Madeline K. Paxson, Ramon A. Silvera Zumaran, Ron Berman, Ilana Brody, Colin F. Camerer, Elizabeth A. Canning, Hengchen Dai, Marcos Gallo, Hal E. Hershfield, Matthew D. Hilchey, Ariel Kalil, Kathryn M. Kroeper, Amy Lyon, Benjamin S. Manning, Nina Mazar, Michelle Michelini, Susan E. Mayer, Mary C. Murphy, Philip Oreopoulos, Sharon E. Parker, Renante Rondina, Dilip Soman, and Christophe Van den Bulte, "A national megastudy shows that email nudges to elementary school teachers boost student math achievement, particularly when personalized," *Proceedings of the National Academy of Sciences*, 2025, *122* (13), e2418616122.

Egami, Naoki, Musashi Hinck, Brandon Stewart, and Hanying Wei, "Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models," in A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds., *Advances in Neural Information Processing Systems*, Vol. 36 Curran Associates, Inc. 2023, pp. 68589–68601.

Elhage, Nelson, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah, "A Mathematical Framework for Transformer Circuits," *Transformer Circuits Thread*, 2021.

Enke, Benjamin and Cassidy Shubatt, "Quantifying Lottery Choice Complexity," Working Paper 31677, National Bureau of Economic Research September 2023.

et al., Victor Sanh, "Multitask Prompted Training Enables Zero-Shot Task Generalization," in "International Conference on Learning Representations" 2022.

Fehr, Ernst and Klaus M. Schmidt, "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 1999, *114* (3), 817–868.

Fish, Sara, Yannai A. Gonczarowski, and Ran I. Shorrer, "Algorithmic Collusion by Large Language Models," 2025.

Friedman, Milton, *The Methodology of Positive Economics*, Chicago: University of Chicago Press, 1953.

Fudenberg, Drew and Annie Liang, "Predicting and Understanding Initial Play," *American Economic Review*, December 2019, *109* (12), 4112–41.

Gandhi, Linnea, Angela L. Duckworth, and Benjamin S. Manning, "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*, 2024.

_ , Anoushka Kiyawat, Colin Camerer, and Duncan J. Watts, "Hypothetical nudges provide directional but noisy estimates of real behavior change," *Communications Psychology*, 2025, *3*, 158.

Gneiting, Tilmann and Adrian E Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American statistical Association*, 2007, *102* (477), 359–378.

Gui, George and Olivier Toubia, "The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective," *SSRN Electronic Journal*, 2023.

Hewitt, Luke, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer, "Predicting Results of Social Science Experiments Using Large Language Models," *Preprint*, August 2024. *Equal contribution, order randomized.

Hirasawa, Toshihiko, Michihiro Kandori, and Akira Matsushita, "Using Big Data and Machine Learning to Uncover How Players Choose Mixed Strategies," June 2022. Preliminary.

Horton, John J., "Price Floors and Employer Preferences: Evidence from a Minimum Wage Experiment," *American Economic Review*, January 2025, *115* (1), 117–46.

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

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

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling Laws for Neural Language Models," 2020.

Korinek, Anton, "Generative AI for Economic Research: Use Cases and Implications for Economists," *Journal of Economic Literature*, January 2023, *61* (4), 1281–1317.

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

Leng, Yan, Yunxin Sang, and Ashish Agarwal, "Reduce Disparity Between LLMs and Humans: Optimal LLM Sample Calibration," April 20 2024. Available at SSRN: https://ssrn.com/abstract=4802019 or http://dx.doi.org/10.2139/ssrn.4802019.

Levitt, Steven D and John A List, "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?," *Journal of Economic Perspectives*, April 2007, *21* (2), 153–174.

Li, Peiyao, Noah Castelo, Zsolt Katona, and Miklos Sarvary, "Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis," *Marketing Science*, 2024, *0* (0), null.

Lilleholt, Lau, "Cognitive ability and risk aversion: A systematic review and meta analysis," *Judgment and Decision Making*, 2019, *14* (3), 234–279.

Lucas, Robert E., "Econometric policy evaluation: A critique," *Carnegie-Rochester Conference Series on Public Policy*, January 1976, *1* (1), 19–46.

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

Ludwig, Jens and Sendhil Mullainathan, "Machine Learning as a Tool for Hypothesis Generation," *The Quarterly Journal of Economics*, January 2024, p. qjad055. _eprint: https://academic.oup.com/qje/advance-article-pdf/doi/10.1093/qje/qjad055/56324173/qjad055.pdf.

_ , _ , and Ashesh Rambachan, "Large Language Models: An Applied Econometric Framework," Working Paper 33344, National Bureau of Economic Research January 2025.

Luo, Yun, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang, "An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-Tuning," *IEEE Transactions on Audio, Speech and Language Processing*, 2023, *33*, 3776–3786.

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

Manning, Benjamin S. and John J. Horton, "General Social Agents," 2025.

_ , Kehang Zhu, and John J. Horton, "Automated Social Science: Language Models as Scientist and Subjects," NBER Working Paper 32381, National Bureau of Economic Research 2024.

McCoy, R. Thomas, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths, "Embers of autoregression show how large language models are shaped by the problem they are trained to solve," *Proceedings of the National Academy of Sciences*, 2024, *121* (41), e2322420121.

MetaAI, "Introducing LLaMA 3," 2024. Accessed: 2024-08-29.

Milkman, Katherine L., Dena Gromet, Hung Ho, Joseph S. Kay, Timothy W. Lee, Pepi Pandiloski, Yeji Park, Aneesh Rai, Max Bazerman, John Beshears, Lauri Bonacorsi, Colin Camerer, Edward Chang, Gretchen Chapman, Robert Cialdini, Hengchen Dai, Lauren Eskreis-Winkler, Ayelet Fishbach, James J. Gross, Samantha Horn, Alexa Hubbard, Steven J. Jones, Dean Karlan, Tim Kautz, Erika Kirgios, Joowon Klusowski, Ariella Kristal, Rahul Ladhania, George Loewenstein, Jens Ludwig, Barbara Mellers, Sendhil Mullainathan, Silvia Saccardo, Jann Spiess, Gaurav Suri, Joachim H. Talloen, Jamie Taxer, Yaacov Trope, Lyle Ungar, Kevin G. Volpp, Ashley Whillans, Jonathan Zinman, and Angela L. Duckworth, "Megastudies improve the impact of applied behavioural science," *Nature*, December 2021, *600* (7889), 478–483.

Olah, Chris, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Shan Carter, and Michael Petrov, "Zoom In: An Introduction to Circuits," *Distill*, 2020.

OpenAI, "GPT-4 Technical Report," 2024.

Oprea, Ryan, "Complexity and Its Measurement," 2024. Prepared for the *Handbook of Experimental Methods in the Social Sciences*.

_ , "Decisions under Risk Are Decisions under Complexity," *American Economic Review*, 2024. Available online at https://www.aeaweb.org/journals/aer/forthcoming.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe, "Training Language Models to Follow Instructions with Human Feedback," in "Advances in Neural Information Processing Systems" 2022.

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein, "Generative Agent Simulations of 1,000 People," 2024.

Pearl, Judea, *Causality: Models, Reasoning and Inference*, 2nd ed., USA: Cambridge University Press, 2009.

Peterson, Joshua C., David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths, "Using large-scale experiments and machine learning to discover theories of human decision-making," *Science*, 2021, *372* (6547), 1209–1214.

Qian, Crystal, Kehang Zhu, John J. Horton, Benjamin S. Manning, Vivian Tsai, James Wexler, and Nithum Thain, "Strategic Tradeoffs Between Humans and AI in Multi-Agent Bargaining," 2025.

Röttger, Paul, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy, "Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models," in Lun-Wei Ku, Andre Martins, and Vivek Srikumar, eds., *Proceedings of the Association for Computational Linguistics*, Association for Computational Linguistics August 2024.

Samuelson, Paul A., "A Note on Measurement of Utility," *The Review of Economic Studies*, 1937, *4* (2), 155–161.

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

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto, "Whose opinions do language models reflect?," in "Proceedings of the 40th International Conference on Machine Learning" ICML'23 JMLR.org 2023.

Sarkar, Suproteem and Keyon Vafa, "Lookahead Bias in Pretrained Language Models," *SSRN Electronic Journal*, 06 2024.

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

Shah, Anand, Kehang Zhu, Yanchen Jiang, Jeffrey G. Wang, Arif K. Dayi, John J. Horton, and David C. Parkes, "Learning from Synthetic Labs: Language Models as Auction Participants," 2025.

Shahidi, Peyman, Gili Rusak, Benjamin S. Manning, Andrey Fradkin, and John J. Horton, "The Coasean Singularity: Demand, Supply, and Market Design with AI Agents," 2025.

Simon, Herbert A., *The Sciences of the Artificial, 3rd Edition* number 0262691914. In 'MIT Press Books.', The MIT Press, September 1996.

Suh, Joseph, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang, "Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions," 2025.

Tranchero, Matteo, Cecil-Francis Brenninkmeijer, Arul Murugan, and Abhishek Nagaraj, "Theorizing with Large Language Models," Working Paper 33033, National Bureau of Economic Research October 2024.

Vafa, Keyon, Ashesh Rambachan, and Sendhil Mullainathan, "Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function," 2024.

_ , **Justin Y. Chen, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan**, "Evaluating the World Model Implicit in a Generative Model," 2024.

_ , **Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan**, "What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models," 2025.

von Neumann, John and Oskar Morgenstern, *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press, 1944.

Wang, Mengxin, Dennis J. Zhang, and Heng Zhang, "Large Language Models for Market Research: A Data-augmentation Approach," 2025.

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

Xie, Yutong, Qiaozhu Mei, Walter Yuan, and Matthew O. Jackson, "Using Language Models to Decipher the Motivation Behind Human Behaviors," 2025.

Yun, Chulhee, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar, "Are transformers universal approximators of sequence-to-sequence functions?," *arXiv preprint arXiv:1912.10077*, 2019.

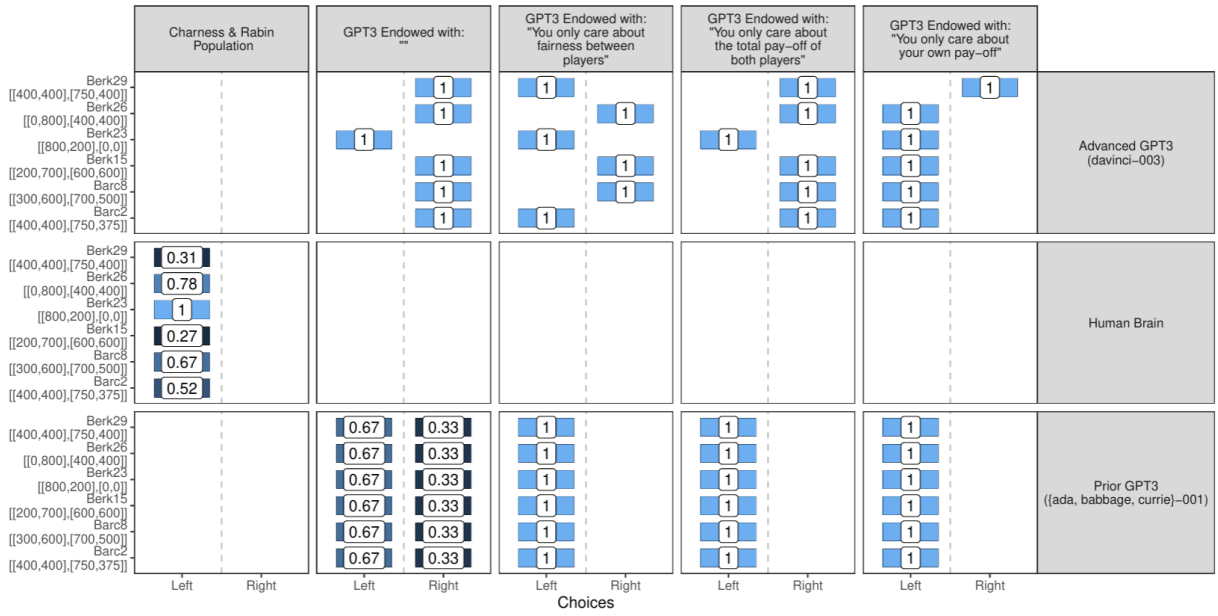
Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen, "A Survey of Large Language Models," 2025.

Zhu, Jian-Qiao, Hanbo Xie, Dilip Arumugam, Robert C. Wilson, and Thomas L. Griffiths, "Using Reinforcement Learning to Train Large Language Models to Explain Human Decisions," 2025.

_ , **Joshua C. Peterson, Benjamin Enke, and Thomas L. Griffiths**, "Capturing the complexity of human strategic decision-making with machine learning," *Nature Human Behaviour*, 2025.

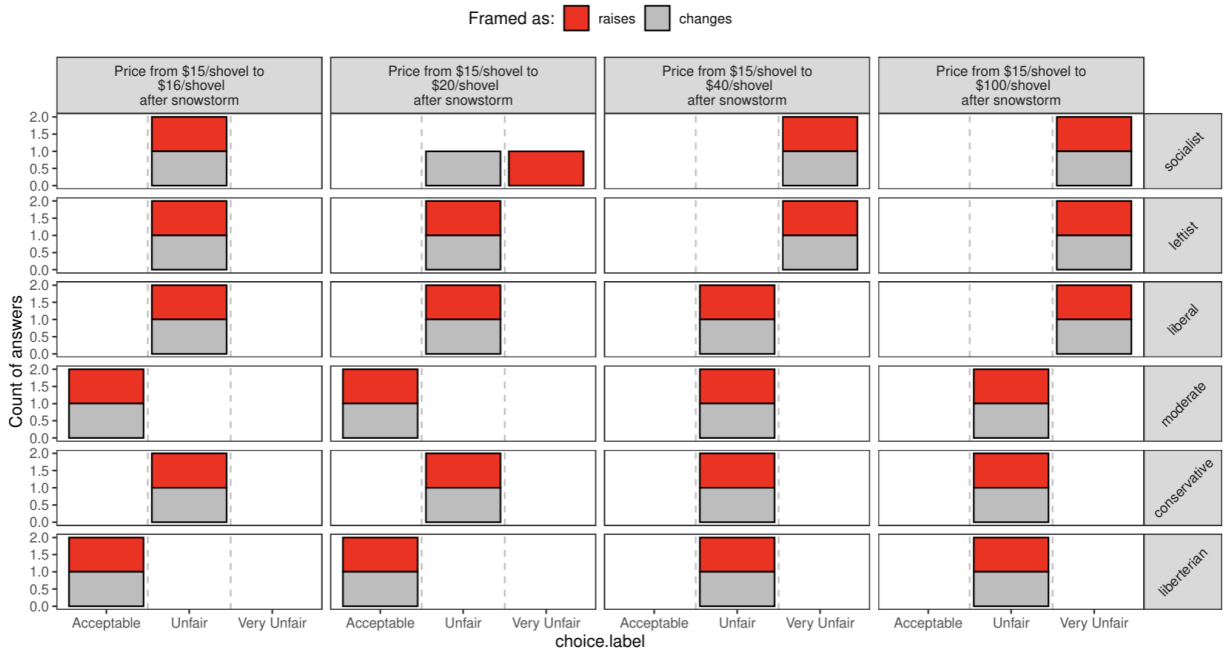
A Figures from the original draft

Figure A1: Charness and Rabin (2002) Simple Tests choices by model type and endowed "personality"



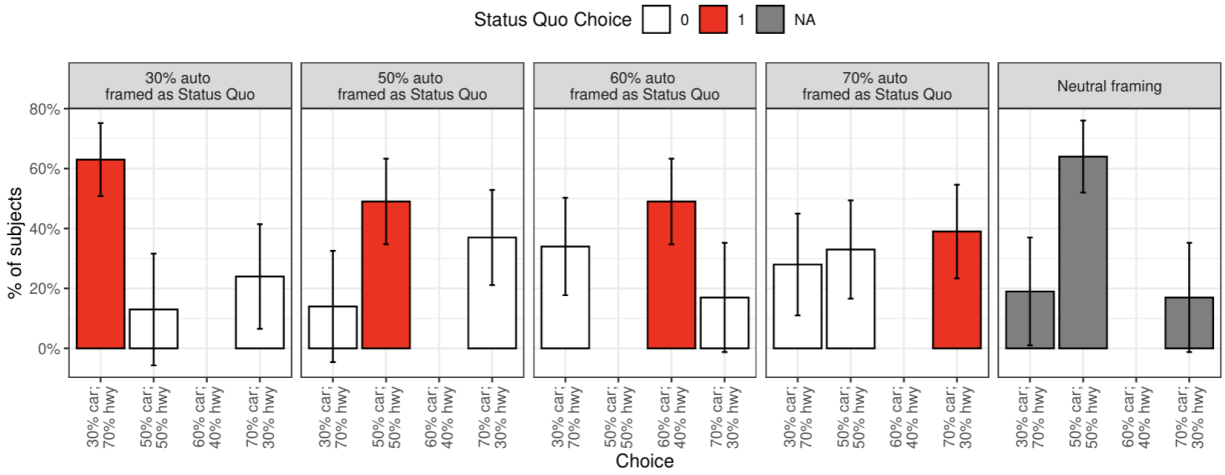
Notes: This shows the fraction of AI agents choosing each option, by framing.

Figure A2: Kahneman et al. (1986) price gouging snow shovel question, with endowed political views



Notes: This shows the fraction of AI agents choosing each opinion, by scenario.

Figure A3: Distribution of preferred car safety budgets, by status quo framing



Notes: This shows the fraction of AI agents choosing each option by framing.

Table A1: Effects of minimum wage on observed wage and hired worker attributes

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

Notes: This reports the results of imposing a minimum wage on the (1) hired worker wage and (2) hired worker experience.

B A brief primer on large language models

LLMs are neural networks with complex architectures and large numbers of parameters, estimated (trained) using massive amounts of text. Broadly speaking, they function as follows. 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.[21]

Training an LLM consists of a "pre-training" phase and a "reinforcement learning with human feedback" (RLHF) phase. In the pre-training phase, the LLM learns to predict the next word in sequences of text in its training data. 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.

After pre-training, the model may undergo one or more post-training stages. The most common post-training step is reinforcement learning with human feedback (RLHF), in which human evaluators rate model outputs, their judgments are used to train a reward model, and the LLM's parameters are adjusted so that its responses receive higher predicted rewards. Other post-training steps include supervised fine-tuning (or instruction-tuning), where human-written examples "teach" the model to follow prompts or exhibit a desired style. In the RLHF phase, the pre-trained LLM is trained further using human feedback. First, human evaluators prompt the LLM and evaluate its responses.[22] 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 fine-tuning to make its responses more consistent with a particular persona or writing style.

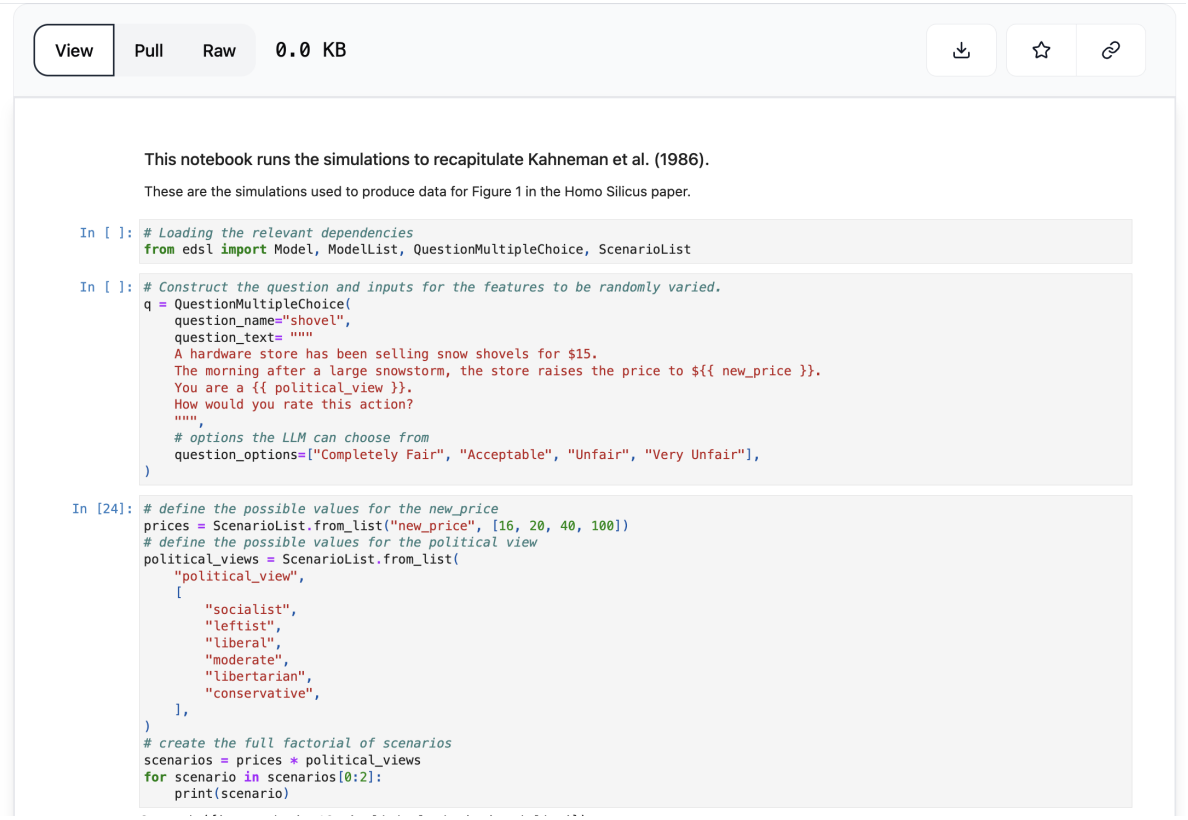
Users can adjust how the LLM responds to input prompts by changing its "temperature" parameter. The temperature parameter controls how the LLM samples from its response distribution. At temperature zero, the LLM is deterministic: it always outputs the highest-probability response. Increasing the temperature makes the output stochastic and the response distribution more uniform.

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

[22] 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.

C Additional Figures and Tables

Figure A4: Example screenshot of linked Jupyter notebooks for running AI simulations



Notes: This figure shows a screenshot of the Jupyter notebook for the Kahneman et al. (1986) experiment. All notebooks in this paper share a similar structure, with instructions and comments throughout the code. Each also includes links to share or download the notebooks so others can easily replicate the experiments or explore their own ideas.

Table A2: Estimates of the effects of political beliefs and price hike levels on AI agents' 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	−0.006*		−0.006*	−0.004*	−0.009*	−0.007*
	(0.001)		(0.0002)	(0.0003)	(0.0002)	(0.0003)
Socialist		−1.437*	−1.437*	−0.598*	−1.005*	−1.172*
		(0.030)	(0.027)	(0.038)	(0.027)	(0.030)
Leftist		−1.383*	−1.382*	−0.536*	−1.065*	−1.113*
		(0.030)	(0.027)	(0.038)	(0.027)	(0.030)
Liberal		−0.733*	−0.733*	0.115*	−0.453*	−0.653*
		(0.030)	(0.027)	(0.038)	(0.027)	(0.030)
Conservative		0.307*	0.307*	0.008	0.422*	0.522*
		(0.030)	(0.027)	(0.038)	(0.027)	(0.030)
Libertarian		0.860*	0.860*	0.848*	1.222*	1.205*
		(0.030)	(0.027)	(0.038)	(0.027)	(0.030)
Constant	2.462*	2.687*	2.860*	2.352*	2.734*	2.576*
	(0.025)	(0.021)	(0.020)	(0.029)	(0.020)	(0.022)
Experiment	Baseline	Baseline	Baseline	Translated	Variations	Adversarial
Observations	2,400	2,400	2,400	2,397	2,400	2,400
R^2	0.043	0.802	0.846	0.460	0.834	0.823

Notes: This table reports regressions where the independent variables are the snow shovel dollar price change relative to its original $15 price, and the AI agents' political beliefs. We use "Moderate" as the reference class. The dependent variable is the fairness assessment of the subjects, which we measure as a continuous variable ranging from 1 ("Very Unfair") to 4 ("Completely Fair"). The Δ Price coefficient represents the change in fairness rating (on the 1–4 scale) for each $1 increase in price above the $15 baseline. 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. Significance Indicator: *p<0.05.