General Social Agents*

Benjamin S. Manning
MIT

John J. Horton
MIT & NBER

September 23, 2025

MOST RECENT DRAFT [HERE]

Abstract

Useful social science theories predict behavior across settings. However, applying a theory to make predictions in new settings is challenging: rarely can it be done without ad hoc modifications to account for setting-specific factors. We argue that AI agents put in simulations of those novel settings offer an alternative for applying theory, requiring minimal or no modifications. We present an approach for building such "general" agents that use theory-grounded natural language instructions, existing empirical data, and knowledge acquired by the underlying AI during training. To demonstrate the approach in settings where no data from that data-generating process exists—as is often the case in applied prediction problems—we design a heterogeneous population of 883,320 novel games. AI agents are constructed using human data from a small set of conceptually related but structurally distinct "seed" games. In preregistered experiments, on average, agents predict initial human play in a random sample of 1,500 games from the population better than (i) a cognitive hierarchy model, (ii) game-theoretic equilibria, and (iii) out-of-the-box agents. For a small set of separate novel games, these simulations predict responses from a new sample of human subjects *better* even than the most plausibly relevant published human data.

*Thanks to Tyler Cowen and the Mercatus Center for generous funding and intellectual support. Thanks to Alex Moehring, Daniel Rock, David Holtz, Drew Fudenberg, Jenny Allen, Jessica Hullman, John List, Kehang Zhu, Leland Bybee, Michael Zhao, Seth Benzell, Sophia Kazinnik, Soumitra Shukla, and Steve Tadelis for their time and helpful comments. We are deeply grateful to Reanna Ishmael for software development support. The experiments in this paper were preregistered on https://aspredicted.org/ numbers 222695, 231091, and 241394. Author contact information, code, and data are currently or will be available at http://www.benjaminmanning.io/. Both authors have a financial interest in https://www.expectedparrot.com/. Horton is an economic advisor to Anthropic. 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

A general, low-cost method for accurately simulating human behavior with AI agents would have wide application in the social sciences (Charness et al., 2025; Jackson et al., 2025; Anthis et al., 2025). Recognizing this potential, a growing literature explores whether large language models (LLMs) can simulate human responses in various settings.[1] Across dozens of experiments, samples of these agents respond with remarkable similarity to humans—even when simulating novel studies that did not appear in the underlying LLM's training corpus (Hewitt et al., 2024; Binz et al., 2024; Li et al., 2024; Tranchero et al., 2024; Suh et al., 2025). Yet within this literature, others find settings where the very same models are poor human proxies.[2] This inconsistency poses a challenge for AI simulations as robust and credible predictive models—particularly in settings where no prior human data exists. The core challenge is not simply achieving a close match between AI and human responses in one setting, but building agents that will generalize reliably.

A natural starting point is improving the instructions given to agents. These instructions, or "prompts," are written descriptions given to the LLM specifying who it is, what it believes, or how it should behave and reason (Horton, 2023). Such second-person instructions (e.g., *"You respond as type-X person"*) can profoundly affect output distributions because advanced LLMs have been explicitly fine-tuned to follow instructions (Bai et al., 2022; Ouyang et al., 2022). With an appropriate prompt (which can be massive),[3] highly capable models can perform complex reasoning and mathematical tasks at levels sometimes better even than those of highly skilled humans.

Despite the existence of these powerful, steerable models, constructing agents whose behavior is similar to that of real humans in a wide variety of settings is nontrivial—even when human data are available to guide the search. The set of possible prompts is vast, ranging from simple combinations of social or demographic traits to complex programmatic instructions related to how humans make decisions (Zhu et al., 2025b; Xie et al., 2025). As in other machine learning applications, the challenge is not only to avoid underfitting but also to guard against overfitting. By iterating through enough prompts, one can almost always find some arbitrary prompt that shifts the LLM's responses to closely match a given human distribution. For example, an LLM instructed *"you randomly offer between $6 and $9"* may perfectly reproduce a distribution of human responses in a $20 dictator game, but such a prompt would be nonsensical for a $5 dictator game. In contrast, a prompt grounded in the underlying behavioral drivers—e.g., *"you are self-interested but fair"*—can perform well in-sample and plausibly extend to a range of allocation games. Standard data-driven approaches, such as a train-test split within a single dataset, cannot reliably distinguish between these two cases; the latter appears better only when tested in truly new settings. If the goal is to predict behavior in settings with no prior human data, how should researchers construct prompts?

[1] (Argyle et al., 2022; Aher et al., 2023; Binz and Schulz, 2023; Brand et al., 2023; Park et al., 2023; Mei et al., 2024; Park et al., 2024; Chang et al., 2024; Manning et al., 2024; Capra et al., 2024; Hansen et al., 2024; Kim et al., 2024; Wang et al., 2025; Cerina and Duch, 2025; Shah et al., 2025; Broska et al., 2025; Bybee, 2025; Fish et al., 2025)

[2] (Santurkar et al., 2023; Atari et al., 2023; Cheng et al., 2023; Gui and Toubia, 2023; Gao et al., 2024)

[3] To date, Google's Gemini 1.5 can accommodate 10 million tokens (Gemini Team, 2024). This is roughly equivalent to 15,000 pages, or about 20 copies of the Handbook of Experimental Economics (Kagel and Roth, 1995).

In this paper, we build agents whose behavior in simulations usefully generalizes to what we see from humans across entire domains. Our approach mirrors what researchers generally try to do in social science, but in reverse. Rather than testing a theory with empirical data, theory is embedded in agents (via natural language instructions), which are then used to generate candidate data. The theory and agent composition is then optimized to reduce error with respect to real-world data from a domain where predictions are desired. Human data from distinct but conceptually similar settings serve as held-out test sets or are incorporated into training to improve generalization. We show that "general" agents constructed and validated in this way can improve the predictive power of AI agents in novel settings. Both steps are essential: without theoretical grounding, optimized prompts may fail to meaningfully improve even in-sample predictions, and without cross-setting validation, they are prone to overfit.

The approach uses two kinds of data: (i) *training data*—existing human data used to optimize AI simulations, and (ii) *validation* or *test data*—existing human data related to but distinct from the training data, used to test whether the optimized agents generalize. The ultimate goal is to produce agents that can more accurately predict human behavior in novel *target settings* where no prior human data exist, but that lie in the same broad domain as the training and validation settings.

The first step of the approach is to limit the "space" of prompts to a subset motivated by some economic theory or causal mechanism relevant to the novel setting of interest. This theory-grounding is analogous to constraining the functional form of the hypothesis class in machine learning. Continuing with the dictator game example, where social preferences likely determine behavior, one might choose the candidate set of prompts characterized by known drivers of the relevant preferences (e.g., *"You are {level}"* for all *levels* $\in$ {*self-interested but fair, altruistic, selfish*}).

The second step is to optimize over this set to best match the human training data.[4] All candidate prompts are put in simulations of the settings that produced the training data (e.g., a $20 dictator game). Optimization effectively filters these candidate prompts down to a final subset whose simulated responses are closest to the human training distribution. We employ two optimization methods: (i) a selection method that identifies the optimal mixture of prompts from the candidate set (Leng et al., 2024; Xie et al., 2025; Bui et al., 2025), and (ii) a construction method that optimizes numerical parameters embedded directly in the prompts. If the final optimized set achieves a good in-sample fit, we have reason to believe they will generalize because they are also grounded in a relevant theory. Poor in-sample fit suggests a mismatch between the theory and the training setting or an inadequate operationalization of that theory. In this case, we revise the candidate prompts and re-optimize.

Given strong in-sample fit, we assess whether the final set of prompts generalize using a train-test split approach inspired by the principles of invariant risk minimization (Peters et al., 2016; Heinze-Deml et al., 2018; Arjovsky et al., 2020). The final set of prompts is placed in simulations of

[4]There is an active literature on optimizing prompts (Khattab et al., 2024).

the settings that produced the human validation data, where we also expect the underlying theory to hold (e.g., optimize on a $20 dictator game, but test on a $5 dictator game). To be clear, this means the validation set necessarily comes from a distinct data-generating process from the one that produced the training data. We then compare the predictions from these simulations to the relevant human distributions. By construction, sets of prompts with strong test performance—those that accurately predict the validation data—are then those that capture generalizable relationships predictive of human behavior across contexts. Those that fail validation likely do not. Consequently, if the novel target setting is governed by the same theory or causal mechanism used to construct the optimized prompts (e.g., the target is a $50 dictator game), we may gain confidence that they will better predict human responses in that setting.

We illustrate this approach and provide evidence of its efficacy with training and validation data drawn from experiments in the behavioral economics literature. All simulations use GPT-4o with the temperature set to 1, though the approach is agnostic to the choice of model and hyperparameters.[5] We first apply the selection method to Arad and Rubinstein (2012)'s 11-20 money request game, where participants request an amount and receive a bonus if they choose exactly one less than their opponent. We use only the original dataset from the paper, which contained fewer than 200 observations. This setting is appealing to study because optimal play depends not only on the focal agent's capabilities, but also on their beliefs about how others will reason. Endowing AI agents with distinct prompts corresponding to varying degrees of strategic reasoning (the theoretical focus of Arad and Rubinstein) produces a mixture that closely matches the original human data.

When we validate these optimized samples on distinct variants of the 11-20 money request game, they are better predictors of initial human play than baseline AI agents with no additional instructions. By contrast, scientifically meaningless or "atheoretical" AI agents derived from historical figures, pseudo-scientific Myers-Briggs personality types, and those instructed to select particular numbers can sometimes match human distributions in one variant of the game but fail to generalize across others.

We next test the predictive power of the optimized agents in target settings where no prior human data exist. To do so, we construct four new games and collect responses from samples of preregistered crowdsourced participants (Horton et al., 2011) on Prolific. These games are derived from the original 11-20 game (and its variants), but adapted to other numeric ranges (1-10 and 1-7). The optimized sample of theory-grounded prompts produces responses that predict the new human data far better than the off-the-shelf baseline. Prediction error is decreased by 53%-73% across the games. Furthermore, these simulations predict the results of the new experiments in some games *better* than the most plausibly relevant human data from Arad and Rubinstein; in one case, the KL divergence is halved. By contrast, the alignment of the atheoretical prompts—which failed validation—with the new human data is often similar to or worse than the baseline AI.

[5] GPT-4o, among the most widely used LLMs, and the temperature setting are defaults for the software used to run our simulations (Horton and Horton, 2024).

What statistical guarantees does this approach afford? Without a correctly specified causal model, no statistical procedure can guarantee performance in arbitrary new environments (Pearl, 2009). Formal guarantees with existing data, like those required for prediction-powered inference (Angelopoulos et al., 2023) and other related methods (Egami et al., 2023; Hardy et al., 2025), would require a strictly firewalled validation set: data never used in the construction of the underlying LLM (Ludwig et al., 2025; Sarkar and Vafa, 2024; Mullainathan and Spiess, 2017; Modarressi et al., 2025). This is a tall order impossible to meet in practice, even with existing public weight LLMs, never mind private models. However, what we can guarantee is prediction performance over a *pre-committed* family of settings.

The setup is very similar to the theoretical framework of Andrews et al. (2025), who study how well predictions from different economic and machine learning models transfer across economic domains. The first step is to establish a clearly defined space of experimental settings—for example, variants of public goods games, dictator games, or other allocation games differentiated by instructions, parameters, or other structural features. This idea is related to that in the literature on program evaluation and external validity, where population-level treatment effects can be estimated by randomizing over a defined set of possible samples (Hotz et al., 2005; Allcott, 2015; Dehejia et al., 2019). From this space, a random subset of settings is drawn and randomly assigned to human subjects, who provide responses.[6] By comparing these human responses with LLM-generated predictions, we can make externally valid estimates across the entire space.

A key distinction from the setups laid out in Hotz et al., Allcott, and Dehejia et al. is that appropriate coverage does not require that the family of settings share a data-generating process or that the underlying samples of human subjects are from the same population. Indeed, the variance of estimates naturally reflects how closely the held-out settings resemble those sampled. If all scenarios are variants of a single setting—e.g., a dictator game with different monetary amounts—performance is tightly bounded; if they span disparate domains—e.g., many types of allocation games—estimates are likely less precise. Note that even if a setting inadvertently appears in the model's pre-training data, the random sampling procedure still accounts for its contribution—such cases may merely reduce prediction error.

We explore this approach to inference by constructing a population of 883,320 novel strategic games. These were inspired by the original 11-20 game but were modified along several dimensions, such as support of possible choices, the size and nature of the bonus, the manner in which money is allocated, etc. The population of games is the full factorial combination of these dimensions. The resulting games differed substantively from each other; in some cases, they were cooperative, others zero-sum, many had a dominant strategy equilibrium, while some had no dominant strategies at all. The vast majority are surely not in any LLM's training corpus to date. From the population, we sample 1,500 unique games and have 4,500 human subjects each play a game in a final preregistered experiment. We take the theoretically-grounded level-k agents—agents constructed months before the novel set of 883,320 games existed—and evaluate their ability to predict the human responses.

[6] This is also similar to the integrative experiment designs of Almaatouq et al. (2024).

The theory-grounded agents are far better predictors of initial human play than the baseline AI off-the-shelf. In the average game, they assign 3.41 times more probability to the actions actually taken by human subjects. The optimized agents are similarly effective compared to two well-established theoretical benchmarks: (i) a symmetric Nash equilibrium calculated for each of the 1,500 games (2.44 times more probability), and (ii) the game-specific predictions from the cognitive hierarchy model of Camerer et al. (2004) (3.02 times more probability). Because the games were randomly sampled, the corresponding confidence intervals over the relative predictive power are externally valid for the much larger population.

The goal of AI simulations in this paper is to harness two extensive sources of information to create better predictive agents: (i) well-established theoretical models from the social sciences, and (ii) the immense knowledge about human behavior that LLMs have implicitly learned during training (Lindsey et al., 2025; Ameisen et al., 2025). In a sense, AI agents are a general vessel through which theory can be flexibly applied to any setting. Our approach aims to generate such agents by conducting a structured trial run across various settings, building evidence that theoretically-grounded prompts generalize effectively to similar yet distinct environments. These principles can be applied to any domain where relevant human data exists—even those with multiple interacting agents (Manning et al., 2024; Qian et al., 2025). One can imagine iteratively identifying novel scenarios where predictions are needed, finding the closest relevant existing data, and continuously optimizing and calibrating agents as needed.[7] Our main contribution is to systematize these ideas for constructing agents and then provide empirical evidence that they can substantially improve the predictive power of AI simulations in new settings.

The remainder of this paper is organized as follows. Section 2 begins with a concrete example of identifying generalizable relationships and constructing prompts that better predict human subjects in new settings. We then describe the approach more generally and specify the optimization methods. Section 3 illustrates the approach and demonstrates its efficacy empirically with several experiments. Section 4 demonstrates how agents can be used in a pre-committed setting to provide externally valid statistical estimates of predictive accuracy at scale. The paper concludes in Section 5.

2 Building prompts that generalize

This section explains the approach for constructing AI agents that better approximate human response distributions in new settings. We begin with an extended example of predicting behavior in a novel public goods game where no previous human data exists. We then discuss the importance of validating across distinct data-generating processes and grounding candidate prompts in relevant social science theories. The section concludes with a summary of the approach and a brief description of the optimization methods used to construct agents.

[7] LLMs could automatically construct candidate prompts as input to such feedback loops (Xie et al., 2025).

2.1 A motivating example

Suppose we want to predict how people will share resources in a novel public goods game (as they do in Alsobay et al. (2025)). In this new game, for which we have no previous data, three participants will be endowed with $5 and can choose to contribute any portion of their endowment, which will be multiplied by 3 and then divided equally among all participants. While we do not have any prior public goods game data, we do have human data from a related $20 dictator game. It is related in the sense that one might reasonably expect some generalizable features of human choice to affect allocations in both games. The observed human offers from the dictator game are {6,6,7,7,8,8,9,9}.

Before LLMs, one might have tried to train a standard machine learning model (e.g., decision trees or regression) solely on these dictator game offers. However, such models would struggle to transfer to the structurally distinct public goods game, as they cannot adapt across different game formats without retraining. LLMs offer a more flexible alternative. Rather than training a new model from scratch, we can prompt an existing LLM to simulate responses by instructing it to behave as a human participant. For instance, we might use a baseline system prompt: *"You are a human."* We can then append game-specific prompts such as *"You are playing a dictator game with $20..."* or *"You are playing a public goods game with $5..."* without any additional training.

Suppose that, when prompted with the dictator game instructions, the LLM produces a response distribution {3,3,3,3,3,4,4,4}. Although well within the allowable offers, this distribution clearly diverges from the observed human data ({6,6,7,7,8,8,9,9}) and leaves us with little hope that it could effectively predict responses to the novel public goods game. Thus, even though the model may capture some general aspects of human behavior as a baseline, i.e., a tendency to offer nontrivial amounts (Henrich et al., 2001), it does so imperfectly.

One might think this problem could be addressed by first randomly splitting the human sample of dictator game offers into training and testing sets. And then identifying the best-performing prompts on the training set, and validating their performance on the test set (Mullainathan and Spiess, 2017; Ludwig et al., 2025). Indeed, an LLM instructed *"You randomly choose numbers between 6 and 9"* could reasonably predict both training and testing sets for any split of the human data better than the baseline LLM with no persona. However, such an approach will almost certainly fail to generalize beyond the specific data-generating process that produced the dictator game offers. It has, in effect, still overfit. Rather than overfitting to the training data, it has overfit to the whole data-generating process.

Now, consider a more theoretically grounded prompt: *"You are self-interested but fair."* Suppose an LLM endowed with this prompt also generates offers in the 6-9 range for the $20 dictator game. Crucially, this prompt aligns with known causal factors that drive human sharing behavior more broadly, like behavioral parameters in empirically tested theoretical models (e.g., Charness and Rabin (2002)). It is a flexible decision-making program that can be effectively applied to various types of allocation games. Yet, without explicit knowledge of the causal drivers governing behavior in the new public goods game, nothing in the data produced by either prompt alone rules out the atheoretical random-number prompt. This illustrates the core challenge. We seek to

construct and select prompts that do not overfit to a single data-generating process, but instead capture the stable behavioral drivers relevant across settings. Then, we might gain confidence that they will better predict human responses in novel target settings governed by the same drivers.

2.2 Identifying generalizable behavioral relationships

As our motivating example illustrates, a traditional train-test split does not adequately guard against overfitting to a single data-generating process. This approach is statistically valid only when training and testing samples are independently drawn from the same distribution (Vapnik, 1998). However, our objective differs fundamentally: we seek prompts that remain predictive even when the underlying data-generating process shifts. By "data-generating process," we refer specifically to the experimental setting (the environment in which behavior occurs), the population from which behavior is observed, and the outcomes they generate. In econometric terms, two processes diverge when the given input covariates appear with different distributions of values or when the covariates themselves are entirely different random variables. For example, a model trained on \$20 dictator games may fail on \$5 dictator games due to different stake values (covariate shift) or when moving from dictator to public goods games (structural shift).

Without direct training data from the novel target setting, no standard statistical procedure ensures predictive accuracy (Ben-David et al., 2010; Klivans et al., 2024). Theoretically, only a fully specified causal model could guarantee accurate predictions (Pearl, 2009). In practice, however, constructing or inferring such causal models generally requires strong, often unverifiable assumptions, which are rarely feasible in complex social science contexts.

Instead, we propose leveraging principles from invariant risk minimization (Arjovsky et al., 2020) to identify behavioral relationships that remain stable despite shifts in the data-generating process. Rather than splitting a single dataset, we deliberately choose training data from one data-generating process (or several related processes) and validate using data from a related but distinct process from those used for training. Prompts that consistently predict behavior across these distinct datasets likely capture generalizable, and possibly causal (Peters et al., 2016; Heinze-Deml et al., 2018), drivers of behavior. As a result, these validated prompts should be more effective in predicting human responses in novel but theoretically similar settings.

Returning to our motivating example, suppose we have additional human data from a \$5 dictator game, with observed offers {1,1,1,2,2,2}. Although stakes differ, both the \$20 and \$5 dictator games likely share a common decision-making process: individuals give slightly less than half the available amount, a reasonable balance more similar to what is observed in humans (Henrich et al., 2001). By using the \$20 dictator game as training data and the \$5 dictator game for testing, we are implicitly filtering for prompts based on their capacity to predict this proportional response. The earlier atheoretical prompt *"You randomly choose numbers between 6 and 9,"* still fits the \$20 game perfectly, but clearly fails validation on the \$5 game. By contrast, the theory-grounded prompt *"You are self-interested but fair,"* likely generalizes effectively across both settings, and most importantly, the novel public goods game.

This validation process becomes even more robust when multiple distinct but theoretically-related datasets are available. Imagine dictator-game responses from \$1, \$5, \$10, \$20, \$50, and \$100 games, all exhibiting offers slightly below half. Optimizing over multiple settings simultaneously further reduces the risk of overfitting. With each new training data-generating process, it is less likely that an arbitrary prompt will generalize in-sample.[8] Thus, if the underlying relationship governing these settings also holds for a novel target setting, the validated prompts should robustly predict behavior there as well.

Yet, even with an effective validation method, a fundamental practical challenge remains: identifying the initial set of candidate prompts. While our motivating example made this step appear straightforward, selecting plausible prompts in more complex settings can be far less obvious. In a perfect world, an optimization procedure would take in a massive set of prompts and filter them down to a smaller set that generalizes. However, this does not yet work in practice because many different prompts can be used to generate the same behavior in a given setting. For example, the number of possible natural language prompts that get an LLM to respond with numbers between 6 and 9 is enormous. If we were to start from a large pool of prompts, then the optimization procedure might endlessly produce sets of prompts that fit in-sample but fail on the validation set. For now, it is essential to ground the candidate prompts with a principled starting point, which—we argue and later demonstrate empirically—economic and behavioral theories offer.

2.3 Theory as guide towards generalizability

A core function of economics is to construct models that capture causal and generalizable relationships that remain stable across environments. For example, the idea that humans make reference-dependent utility choices is not specific to one particular economic environment, but has been shown to broadly apply to decision-making under uncertainty (Kahneman and Tversky, 1979). Conveniently, these are exactly the types of generalizable relationships that we would expect to better predict human behavior in new settings when supplied to an LLM. Our approach is to narrow the search space of possible prompts by grounding candidate prompts in such theories. By doing so, we might increase the likelihood of identifying prompts that we have prior reason to believe reflect genuine, stable behavioral patterns. Without doing so, we risk dramatically underfitting and failing even to accurately predict the training data.

Economic theories—and theories from other social sciences that embrace methodological individualism (Friedman, 1953)—are well-suited to be cast into prompts, since good theory rests on the choices and actions of individuals. Specifically, we consider a prompt to be "theoretically grounded" when it instructs the LLM to make predictions about how a human would respond based on some underlying theory or model of human behavior. For example, a utility function that trades off one's own payoff against inequality might become *"You value your own earnings but dislike outcomes where you earn far more (or less) than others."* A highly capable and tool-using LLM could even

[8] We do not offer empirical examples of optimizing over multiple training settings in the main text (only single settings in Section 3). Appendix A provides such analyses, including an additional preregistered experiment.

be given that utility function and parameters directly (e.g., $u(x_{self}, x_{other}) = x_{self} - |x_{self} - x_{other}|$). A prospect-theoretic model reasonably maps to *"You are risk averse in gains, risk seeking in losses, and probability weight the very likely and very unlikely outcomes"*—and of course these could be parameterized as well.

This definition is necessarily broad because what constitutes a theory is also broad. One might consider good theories to be those that are parsimonious, predictive, and interpretable. We view these as effective guides for constructing prompts that are grounded in theory. But just as there is no universal rule as to what makes a theory good, there is not a singular all-encompassing playbook for candidate prompts. Here, the researcher must leverage their expertise appropriately to determine which theories are most relevant to the domain where predictions are desired. This should be relatively straightforward in practice. One can even simply ask an LLM to construct a candidate set of prompts based on a given paper and then adjust the prompts accordingly. For example, the linked notebook—which took 15 minutes to code—contains an example where we take PDFs of 5 well-known papers in the behavioral economics literature and generate 10 agents from each.[9]

It is worth highlighting just how flexible such agents are as predictive models. They can make predictions in response to *any* setting described in natural language. This is usually impossible for traditional machine learning and mathematical economic models, which operate within a fixed parameter space. One cannot introduce additional covariates to a regression model once it is trained. Similarly, a prospect-theoretic model for evaluating risky gambles cannot instantaneously incorporate other contextually relevant factors—the recent performance of financial markets, the demographic composition of the experimental sample, or even something as mundane as whether participants are making decisions before or after lunch. AI agents, by contrast, can interpolate over such details through natural language instructions. Whether these agents actually leverage this flexibility effectively is, of course, an empirical question, but one that is easily testable.

Once a set of prompts is defined, we can optimize them against an appropriate sample of relevant human decisions. This may involve searching for an optimal mixture of prompts, or adding continuous parameters directly in a single prompt and adjusting them until the simulated distribution aligns with the training data. From a machine learning perspective, the candidate set of prompts defines the functional form of our hypothesis class, with theory guiding this specification to effectively navigate the bias-variance tradeoff.[10] The set is then pruned (or tuned, depending on the method) to best fit the training data.

As an example, consider the extensive theoretical and empirical economic literature studying how people reason strategically (Stahl and Wilson, 1994, 1995; Arad and Rubinstein, 2012; Camerer et al., 2004). Level-k models, for instance, predict that individuals best respond based on their

[9] https://expectedparrot.com/content/32751dae-7a69-430d-9f1a-0c3f50ccef5b

[10] One could imagine an "oracle" prompt akin to a perfect program, describing every preference, heuristic, and belief update rule—allowing the LLM to accurately produce responses in all contexts for a person. In effect, our approach is to identify portions of this oracle that are most relevant to the given training and testing data. Increasing context windows suggest that such highly generalizable agents may be possible in the not-so-distant future.

beliefs about others' levels of reasoning. In the $\frac{2}{3}$ guessing game, players aim to pick $\frac{2}{3}$ of the average number chosen by the group (Nagel, 1995; Keynes, 1936). If we had human data from such a game, we could construct a series of prompts that specify different levels of strategic thinking (e.g., *"You are a level-0 reasoner," "You are a level-1 reasoner,"* etc.). We then identify the mixture that best fits the training distribution and test whether it generalizes to other variants of the game (e.g., different values than $\frac{2}{3}$). Then, the optimized prompts can be used to predict other related distributions in new settings.

Of course, constructing and validating these prompts is not a perfectly specified problem. There are various ways to translate a theory into natural language, and multiple theories may apply to a given setting. The boundaries of a theory and the settings it plausibly governs are not always well-defined. Nor can we guarantee that the data-generating processes of the training and testing sets are either sufficiently similar or sufficiently distinct to yield reliable validation. Yet, as we will show empirically, these challenges can be overcome in practice, at least in some domains.

Ultimately, applying a prompt—or a set of prompts—to a new environment requires an unavoidable inductive leap. What we can do is try to make the leap explicit and interpretable. LLMs are extremely good at following explicit, well-defined instructions. Because each prompt is a set of these flexible natural language instructions, researchers can both evaluate performance and reasonably assess its relevance for a new setting.[11] This mirrors how economic theory is used more broadly with real humans: it is tested against data, observed where it succeeds or fails, and cautiously extrapolated to settings just beyond current evidence. When a new tariff is proposed, for example, the theory motivating the tariff is not known to be universally "correct" in a complex dynamic world ex ante. It is a disciplined guess, shaped by assumptions and previous evidence, which the economist then uses to make predictions about the world. Theory-guided prompts, validated across environments, bring the same kind of disciplined reasoning to simulated AI agent predictions of human behavior.

2.4 A brief summary of the approach

We can summarize our approach into the following steps.

1. **Select Training and Testing Data.** Identify distinct samples of human-generated data that are presumably generated by the same mechanisms as the novel target setting(s) of interest. When possible, use multiple distinct datasets for both training and testing to increase confidence and minimize the possibility of selecting spurious prompts.

2. **Propose Theory-Driven Candidate prompts.** Generate a broad set of prompts that are plausibly consistent with the proposed theory (or causal mechanisms if known) related to the training, testing, and novel settings.

[11] Such flexibility also makes AI agents less susceptible to the Lucas critique (Lucas, 1976)—if at all. Unlike traditional agent-based models or mathematical theories, which have hard-coded predetermined processes to generate predictions, LLMs can, as we will show, interpret new policies in context, reason through their implications, and adaptively formulate responses.

3. **Optimize prompts on Training Data.** Optimize the candidate prompts to best match the training data. This might involve selecting a mixture of prompts or adjusting trait parameters to minimize some statistical distance from observed responses. Confirm that the optimized sample outperforms the baseline LLM off-the-shelf on the training data.

4. **Validate prompts on Testing Data.** Apply the optimized prompts to the testing data and evaluate their performance relative to the baseline LLM.

Broadly speaking, our approach provides a disciplined "trial run" to confirm whether a given sample of AI agents can reliably predict human behavior across multiple related settings. Similar to applying economic models estimated from past data to inform predictions in new but structurally similar environments, the success of our approach depends critically on identifying and leveraging underlying stable behavioral relationships.

In Section 3, we demonstrate the approach empirically on a set of strategic games using training data from Arad and Rubinstein (2012). Appendix A provides another example using training data from (Charness and Rabin, 2002) and allocation games. Both applications show that the approach substantially reduces prediction error relative to baseline LLM predictions in preregistered experiments with novel human samples. We emphasize throughout that the two key elements of the approach—grounding candidate prompts in economic or behavioral theories, and validating across distinct but related datasets—each independently address critical pitfalls. Without theoretical grounding, optimized prompts may fail to meaningfully improve even in-sample predictions; without validating across multiple related settings, optimized prompts are prone to overfitting a single training context.

2.5 Methods for optimizing prompts in-sample

We briefly describe two methods for optimizing prompts with a given set of training data. The first, the selection method, assumes we have a finite library of candidate prompts and selects (or mixes) them to best fit the training data. We apply this method to the experiments presented in Section 3 with data from Arad and Rubinstein (2012). An early version of this idea was suggested by Horton (2023), and several others have since explored applications (Leng et al., 2024; Xie et al., 2025; Bui et al., 2025). The second, the construction method, parameterizes a prompt template with numeric trait dimensions and optimizes those parameters. To the best of our knowledge, this method is novel. An application of the method is presented in Appendix A using data from Charness and Rabin (2002).

Selection Method. A finite set of unique candidate natural language prompts is first specified. For each prompt $\theta \in \Theta$, the LLM is used to generate a predictive distribution $\hat{P}_\theta$. Let P represent the observed ground-truth human distribution. The objective is to solve

$$\min_{\mathbf{w}} \; d\Big(P, \sum_{\theta \in \Theta} w_\theta \, \hat{P}_\theta\Big) \quad \text{subject to} \quad \sum_{\theta \in \Theta} w_\theta = 1, \quad w_\theta \geq 0,$$

where d is a chosen distance measure (e.g., KL divergence or the mean absolute distance between distributions). Once solved, these weights can be used to scale the appropriate mixture of prompts (i.e., $\boldsymbol{\theta}^\star$) and applied to new settings.

Construction Method. Alternatively, a prompt template can be parameterized by numeric trait variables. This is best illustrated with an example. Suppose ϕ_1 and ϕ_2 capture degrees of self-interest and inequity aversion, respectively. The prompt could be:

> $\theta(\phi_1, \phi_2)$ = *"You weigh your own payoff with weight* $\{\phi_1\}$*, and you dislike creating disadvantageous inequality at level* $\{\phi_2\}$*. Please respond accordingly."*

$\hat{P}_\theta$ denotes the distribution induced by the LLM under parameter vector $\theta = (\phi_1, \phi_2)$. Given an observed human distribution P, the optimal parameters $\boldsymbol{\theta}^\star$ are found by solving $\min_\theta \; d\big(P, \hat{P}_\theta\big)$. In practice, this can be solved using any derivative-free optimization algorithm, such as Bayesian optimization or evolutionary algorithms. Note that one need not be limited to a single prompt; optimization can be solved using multiple templates or sets of agents with different values for the same template.

Measuring Performance. Let $\hat{P}_{\boldsymbol{\theta}^\star}$ denote the LLM's distribution of responses under $\boldsymbol{\theta}^\star$, and $\hat{P}_0$ the baseline LLM off-the-shelf without any additional prompting. Training loss is $d(P, \hat{P}_{\boldsymbol{\theta}^\star})$ and validation loss is defined analogously on the held-out test setting.

To assess whether $\boldsymbol{\theta}^\star$ generalizes, we compare predictive fit against the baseline $\hat{P}_0$. This is an appealing reference measure because many distance metrics between distributions (like the KL divergence) are not easily interpretable. Furthermore, improvement over the baseline provides direct evidence that designing and optimizing prompts is a worthwhile endeavor in the first place.[12]

Relative improvement can be computed as a difference-in-distances: $d\big(P, \hat{P}_0\big) \; - \; d\big(P, \hat{P}_{\boldsymbol{\theta}^\star}\big)$. A positive difference indicates that the prompts provide better predictive power than the baseline. When we optimize over a set of candidate prompts, we seek those that yield this positive difference on both the training and testing data—that would suggest that the set generalizes.

The same framework applies across multiple training and evaluation settings, with optimization performed by averaging distances if needed. It also allows direct comparison of optimized prompts to alternative prompt sets or benchmark models (e.g., game-theoretic equilibria) using the same measurement tools.

3 Predicting behavior in novel strategic games

Thus far, we have argued that theory-grounded prompts, validated on distinct but related datasets, offer an approach for predicting human responses in entirely novel settings. To empirically test the efficacy of our approach, we now turn to predicting human behavior in a set of strategic games adapted from Arad and Rubinstein (2012)'s (AR) study of strategic reasoning.

[12] Strong baseline performance would simply reflect the LLM's already high off-the-shelf predictive capacity.

We begin by briefly reviewing the level-k model of strategic reasoning that originally motivated AR. This model provides plausible theoretical underpinnings linking the training, testing, and novel games in this section. We then describe the structure of the original 11-20 money request game and outline our procedure for constructing theory-grounded AI agents: optimizing their parameters on human data from AR's original experiment, and validating their predictive performance on distinct but related variants also studied by AR. Finally, we introduce an entirely new set of games adapted from AR's original design but featuring distinct numeric ranges and a novel participant sample recruited from Prolific. Agents are evaluated on their ability to predict human behavior in these never-before-seen games.

Although this section focuses on strategic reasoning games, the same procedure can be applied in any setting. In Appendix A, we replicate the process using the allocation games from Charness and Rabin (2002), which explore social preferences. Two other important differences in that section are that we: (i) optimize across multiple related settings rather than a single training setting, and (ii) utilize the construction method to build agents. We also present an additional preregistered experiment with human subjects to validate these agents in new settings.

3.1 Arad and Rubinstein (2012)'s 11-20 money request game

The level-k model posits that players differ in how many steps ahead they consider when forming their strategies (Nagel, 1995; Stahl and Wilson, 1994, 1995). The model defines different types of players, from level-0 to level-k. Level-0 players use some predefined arbitrary decision rule, while level-k players ($k \geq 1$) best respond assuming others are level-$(k-1)$ reasoners. Such a model highlights the idea that players' behavior depends not only on their decisions but also on their beliefs about how other people think.

To measure the distribution of level-k thinkers in human populations, AR developed the 11-20 game. The instructions are:

> *You and another player are playing a game in which each player requests an amount of money. The amount must be (an integer) between 11 and 20 shekels. Each player will receive the amount he requests. A player will receive an additional amount of 20 shekels if he asks for exactly one shekel less than the other player. What amount of money would you request?*

This *basic* version of the game clearly maps players' levels of reasoning to their choices. A natural starting point is to choose 20 shekels. This maximizes the guaranteed payment and provides an obvious starting point for level-0 thinking. Next, level-1 players, anticipating level-0 players, will choose 20, best respond by requesting 19 shekels to earn the bonus. Level-2 players, expecting level-1 behavior, choose 18 shekels, and this pattern continues down to the minimum of 11. More generally, a player choosing $(20-k)$ shekels plausibly reveals themselves as a level-k thinker.

Interestingly, the 11-20 game does not have a pure strategy Nash equilibrium. The best response to any choice greater than 11 is to undercut the opponent by 1. But if the other player chooses 11, also selecting 11 is strictly dominated by every other strategy. The top row of Table 1 shows

the unique symmetric mixed strategy Nash equilibrium for the game, with most of its density lying between 15 and 17 (levels 3 to 5).

Table 1: Original results from Arad and Rubinstein (2012)

Shekels Requested	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Level-k	***L*9**	***L*8**	***L*7**	***L*6**	***L*5**	***L*4**	***L*3**	***L*2**	***L*1**	***L*0**
Nash Eq. Prediction (%)	*0*	*0*	*0*	*0*	*25*	*25*	*20*	*15*	*10*	*5*
Basic ($n = 108$) (%)	4	0	3	6	1	6	32	30	12	6
Cycle ($n = 72$) (%)	1	1	0	1	0	4	10	22	47	13
Nash Eq. Prediction (%)	*0*	*0*	*0*	*10*	*15*	*15*	*15*	*15*	*15*	*15*
Costless ($n = 53$) (%)	0	4	0	4	4	4	9	21	40	15

Notes: This table reports the empirical PMFs for three versions of the 11-20 game from Arad and Rubinstein. In the basic version of this game, two players each request a number between 11-20 shekels and they receive that amount. If a player requests exactly one less than their opponent, they win their request plus a 20 shekel bonus. In the cycle version, players also receive a 20 shekel bonus if they select 20 and their opponent selects 11. The costless version is identical to the basic version, except that players receive 17 shekels if they select any amount other than 20. The basic and cycle versions of the game share a unique symmetric mixed strategy Nash equilibrium, which is shown in the first row. The unique mixed strategy Nash equilibrium for the costless version is shown in the 4th row.

Table 1 also shows that when AR tested this game on pairs of college students, they deviated significantly from the Nash equilibrium (see row titled Basic). Most notably, 73% of participants chose between 17 and 19 shekels, whereas only 45% of the density for the mixed strategy Nash is on these values.

Finally, Table 1 provides results from two additional variants of the game from AR. Both variants of the game still involve two players selecting numbers between 11 and 20. They differ from the basic version in their payouts (see Appendix B for the full instructions). In the *cycle* version, players can earn a bonus of 20 shekels by undercutting the other's request by exactly one or by selecting 20 when the other selects 11. This version has the same unique mixed strategy Nash Equilibrium as the basic game because the only affected strategy profiles are (20,11) and (11,20), which are outside the equilibrium support. However, if there is a distribution of level-k thinkers—as opposed to perfectly rational game-theoretic best-responders—such a change might significantly impact play.

The *costless* version has the same bonus structure as the basic game, but with a different payout for choices below 20. Here, requesting 20 yields 20 shekels outright, while choosing a lower amount guarantees 17 shekels plus a 20-shekel bonus if the lower request is exactly one less than the other player's. It is comparatively "costless" to continue undercutting. This game induces a symmetric mixed strategy Nash equilibrium that is more uniform across the choice set than the basic and cycle versions.

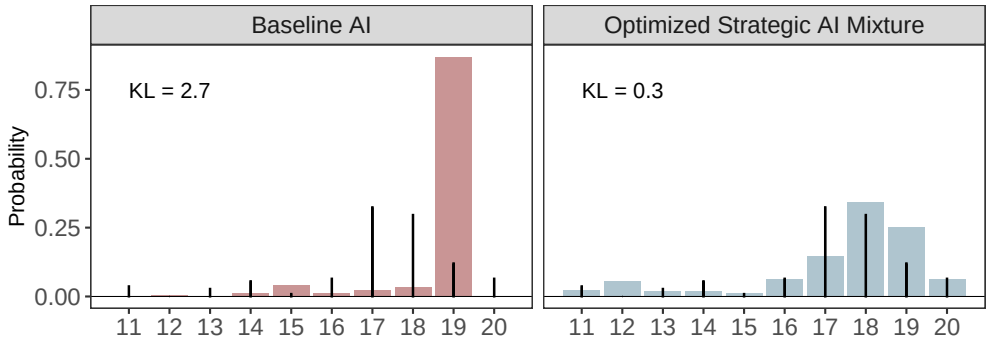
The human responses from these game variants, also from a similar sample of college students, are noticeably shifted towards 18-20 shekels—even with the basic and cycle sharing the same equilibrium. AR attributes this to the increased salience and payoff of selecting a higher number. AR concluded that the collective results from these three experiments are best explained by a mix of strategic types consisting of level-0, level-1, level-2, level-3, and random choosers.

The empirical human distributions from these three games comprise our training (the basic version) and validation (the costless and cycle versions) data. The games are distinct enough such that the human response distributions are different, but still all likely well-explained by similar underlying human choice processes.

3.2 Optimizing AI agents in-sample

We begin by eliciting the baseline AI's response distribution ($\hat{P}_0$). We prompt GPT-4O to play the basic version of the game 1,000 times without any additional instructions, setting the temperature to 1. Figure 1 displays the results. The left panel shows the empirical PMF of the baseline AI responses (red), along with the empirical human distribution P from AR's original experiment (vertical black lines). The baseline AI almost exclusively selects 19 shekels (87%), demonstrating limited variability. Using the forward KL divergence as $d(\cdot, \cdot)$ with the human distribution as the reference, the distance between these distributions is $d(P, \hat{P}_0) = 2.7$.

Figure 1: Response distributions for the basic version of the 11-20 game



Notes: This figure displays empirical PMFs for three samples playing the basic 11-20 money request game: human subjects from Arad and Rubinstein (the vertical black lines in both panels), the off-the-shelf baseline (left panel), and responses from our selected AI agents based on the weights (right panel). The KL divergence between the human and AI distributions is reported in the upper left corner of each panel.

Such a poor baseline result is unsurprising. In a related exercise, Gao et al. (2024) also elicit responses from various LLMs playing the same 11-20 game. Even after applying diverse prompting strategies, fine-tuning, and endowing agents with distributions of demographic traits, they similarly find that LLMs strongly index on choosing 19 shekels. While this outcome is not inherently problematic—19 shekels lies within the support of both the Nash equilibrium and the empirical human distribution—it underscores a limitation: demographic prompts (and the other techniques they employ) alone provide little leverage in predicting strategic human reasoning. They do not provide any reliable, flexible program for the LLM to follow, nor do they allow for heterogeneity within the simulated sample. In contrast, the level-k model implies that optimal predictions require explicitly accounting for how individuals reason about others' decisions. And because we know from AR that there is likely a distribution of reasoning levels in their human sample, we should apply various levels of reasoning to construct a heterogeneous sample of agents. This may, in turn, better

reflect the distribution of human strategic reasoning processes.

To operationalize the level-k reasoning explicitly, we construct a set of natural language prompts ($\Theta_{Strategic}$) corresponding to varying levels of strategic reasoning. These candidate prompts, listed in Table 2, specify how far ahead each AI agent reasons about the opponent's decisions, effectively encoding beliefs about others' strategies.

We elicit response distributions $\hat{P}_\theta$ for each candidate prompt $\theta \in \Theta_{Strategic}$ by prompting GPT-4o 100 times per prompt.[13] We then employ the selection method to identify the optimal mixture of prompts that minimizes the absolute difference between the CDFs implied by the empirical distribution of the human responses (P). This distance can be minimized using simple nonlinear programming techniques. The weights $\mathbf{w}^*$ corresponding to the optimal mixture appear in the second column of Table 2.[14]

Most mass concentrates on two prompts: one that reasons between levels 1 and 3 (47%), and another varying more broadly from levels 0 to 5 (34%). The remaining weight falls on more extreme behaviors (random choices or the safest guaranteed option). This aligns closely with AR's findings, whose human subjects predominantly exhibited level-0 through level-3 reasoning or made random choices.

Table 2: Proposed AI agent prompts and resulting mixture weights from the selection method

Persona	Weight
You are generally a 0-level thinker—picking the option with the most guaranteed money.	0.065
You vary between a 0 and 1-level thinker.	0.000
You vary between a 1 and 2-level thinker.	0.000
You vary between a 0, 1, and 2-level thinker.	0.000
You vary between a 0, 1, 2, and 3-level thinker.	0.000
You vary between a 1, 2, and 3-level thinker.	0.469
You vary between a 0, 1, 2, 3, and 4-level thinker.	0.013
You vary between a 0, 1, 2, 3, 4 and 5-level thinker.	0.339
You randomly pick between lower numbers because you think that's the best way to win.	0.114
You are Homo Economicus.	0.000

Notes: This table shows the set of prompts $\Theta_{Strategic}$ used as input to the selection method. The right column shows the optimized mixture weights $\mathbf{w}^*$ that minimize the absolute difference between the CDFs of the human distribution P_s and the distribution of responses from the AI agents. Prepended to all the prompts is: *You are a human being with all the cognitive biases and heuristics that come with it.* We also include an explanation in the prompts for k-level reasoning for all prompts besides the random one: *A k-level thinker thinks k steps ahead. A 0-level thinker thinks 0 steps and would, therefore, just select the maximum amount that guarantees money.*

Using these weights, we generate the sample $\boldsymbol{\theta}^*$ of 1,000 AI agents by assigning each agent to one of the 10 prompts with probability equal to its corresponding weight in Table 2. The resulting empirical distribution of responses $\hat{P}_{\boldsymbol{\theta}^*}$ produced by this sample of 1,000 agents $\boldsymbol{\theta}^*$ appears in the right panel of Figure 1 (blue). The improvement over the baseline AI is substantial: $d(P, \hat{P}_{\boldsymbol{\theta}^*}) = 0.3$

[13] We then elicit each AI agent's responses using Chain-of-Thought prompting, which encourages step-by-step reasoning before producing a final answer (Wei et al., 2024). We implement this through two sequential prompts. **Prompt 1**: {*11-20 game instructions*}. *Reason out a few settings according to your personality and how others might respond.* **Prompt 2**: {*11-20 game instructions*}. *You previously had the following thoughts:* {*Response to prompt 1*}. *What amount of money would you request?*. This procedure creates the agents (and their response distributions) over which we then optimize.

[14] See https://www.expectedparrot.com/content/6f58d11f-98cc-4de5-bb89-edcf78042d79 for the agents.

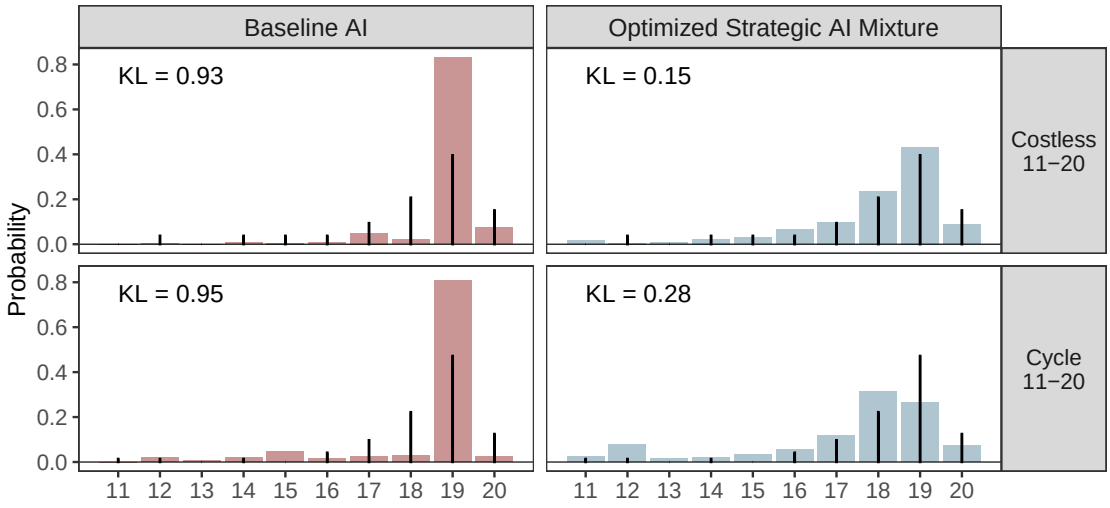
is 89% smaller than $d(P, \hat{P}_0) = 2.7$, demonstrating a strong in-sample fit.

3.3 Validation using game variants

To validate these agent sets, we elicit their response distributions to the costless and cycle versions of the game. Both games maintain our fundamental requirement of similar but distinct data-generating processes. They involve reasoning well-explained by level-k thinking, but feature payoff structures and incentives that differ from those of the basic game. Most important, humans do not play these games in the same way as the basic version. The forward KL divergence between the basic and costless game is 1.26, and between the basic and cycle game is 1.00 (see Table 1 for the human response distributions for all three games). These divergences are substantial: each is more than three times larger than the divergence between the human and optimized agent distributions in the right panel of Figure 1. Accurate predictions on these distinct games would indicate that our optimized AI agents capture generalizable patterns rather than merely replicating the original training scenario.

We elicit responses from all 1,000 AI agents in our optimized sample $\boldsymbol{\theta}^*$ for both the costless and cycle versions of the game. As a benchmark for evaluating relative performance, we also elicit responses from the baseline AI 1,000 times per game. Figure 2 presents these results, comparing the empirical distributions from AR's original human experiments (black lines) with those generated by the baseline AI (red) and the optimized mixture of theory-grounded AI agents (blue).

Figure 2: Response distributions for the cycle and costless versions of the 11-20 game



Notes: This figure displays empirical PMFs for the costless (top row) and cycle (bottom row) variants of the 11-20 game. The columns correspond to the baseline (red), the optimized AI agents (blue). Within each panel, the empirical PMFs from Arad and Rubinstein are imposed in black. The KL divergence between each human and the AI response distribution is displayed in each panel. For both variants of the game, the selected AI agents (blue) are far closer to the human distribution than the baseline (red), even though the selected AI agents are constructed using only the basic version of the game.

Consistent with the basic version of the game and Gao et al. (2024), the baseline AI overwhelm-

ingly selects 19 shekels in both variants, a considerable divergence from AR's human subjects. In contrast, $\boldsymbol{\theta}^*$ is a far better predictor of both validation settings. The costless game in particular shows substantial improvement, with the KL divergence decreasing by 84% ($d(P, \hat{P}_{\boldsymbol{\theta}^*}) = 0.15$ vs. baseline $d(P, \hat{P}_0) = 0.93$). It is almost perfectly predictive of the human responses when comparing the empirical PMFs. In the cycle game, the KL divergence between the optimized AI and human responses is also reduced substantially, by 71% relative to the baseline ($d(P, \hat{P}_{\boldsymbol{\theta}^*}) = 0.28$ vs. $d(P, \hat{P}_0) = 0.95$). Taken together, these results are interpreted as evidence that $\boldsymbol{\theta}^*$ has effectively generalized to the validation data—data that was not used to construct its mixture. We therefore gain confidence that these agents may better predict entirely new settings that call for similar strategic reasoning.

3.4 Optimizing among atheoretical prompts

We now generate sets of arbitrary, atheoretical AI agents using the basic version of the game, which ultimately fail validation on the cycle and costless versions. These agents will offer a substantial contrast with $\boldsymbol{\theta}^*$ when applied to entirely novel games in the next subsection. This exercise also highlights two potential pitfalls of AI simulations addressed by our approach: (i) without theoretically motivated candidate prompts, optimization may fail to even improve predictive power in-sample over the baseline, (ii) atheoretical candidate sets can be optimized to effectively match particular samples of human data—even when such samples are obviously overfitting. The former is addressed by using samples of AI agents grounded in plausible theory, and the latter is addressed by using training and testing data from distinct settings (or training and testing both across many different settings).

To illustrate these points concretely, we introduce three new sets of candidate prompts, none having any plausible relationship to strategic reasoning or the choices made in the different 11-20 games. These are shown in Table D1 in the appendix. The first set consists of historical figures (Θ_{Hist});[15] the second has the 16 Myers-Briggs personality types (Θ_{MB}); and the third set comprises 10 "Always Pick 'N"' agents (Θ_N), each of which is instructed to exclusively select a given integer from 11 to 20.[16] We apply the exact same selection procedure used in Section 3.2 to find optimized weights for each set, using only the human data from the basic version of the 11-20 game.

Table D1 also shows the resulting weights. For the historical figures, nearly all weight (89.1%) collapses onto Julius Caesar and a small remainder on Confucius (10.1%). In the Myers-Briggs set, all weight concentrates on ENFP.[17] While Julius Caesar is historically renowned for his strategic military prowess, it is unclear how a generic reference to his name translates into a meaningful prompt for this game. Likewise, Myers-Briggs constructs are widely considered to be pseudo-scientific and meaningless.

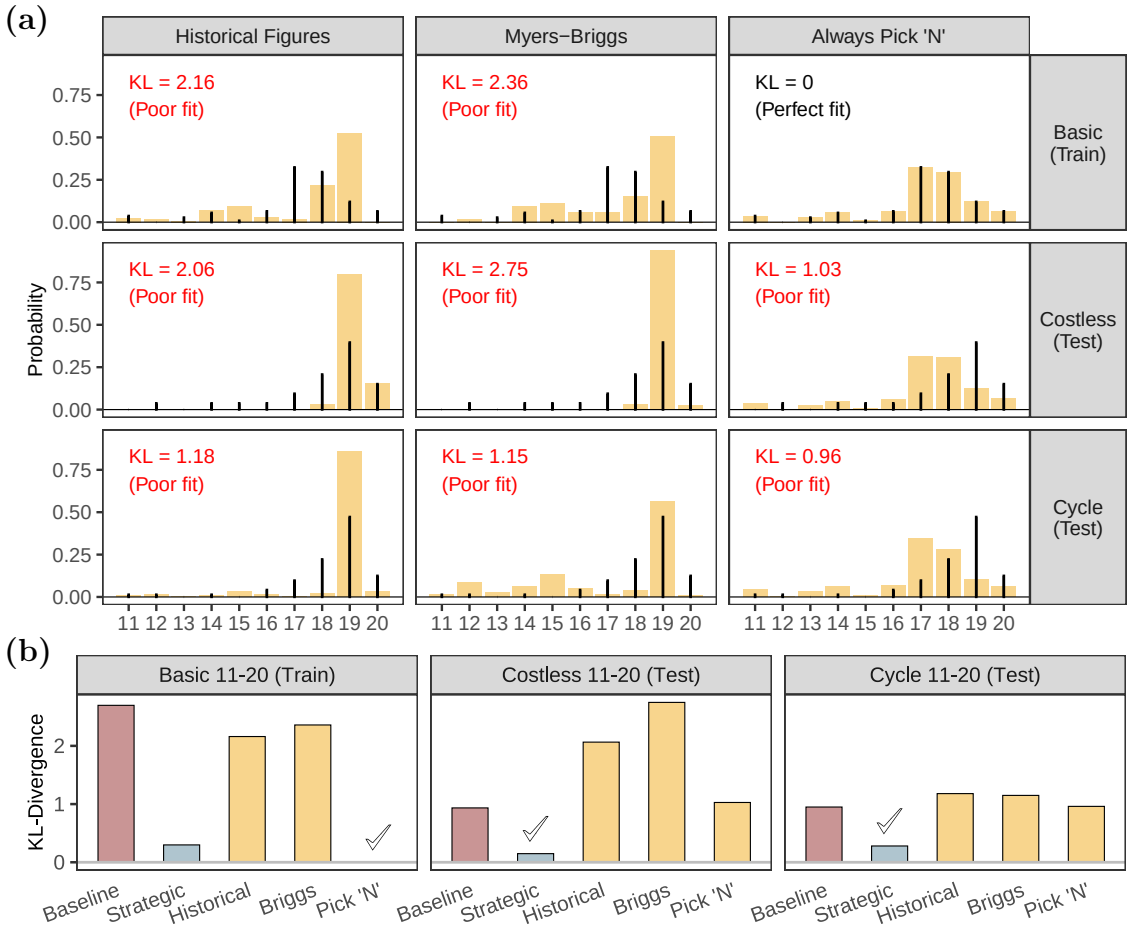
[15] Cleopatra, Julius Caesar, Confucius, Joan of Arc, Nelson Mandela, Mahatma Gandhi, Harriet Tubman, Leonardo da Vinci, Albert Einstein, Marie Curie, Genghis Khan, Mother Teresa, Martin Luther King, Frida Kahlo, George Washington, Winston Churchill, Mansa Musa, Sacagawea, Emmeline Pankhurst, and Socrates.

[16] These agents all take the form of *"You always pick N"* for $N \in \{11, \ldots, 20\}$.

[17] ENFP is Extraversion, Intuition, Feeling, Perceiving (wikipedia.org/wiki/Myers-Briggs_Type_Indicator).

Figure 3a shows that these two selected samples do not even offer a good in-sample fit. Each row corresponds to a different variant of the game—the top row is the basic. The columns represent different AI agent types, with the empirical PMFs from AR superimposed in black. The KL divergence between the distributions in each panel is shown in the top left of each panel. After optimization, the in-sample KL divergence between the selected AI agents and the humans in AR is $d(P, \hat{P}_{\boldsymbol{\theta}^*_{Hist}}) = 2.16$ and $d(P, \hat{P}_{\boldsymbol{\theta}^*_{MB}}) = 2.36$ for the historical figures and Myers-Briggs, respectively. These are not much better than the baseline $d(P, \hat{P}_0) = 2.7$ and far worse than the strategic AI agents $d(P, \hat{P}_{\boldsymbol{\theta}^*}) = 0.3$.

Figure 3: Response distributions for the 11-20 games with atheoretical AI agents



Notes: (a) shows empirical PMFs for three variants of the 11-20 game from Arad and Rubinstein, compared to selected atheoretical AI agent samples optimized using only the basic version. Rows correspond to game variants, and columns correspond to AI agent types, with human data superimposed in black. Historical figures and Myers-Briggs subjects poorly match human distributions across all variants. The "Always Pick 'N"' set matches human data perfectly in-sample but fails to generalize. (b) shows KL divergence between human and AI responses for games from Arad and Rubinstein. The lowest KL divergence in each panel is indicated by a checkmark. Only the strategically-selected AI agents consistently improve over the baseline in all games.

The core issue is that these "theories" are bad: historical personas and pseudo-scientific personality types are not causally related to how humans play these games, which means the prompts cannot generate meaningful variation in the dependent variable. The hypothesis classes are not

flexible enough, and the optimization produces a predictive model that severely underfits. To make an analogy, if x covaries with y, then $y = mx + b$ may effectively fit a range of (x, y) pairs, but $y = b$ cannot.

When validated out-of-sample on the costless and cycle variants (Figure 3a, bottom rows), these atheoretical personas perform even worse relative to the baseline. Both historical figures and Myers-Briggs types simply default to selecting 19 shekels, severely diverging from the shifted human response distributions.

The third atheoretical set ('Always Pick 'N"') initially appears successful, achieving a perfect in-sample fit ($d(P, \hat{P}_{\boldsymbol{\theta}^*_N}) = 0$). However, this is misleading—such personas offer no flexibility. Each agent always selects its assigned integer, between 11 and 20, regardless of setting changes, clearly overfitting to the training data. Unsurprisingly, when validated on the human data from the new variants, these largely fail to improve over the baseline, merely reproducing their training distribution and failing to capture shifts in human responses (rightmost column of Figure 3a).

Figure 3b succinctly compares these results along with the optimized mixture of strategic AI agents and the baseline, marking the best-performing sample in each setting. Only $\boldsymbol{\theta}^*$ consistently outperforms the baseline and generalizes across all settings. All three atheoretical samples are strictly worse than the baseline on both validation games.

These results underscore the importance of theory-driven candidate prompts and validation across related but distinct settings. We next show that failure to pass validation bodes poorly for predicting responses in new settings.

3.5 Predicting the new games

We now introduce the four novel strategic games. To the best of our knowledge, these games are not in the LLM's training corpus, thus providing a stringent testing ground for the AI agents we have explored so far. Three of the games parallel AR's games in strategic structure but modify the implementation: participants choose between 1 and 10 (rather than 11 and 20) and earn points instead of shekels. The instructions for the "basic" version of this 1-10 game highlight these differences:

> *You are going to play a game where you must select a whole number between 1 and 10. You will receive a number of points equivalent to that number. After you tell us your number, we will randomly pair you with another player who is also playing this game. They will also have chosen a number between 1 and 10. If either of you selects a number exactly one less than the other player's number, the player with the lower number will receive an additional 10 points.*

We adapted the costless and cycle variants similarly. The fourth "1-7 game" introduces an entirely new variant with a restricted choice set (see Appendix B for the full instructions for all games).

Of the 1,000 participants we recruited from Prolific, 955 passed the validation check and were randomly distributed across the four games. To ensure incentive compatibility, participants were paid $1 for completion and had a 10% chance of receiving the dollar value of their earned points

from the single game they played. We preregistered our complete experimental design, including all prompts for both the baseline and selected AI agents—the latter using only the weights optimized on the basic 11-20 game. We have $\boldsymbol{\theta}^*$, $\boldsymbol{\theta}^*_{Hist}$, $\boldsymbol{\theta}^*_{MB}$, $\boldsymbol{\theta}^*_{N}$, and the baseline play each game. All AI agent responses are elicited using GPT-4o with the temperature set to 1. All AI agent samples played these games *before* the human subjects' data was collected.

Figure 4a presents the response distributions of initial play for all subject samples—both human (top two rows) and AI (remaining rows)—across the four novel games. The responses from AR are all shifted down by 10 for ease of comparison—20 is now 10, 19 is 9, etc. Response distributions from the Prolific sample (thin black bars) are superimposed on all other samples.

The Prolific responses differ notably from AR's original results. Specifically, in the basic 1-10 game, the Prolific sample's distribution is more uniform, with a modal choice at 8 rather than AR's mode at 7; in the costless variant, Prolific responses peak at 10 instead of AR's 9 and 8; and the cycle variant yields a more uniform distribution relative to AR's original sample. The newly introduced 1-7 game lacks an analogous comparison in AR, but the modal response is 5, and most other choices are on 4, 6, and 7.

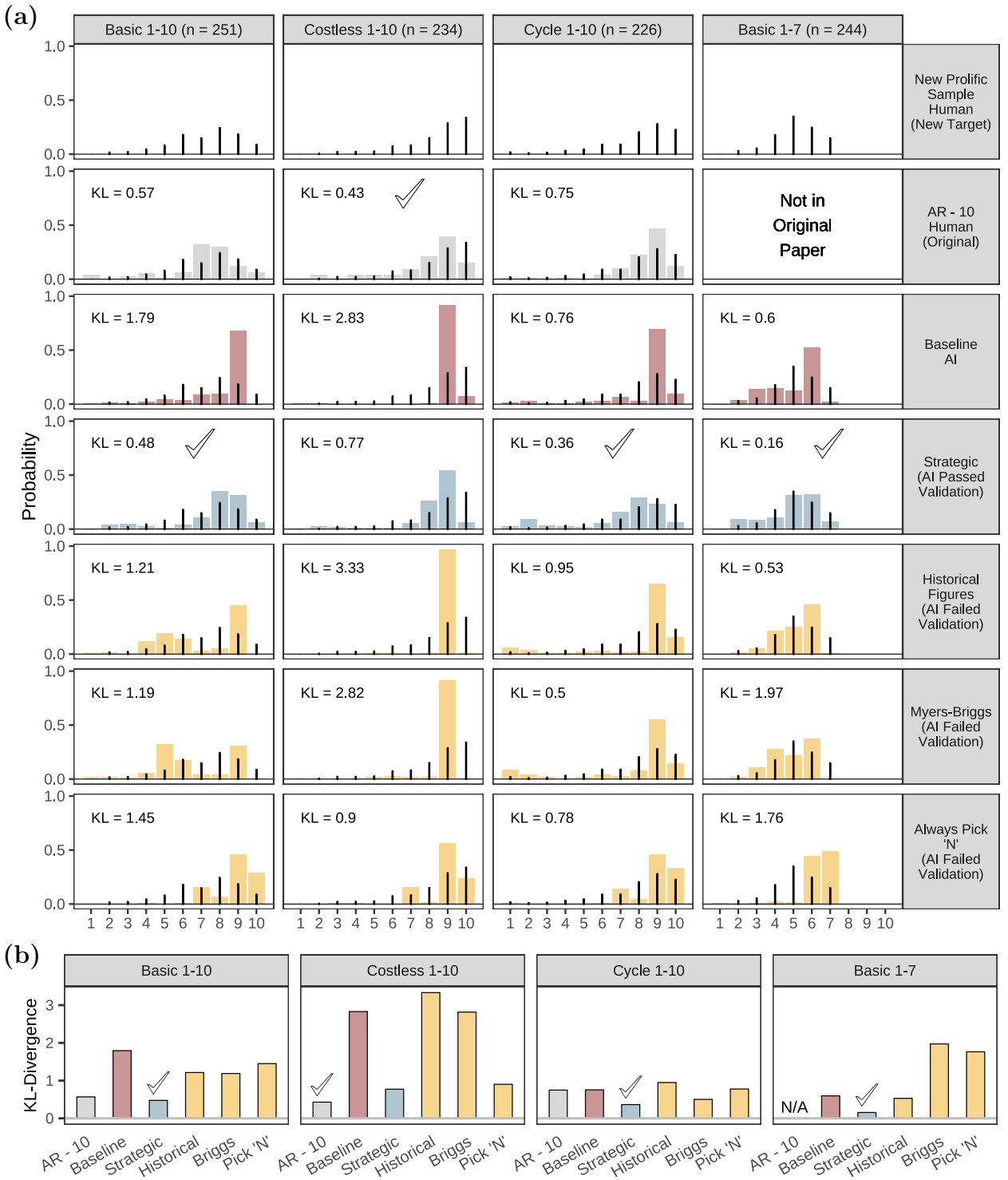
Assuming the difference is not due to sampling variation, the gap between the Prolific data and Arad and Rubinstein is somewhat surprising. Both sets of games share the same fundamental strategic structure. Although the games differ in their Nash equilibria (which we know is not how actual humans play anyway), a fixed set of k-level reasoners would play them identically: level-0 would naturally choose the highest number (10), level-1 the second-highest (9), and so on. The divergence may therefore reflect setting-specific factors or differences in beliefs about others' reasoning—precisely the kinds of factors that otherwise make theory difficult to apply. In the end, even highly relevant human data proves to be an imperfect predictor.

Moving down Figure 4a, the remaining rows show response distributions from various AI samples. Whether or not the AI sample successfully passed validation on the costless and cycle games is indicated below the sample name. The baseline AI (red) generally provides a poor fit to the Prolific data, frequently concentrating choices on 9. The notable exception is the 1-7 game, where it disperses responses across several choices.

Critically, the optimized sample of strategically-motivated AI agents robustly generalizes to these novel settings. This is the only sample of AI agents that was successfully validated on all of the data from AR's original games (the atheoretical samples each failed to improve in at least one of these games). Figure 4b reports the KL divergence between the Prolific data and each AI-generated distribution: the strategic agents consistently outperform the baseline AI by at least 53% in every variant (Appendix Figure D2 shows the same comparisons for several alternative distance metrics). Especially notable is the strategic agents' near-perfect alignment in the entirely novel 1-7 game (KL divergence = 0.16). Indeed, the strategically motivated AI even outperforms AR's original human data in predicting our Prolific participants' responses in the basic and cycle games.

In stark contrast, atheoretical AI agents fail to generalize. All arbitrary samples predict the human responses no better than the baseline in at least two of the four games. The "Always Pick

Figure 4: Analysis of novel 1-10 games: response distributions and KL divergences



Notes: (a) plots human and AI response distributions for the four games. Prolific data are in black superimposed on all other distributions. Arad and Rubinstein data occupy the second row, shifted down by 10 to fit the 1-10 format. The remaining rows show the various AI agent samples named in the right-hand column. (b) Reports the KL divergence between Prolific data and each other distribution. The minimum in each panel is flagged by a checkmark.

'N"' agents are particularly nonsensical as they are solely instructed to select integers between 11 and 20, highlighting a severe case of overfitting to a particular data-generating process.

Overall, these results underscore our central claim: carefully identifying theoretically-grounded candidate prompts and validating their predictive utility in related but distinct contexts can substantially enhance predictive accuracy in novel, unseen settings. Only agents subject to this approach generalized to the novel strategic games.

4 External validity in pre-committed novel settings

We now turn to making guarantees about inference in novel settings. This is made possible when we have a pre-committed family of settings from which we can randomly sample. In particular, this framework allows us to compare the relative accuracy of two models in predicting human responses even for unsampled settings—where we have no previous data—within the same general domain.

The setup is related to that of Hotz et al. (2005), Allcott (2015), and Dehejia et al. (2019), where treatment effects from various "sites" are used to evaluate the external validity of a given intervention at the population level. However, they assume a common underlying intervention—analogous to a single setting in our framework—across all sites. When there are heterogeneous interventions, only special instances with strong additional assumptions allow for appropriate inference. The following, which is very similar to the theoretical framework in Andrews et al. (2025), requires no such assumptions.

Let $X = \{x_1, \ldots, x_M\}$ denote the pool of candidate settings for which we wish to make predictions. $P(y|x)$ denotes the true human response distribution for $y \in Y$—the set of allowable responses. We define a predictive distribution for some flexible model θ—an LLM, an economic model, etc.—as $\hat{P}_\theta(y|x)$. For a given setting x, the expected log-likelihood that the human distribution could have been produced by θ is

$$\ell(x;\theta) = \mathbb{E}_{y\sim P(\cdot|x)}\big[\log \hat{P}_\theta(y|x)\big].$$

Then the comparative predictive power of two models θ' and θ'' can be measured via

$$\Lambda(x) = \ell(x;\theta') - \ell(x;\theta''). \tag{1}$$

A positive $\Lambda(x)$ means that θ' assigns more probability mass to the human responses than θ'' for x. Averaging over all settings in X yields the population estimand

$$\bar{\Lambda} = \mathbb{E}_{x\sim\pi}\big[\Lambda(x)\big] \tag{2}$$

where π is some distribution over X. A positive $\bar{\Lambda}$ is interpreted as evidence that θ' is, on average, more predictive of human behavior than θ'' across the entire population of X.

Suppose we observe a sample $S = \{s_1, \ldots, s_n\} \subset X$ and, for each $s \in S$, independent human

responses $y_s = (y_{s,1}, \ldots, y_{s,m_s})$. Identification to estimate equations 1 and 2 from these samples requires the following assumptions.

Assumption 1. (Unconfounded settings). The observed settings $S = \{s_1, \ldots, s_n\}$ are randomly sampled from distribution π with full support over X such that $\pi(x) > 0$ for all $x \in X$.

Assumption 2. (Random assignment and within-setting independence). Humans are randomly assigned to settings in S. Human responses within a setting are independent draws from $P(y \mid s)$.

Assumption 3. (Positivity and finite second moment). Whenever $P(y \mid x) > 0$, then $\hat{P}_{\theta'}(y \mid x) > 0$ and $\hat{P}_{\theta''}(y \mid x) > 0$. Moreover, $\mathbb{E}_{x\sim\pi}\{\mathbb{E}_{y\sim P(\cdot|x)}[(\log \hat{P}_{\theta'}(y \mid x) - \log \hat{P}_{\theta''}(y \mid x))^2]\} < \infty$.

The first two assumptions are basically identical to the assumptions of *unconfounded location* and *random assignment* in Hotz et al.. They are also very similar in spirit to Assumption 1 in Andrews et al.. The first part of Assumption 3 is similar to the covariate overlap assumption in causal inference; without it, some observed responses would have $\log 0$. The second portion of Assumption 3 is a standard finite second-moment condition.

For every setting $s \in S$, define the sample analogue to equation 1 as:

$$\hat{\Lambda}_s \;=\; \frac{1}{m_s}\sum_{j=1}^{m_s}\Big[\log \hat{P}_{\theta'}(y_{s,j} \mid s) - \log \hat{P}_{\theta''}(y_{s,j} \mid s)\Big]. \tag{3}$$

Aggregate across settings to produce the sample analogue to equation 2:

$$\bar{\Lambda}_S \;=\; \frac{1}{n}\sum_{s\in S}\hat{\Lambda}_s. \tag{4}$$

Proposition 1 (Unbiasedness and asymptotic normality)**.** Suppose Assumptions 1–3 hold. Then

$$\mathbb{E}[\bar{\Lambda}_S] = \bar{\Lambda} \quad \text{and} \quad \sqrt{n}\,(\bar{\Lambda}_S - \bar{\Lambda}) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2\big),$$

where $\sigma^2 = \mathrm{Var}_{x\sim\pi}\big[\Lambda(x)\big] \;+\; \mathbb{E}_{x\sim\pi}\Big[\frac{1}{m_x}V_x\Big]$ and $V_x = \mathrm{Var}_{y\sim P(\cdot|x)}\big[\log \hat{P}_{\theta'}(y \mid x) - \log \hat{P}_{\theta''}(y \mid x)\big]$.[18]

Notably, this does not rely on a large sample size of humans for any particular setting. The asymptotic variance in Proposition 1 decomposes into two conceptually distinct parts. The first term, $\mathrm{Var}_{x\sim\pi}\big[\Lambda(x)\big]$, reflects heterogeneity in model performance across settings. The second term, $\mathbb{E}_{x\sim\pi}\big[\frac{1}{m_x}V_x\big]$, is sampling noise that arises because we estimate $\Lambda(x)$ with a finite number m_x of human draws. As long as $m_x \geq 1$, this component is finite. Consequently, once a few human

[18] *Proof.* (Unbiasedness). For any setting s, Assumption 2 implies $\mathbb{E}\big[\hat{\Lambda}_s \mid s\big] = \Lambda(s)$. Hence $\mathbb{E}\big[\bar{\Lambda}_S \mid S\big] = \frac{1}{n}\sum_{s\in S}\Lambda(s)$. Taking expectation over the i.i.d. draw of the settings (Assumption 1) yields $\mathbb{E}[\bar{\Lambda}_S] = \bar{\Lambda}$. (Asymptotic normality). The random variables $\{\hat{\Lambda}_s\}_{s\in S}$ are i.i.d. across settings with finite variance $\mathrm{Var}(\hat{\Lambda}_s) = \mathrm{Var}_{x\sim\pi}[\Lambda(x)] + \mathbb{E}_{x\sim\pi}\big[\frac{1}{m_x}V_x\big] < \infty$, where the decomposition follows from the law of total variance and the independence of human draws within each setting. Because the moment condition in Assumption 3 guarantees $V_x < \infty$, the central-limit theorem gives $\sqrt{n}\big(\bar{\Lambda}_S - \bar{\Lambda}\big) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2\big)$. (Regularity). By Assumption 3, $\log \hat{P}_{\theta'}(y \mid x)$ and $\log \hat{P}_{\theta''}(y \mid x)$ are finite, so all moments used above exist.

observations are obtained per setting, the precision of $\bar{\Lambda}_S$ is governed primarily by n because each setting contributes only a single observation $\hat{\Lambda}_s$. This also means there is no within-cluster correlation left to adjust for, so the standard sample variance estimator ($\hat{\sigma}^2 = \frac{1}{n-1}\sum_{s\in S}(\hat{\Lambda}_s - \bar{\Lambda}_S)^2$) across settings already yields valid standard errors without needing to adjust for clustering. As such, standard z-tests or Wald confidence intervals follow immediately when estimating if $\bar{\Lambda}_S$ is significantly different from zero.

Crucially, this construction imposes *no assumption* that all settings share a single data-generating process. The settings can be an arbitrarily eclectic mixture—public goods games, dictator games, or entirely unrelated tasks. Inference remains valid even when the model's performance differs sharply across sub-domains; the variability of estimates widens (or narrows) in proportion to the observed heterogeneity.

The remainder of this section is devoted to implementing the above framework on a pre-committed family of strategic games. This set comprises 883,320 novel and unique permutations of Arad and Rubinstein's money request game. We randomly sample 1,500 of these games for human subjects and AI agents to play in a preregistered experiment. We use this data to estimate the relative capacity of different AI agents to predict human responses at scale. In particular, we return to the strategic sample of optimized level-k AI agents presented in Table 2 from Section 3. We compare these agents' ability to predict human responses across the 1,500 games to the baseline AI, a cognitive hierarchy model, and symmetric Nash equilibria. Because the games (and human subjects) are randomly sampled from the population according to a known distribution, confidence intervals over the comparisons are valid over the 883,320 games.

4.1 A pre-committed family of strategic games

The pre-committed family of games generalizes the original 11-20 money request game by parameterizing it into six independent variable components.[19] Each symmetric game preserves the core idea behind the original: two players simultaneously select a whole number between two bounds, earning guaranteed points based on their individual choice plus a potential bonus determined by both players' choices. The six parameters—lower bound, upper bound, gap to achieve the bonus, bonus size, rule to award guaranteed points, and bonus rule—are detailed in Table 3. The table's upper portion enumerates the possible values for the first five parameters, while the bottom section presents the eleven possible bonus rules.

To illustrate how these parameters translate into actual games, consider the following example. If the lower bound is 5, upper bound is 14, gap is 6, bonus size is 10, points rule is # - 2, and bonus rule is the Gap Abs. rule, participants see:

> *You are going to play a game where you must select a whole number between **5 and 14**. A player will receive a number of points equivalent to **that number minus two**. After you tell us your number, we will randomly pair you with another Prolific worker who is*

[19] Alsobay et al. (2025) similarly generate various public goods games and Zhu et al. (2025a) do so for over 2,000 2-by-2 strategic games.

Table 3: Game parameters and possible values

Parameter	**Possible Values**	**Description**
Lower Bound	The minimum number players can select	$\{1, 2, \ldots, 20\}$
Upper Bound	lower bound + $\{4, 5, \ldots, 20\}$	Max number players can select
Bonus Size	Points awarded when bonus condition is met	$\{1, 2, \ldots, 20\}$
Gap	Difference parameter used in certain bonus rules	$\{1, 2, 3, 4\}$
Points Rule	Rules to award guaranteed points (# is number participants select)	{# - 2, # - 1, #, # + 1, # + 2, costless - 2}
Bonus Rule	**Bonus awarded to player when...**	**Mutual vs. Competitive**
Gap Low	they select exactly $\{gap\}$ less than the opponent	Competitive
Gap High	they select exactly $\{gap\}$ more than the opponent	Competitive
More Than	they select a number $\{gap\}$ more than the opponent's number	Competitive
Gap Abs.	the absolute difference from the opponent's number equals $\{gap\}$	Mutual
Equal	they select the same number as the opponent	Mutual
Unequal	they select a different number than the opponent	Mutual
Sum Even	the sum of both numbers is even	Mutual
Sum Odd	the sum of both numbers is odd	Mutual
Coord. Low	both players select the lower bound	Mutual
Sum Upper	the sum of both numbers equals the upper bound	Mutual (or not achievable)
Less Upper	the sum of both numbers is less than upper bound	Mutual (or not achievable)

Notes: This table shows the possible values for the six parameters of the pre-committed family of games. The rightmost column indicates whether the bonus is mutually achievable (both players can receive it together) or competitive (only one player can receive it, if at all). The "Sum Upper" and "Less Upper" bonus rules are not achievable when the lower bound exceeds half the upper bound. The naive Cartesian product of all parameter values yields $20 \times 16 \times 4 \times 20 \times 6 \times 11 = 1{,}689{,}600$ combinations. However, seven of the eleven bonus rules do not use the *gap* parameter; for these rules, varying the gap value produces mechanically identical games. Collapsing such duplicates leaves 883,320 unique games in the final population X.

> *also playing this same game. They will also have chosen a number between 5 and 14. Both players will receive* ***an additional 10 points*** *if their requested numbers* ***differ from each other by exactly 6****. What number would you request?*

The full factorial product of this parameterization yields 1,689,600 games. However, many of these games are mechanically identical because seven of the bonus rules do not use the *gap* parameter. Accounting for these duplicates, we have 883,320 unique games in total—the pool of candidate settings X. This family includes the original 11-20 game as a special case (lower bound 11, upper bound 20, bonus 20, gap 1, points rule #, first bonus rule). Besides this and the other games in Section 3, to the best of our knowledge, all of these games are novel. They cannot be found in GPT-4O's training data—the model we use to generate AI responses.

Notably, the games exhibit dramatic variation in strategic difficulty. With bonus rules such as "Unequal," the vast majority of strategy profiles lead to the bonus. Conversely, for rules like "Sum Upper" or "Less Upper" bonuses are sometimes unattainable (when the lower bound exceeds half the upper bound). The games also vary in their incentive alignment between players. As indicated in the rightmost column of the bottom panel of Table 3, some bonus rules are mutually achievable, while others are competitive—only one player can receive the bonus. This heterogeneity creates a particularly stringent test of agents' predictive power, as successful generalization demands flexibility along multiple dimensions.

To construct S, we randomly sampled 1,500 games from X. The intended design was uniform

sampling for π across all 883,320 unique games. A very minor miscalculation in the deduplication process caused small deviations from uniformity: the seven bonus rules that do not use the *gap* parameter were each sampled with probability ≈ 0.086, while the four gap-using bonus rules were each sampled with probability ≈ 0.010.[20] For points rules, the "costless" variant was sampled with probability ≈ 0.095, and each of the remaining rules with probability ≈ 0.18. All other parameters were sampled uniformly. Consequently, the estimand in this section is technically the expected relative predictive power of the models over the 883,320 games under this slightly non-uniform distribution. However, robustness checks will later show that the relative predictive power of the models is not particularly sensitive to the points rule or bonus rule, indicating that this minor departure from uniformity has no substantive effect on our conclusions. The results are therefore likely to hold for many distributions π over X.

4.2 Eliciting AI agent responses

We generate AI responses for each of the 1,500 games in the set S using two distinct samples of AI agents. As a baseline sample, we prompt GPT-4o at temperature 1 to independently play each game 100 times, without providing any additional instructions. For the optimized strategic sample, we use the same 10 prompts from Table 2, which were optimized using human experimental data from Arad and Rubinstein. To generate this strategic sample, we proportionally scale the optimized persona weights to create a total of 100 AI agents, each of which plays each game exactly once using GPT-4o (the "strategic" sample of AI agents hereinafter).[21]

This procedure produces an empirical distribution for both the strategic level-k sample $\hat{P}_{\boldsymbol{\theta}^*}$ and the baseline AI $\hat{P}_0$ for every game $s \in S$. These samples correspond exactly to those described in Section 3 ($\hat{P}_{\boldsymbol{\theta}^*}$ and $\hat{P}_0$, respectively, in the notation of that section), differing only in the number of agents—here, each distribution is generated with 100 agents rather than 1,000. To be clear, these agents were constructed *months* before they played any of the 1,500 games. In total, the elicitation procedure produces approximately 300,000 individual AI agent responses.

4.3 Predictive benchmarks

A key limitation of our statistical framework is that comparing the predictive power of different AI simulations provides no absolute benchmark for how well these samples predict human responses in general. Thus, we require a suitable theoretical or statistical benchmark for a more comprehensive analysis. However, due to the scale and heterogeneity of our games, several appealing benchmarks are impractical.

[20] This miscalculation was only noticed after the experiment. As such, the preregistration (aspredicted #241394) states that we were sampling from a larger number of games. The only difference is that the correct number is 883,320. The random sample of 1,500 was still chosen before the experiment and is available here: www.expectedparrot.com/content/db984e24-2810-4b21-be4e-91efde378e21—the same link given in the preregistration.

[21] A very small fraction (less than 0.1%) of the AI agent responses were invalid due to stochasticity inherent to the LLM at temperature 1. Following our preregistered analysis plan, we discard these invalid responses without resampling.

Ideally, we would apply the standard level-k model from Section 3 to generate predictive distributions across these games. Unfortunately, no existing mechanical method reliably identifies which choices correspond to specific levels of reasoning across such diverse contexts. For example, the "obvious" choice for a level-0 player is not always the highest number—particularly when bonuses are large or the bonus rule involves selecting number 11. Consequently, higher-level reasoning does not follow the intuitive progression featured by the original game in Arad and Rubinstein.

Another seemingly attractive alternative would be to follow the approaches of Fudenberg and Liang (2019), Hirasawa et al. (2022), or Andrews et al. (2025), which involve training a bespoke supervised machine learning model on past game data to predict responses. However, this method is infeasible for our purposes because of two problems. First, it relies heavily on having access to a large and representative dataset, which we currently do not possess.[22] Second, as discussed in Section 2, traditional machine learning models (e.g., regression, decision trees, or generic neural networks) are not well-suited for predicting human behavior in novel settings because they cannot adapt to structural differences across settings. Even with substantial representative data, if a game had even one new parameter value (e.g., a new bonus rule or an upper bound above the highest from the training data), the model would need to be retrained from scratch.

We employ two complementary benchmarks that each address a different part of the above limitations. First, we use the cognitive hierarchy model of Camerer et al. (2004). Throughout the economics literature, it is one of the most empirically accurate predictors of human responses in one-shot strategic games. It can also produce predictions for all the structurally distinct games in S, but does require some prior information, which we discuss in the next subsection. Second, we use Nash equilibrium with Harsanyi-Selten selection, which requires no empirical calibration, but, along with most equilibrium concepts, has not been particularly successful at predicting initial human play for static games in the literature.

The challenge of finding an ideal benchmark actually highlights a unique strength of AI agents: they can generate plausibly accurate predictions in virtually any setting without extensive training data or pre-specified behavioral parameters. To our knowledge, no other well-known benchmarks are both likely to have predictive accuracy and flexible enough to be applied to such a wide range of structurally distinct games.

4.3.1 Cognitive hierarchy model as a benchmark

The cognitive hierarchy model of Camerer et al. (2004) is a well-established behavioral model for predicting human behavior in strategic games. It has a similar structure to the level-k model from Arad and Rubinstein, assuming a distribution of players with different reasoning levels from zero through k. However, there are three key differences. First, rather than assuming level-0 players choose the highest or most obvious choice, it assumes they choose uniformly at random. Second, level-k players best respond to the mixture of all lower-level players (levels 0 through $k-1$), not

[22] The combined approach of Zhu et al. (2025a)—using machine learning to flexibly estimate parameters in a well-defined behavioral model and then using that model to predict responses—is infeasible for the same reason.

just level-$(k-1)$. Specifically, each level chooses a pure strategy best response, and if multiple pure strategies are equally optimal, players at that level mix uniformly between them. For example, in the original 11-20 money request game, the cognitive hierarchy model assumes level-0 players choose uniformly at random, level-1 players best respond to this uniform distribution, while level-2 players best respond to the weighted mixture of level-0 and level-1 players, and so on. Third, the number of level-k players is assumed to follow a Poisson distribution ($f(k) = \frac{\tau^k e^{-\tau}}{k!}$).

The key limitation of this model in our setup—shared by the aforementioned supervised machine learning approaches—is that it requires a pre-specified value for τ to parameterize the distribution of different-level players. Since our goal is to make predictions in the 1,500 games where we have no prior human data, we have no compelling value to choose for τ other than that from the literature. In their meta-analytic validation across many strategic games and subject pools, Camerer et al. (2004) find that $\tau = 1.5$ is often an accurate predictor of human response distributions. As such, we adopt $\tau = 1.5$ as our ex ante parameter value. This estimate implies the modal player is level-1, with 90% of mass between levels 0 and 3.

With this as the distribution over reasoning levels, we mechanically calculate the cognitive hierarchy model's predictive distribution for each game in S. Since this is mechanical for a given game, it does not affect statistical inference for $\bar{\Lambda}_S$. Figure 5 illustrates the diversity of predictive distributions across games. Each panel corresponds to a predictive PMF for one of the 1,500 games, which are ordered top-left to bottom-right by their variance. The x-axes are scaled freely, so games with different action spaces (e.g., 5, 11, or 20 options) span the same horizontal width. The y-axes are also scaled freely, ensuring that distributions with small probabilities remain visible.

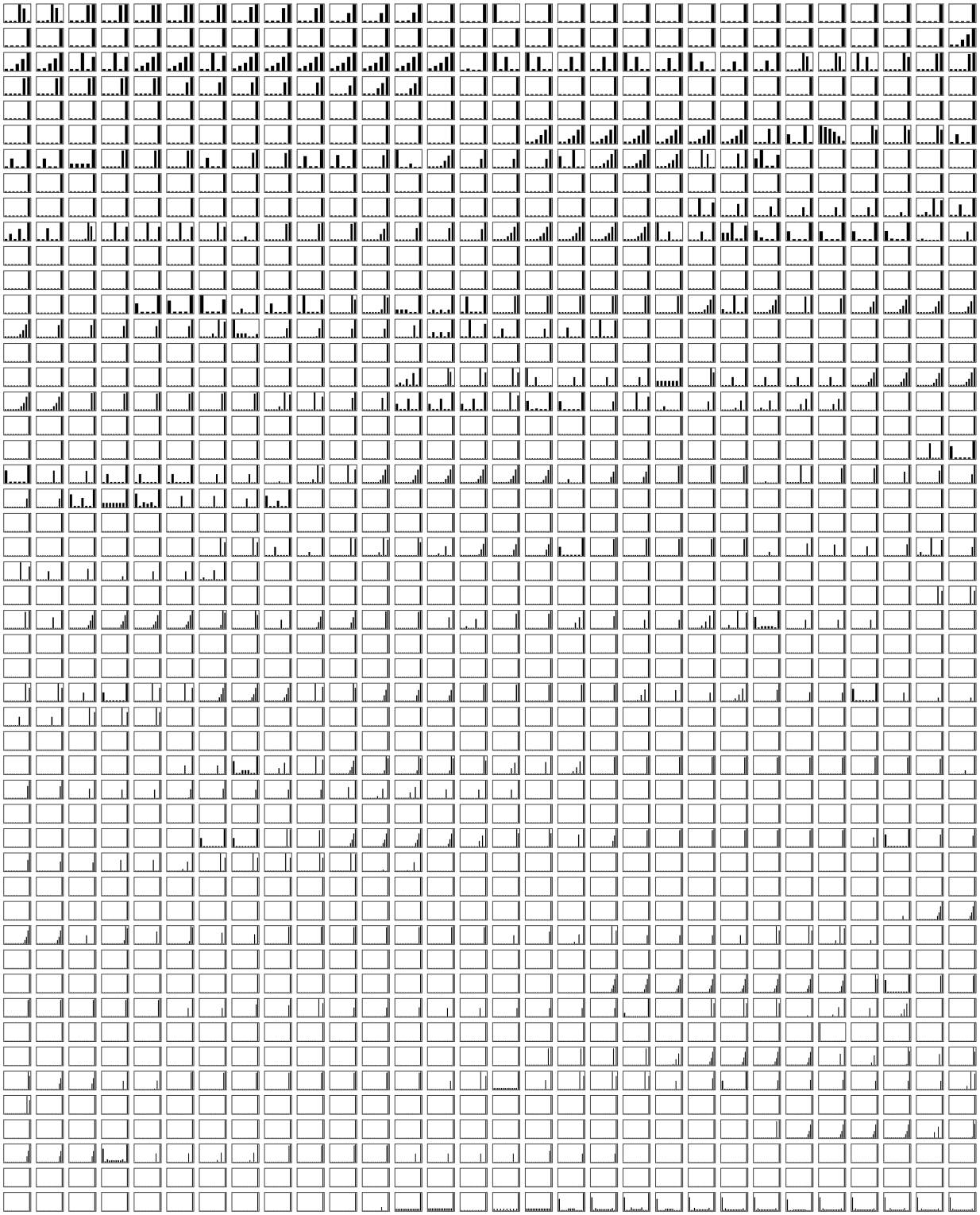
From the modest natural-language permutations in Table 3, we obtain a striking range of predictive distributions from the cognitive hierarchy model. Some spread probability across many options, while others concentrate only on the extremes—the highest and lowest actions. A few resemble uniform distributions, others exhibit sharp spikes on a small set of actions, and still others increase monotonically with the action number.

4.3.2 Harsanyi-Selten Nash equilibria as a benchmark

The second benchmark is symmetric Nash equilibria. Unlike the cognitive hierarchy model, this solution concept does not require any previous data to generate predictions. It is also flexible: symmetric Nash equilibria exist for any symmetric two-player game with a finite number of actions (Nash, 1951). They can be computed systematically across all game types in our dataset. Furthermore, in our setting, all games are played independently by participants (AI and human), making symmetry a natural assumption that suggests a "consistent common belief" across the population (Stahl and Wilson, 1994).

Since many games have multiple symmetric Nash equilibria, we require a systematic method to select a single equilibrium prediction. Indeed, one game in S has $10,051$ symmetric equilibria. Unfortunately, there is no universally agreed-upon criterion for selecting the "optimal" symmetric equilibrium across all games (Camerer, 2003; Tadelis, 2013).

Figure 5: Predictive distributions from the cognitive hierarchy model for all 1,500 games in S



Notes: This figure shows the predictive distributions of Camerer et al. (2004) Poisson cognitive hierarchy model ($\tau = 1.5$) across all 1,500 sampled games in S, with each panel corresponding to one game.

We employ a slightly modified version of the equilibrium selection procedure developed by Harsanyi and Selten (1988), which provides a principled approach grounded in stability and focal-point considerations. It guarantees the selection of a unique Nash equilibrium for every game in our dataset and can be slightly modified to ensure symmetry. The procedure also prioritizes equilibria that are stable in the sense of Schelling (1960), specifically favoring equilibria that are either payoff dominant (maximizing joint welfare) or risk dominant (robust to strategic uncertainty). This is appealing because many of the games—particularly those with bonus rules "Equal" or "Coord. Low"—often have clear equilibria that are both payoff and risk dominant, which humans often choose (Camerer, 2003).

The Harsanyi-Selten procedure operates through a multi-stage filtering process, progressively narrowing the set of candidate equilibria. The procedure first identifies all Nash equilibria, then applies filters based on Pareto efficiency, symmetry requirements, and risk dominance, and finally employs tracing methods to resolve any remaining ties. See the pseudocode in Appendix C for details on the implementation.

We first use open-source software (Savani and Turocy, 2025) to calculate all Nash equilibria for the 1,500 games in set S. We then apply the Harsanyi-Selten procedure to each game's set of equilibria, producing a single equilibrium distribution for 1,487 of the games.[23] Of these 1,487 games, 467 have unique symmetric equilibria. The selection procedure was unnecessary in these cases. Among the remaining games with multiple symmetric equilibria, the procedure selects payoff-dominant equilibria—those Pareto superior to all alternatives—in 328 games, and risk-dominant equilibria in 1,026 games. Finally, 59% of the equilibria selected by the Harsanyi-Selten procedure have pure strategies, with the remainder employing mixed strategies.

Similar to the predictions from the cognitive hierarchy model, the equilibrium distributions are highly diverse across games. This can be seen in Figure D3 in the appendix, which is of the same format as Figure 5.

4.4 Eliciting human responses

We collected human data from a sample of 4,500 Prolific workers using a custom online survey platform, which allows us to generate any game programmatically (Horton and Horton, 2024). This human data supplies y_s for each game $s \in S$, with which we can then estimate the relative predictive power of the different AI models and benchmarks. The entire experimental design, all AI agent responses, and the statistical analysis in this section were preregistered before collecting the human subjects' data.[24] Each Prolific worker was randomly assigned one of the 1,500 sampled games such that each game had approximately three human players.

The survey flow began with a very simple attention check. Participants were then shown the rules of their assigned game. This was followed by a comprehension check, which asked participants

[23] In fewer than 1% of games, degeneracy issues prevented the code from converging. Following our preregistration, we discard these games in our analysis when comparing the equilibria to the strategic AI agents.

[24] The sole exploratory analysis outside our preregistration is the comparison to the cognitive hierarchy model, added in response to suggestions we received after posting the paper online.

to calculate the correct number of points for a hypothetical outcome of their game. They were then asked to make their strategic choice. Participants received a fixed payment of \$0.50 for completing the survey. To align incentives with the game structure, 1% of participants were randomly awarded performance-based bonus payments, with each point earned in their assigned game converted to US dollars at a 1:1 rate. These bonuses were substantial, averaging \$23 across those who received them, with one participant earning \$48.

After a preregistered filtering based on the first attention check, removing participants who timed out on our platform, or those who selected a final number outside of the range of their game or did not select a whole number, our final sample size was 4,249, each playing one of the 1,490 unique games. These 1,490 games comprise the sample S we use for analysis.

4.5 Estimation

We estimate the relative predictive power of the different AI samples and the theoretical benchmark in three steps. These are: (i) construct smoothed predictive distributions for each benchmark (besides the cognitive hierarchy model) in every sampled game; (ii) evaluate the strategic sample of AI agents compared to each other model with per-game log-likelihoods and their paired differences; (iii) attach sampling-error bounds that are externally valid for the full population of nearly one million games. We address these steps in turn.

Many of the Harsanyi-Selten equilibria—mainly those pure strategies—and to a lesser extent the samples of AI agent place zero probability on strategies that humans sometimes take. For these models, the product of all likelihoods would be zero, making the log-likelihood $-\infty$ and violating Assumption 3. Dropping these games would heavily bias the results away from any pure strategy equilibria or the samples of AI agent where the agents mostly pick a single action—even when most people do select that action.

To address this, we follow the convention in game theory where players are assumed to follow the equilibrium strategy with probability $1-\varepsilon$ and choose uniformly at random the remaining ε of the time. Mathematically, this is: $\tilde{P}_\theta(y \mid s) = (1-\varepsilon)\,\hat{P}_\theta(y \mid s) + \frac{\varepsilon}{K_s}$, where K_s is the number of feasible actions in game s. Setting $\varepsilon = 0.2$ implies that players follow their model 80% of the time and choose uniformly at random the remaining 20%. We apply this smoothing to all models except the cognitive hierarchy model, which already incorporates random play and full support through its level-0 players (approximately 22% when $\tau = 1.5$)—a feature specifically designed to capture this behavior (Camerer et al., 2004). Further smoothing is therefore unnecessary for this benchmark.

This additional smoothing is not without both empirical and theoretical support (McKelvey and Palfrey, 1992, 1995). In the original 11-20 money request game, Arad and Rubinstein estimate that 32% of participants choose uniformly at random in their best fitting model. In Stahl and Wilson, at most one out of 40 participants is best explained by random choosing. Across these and other studies, the evidence suggests that between 20-30% random play captures observed behavior well. We therefore report all main results with $\varepsilon = 0.2$ but provide robustness checks with $\varepsilon \in \{0.05, 0.1, 0.3\}$ in Appendix D.

For each game s, we calculate Equation 3 five times. In all five cases, θ'—the numerator of the log-likelihood ratio $\hat{\Lambda}_s$—is the predictive distribution for the strategic sample of AI agents. The denominator θ'' is one of five benchmarks: (i) the baseline AI, (ii) the Poisson cognitive hierarchy model (with $\tau = 1.5$), (iii) the Harsanyi-Selten Nash equilibria, (iv) a uniform distribution over all possible strategies, or (v) a randomly selected pure strategy distribution. To be clear, this means all comparisons are made with respect to the strategic sample and $\hat{\Lambda}_s > 0$ implies the strategic sample is the best predictor of initial play for game s. We then take the average across the games to estimate $\bar{\Lambda}_S$ from Equation 4 for each benchmark.

Proposition 1 holds for each of these sample averages. Games were drawn via a known and fully supported distribution over the population X (Assumption 1). Human respondents were randomly assigned these games, and their answers were independent (Assumption 2). Smoothing guarantees the first part of Assumption 3, and the possible human responses form bounded, discrete distributions—so the required second-moment condition is satisfied. This means that confidence intervals must cover appropriately, and the results are externally valid for the population of all 883,320 games.

We report bootstrapped confidence intervals for the five $\bar{\Lambda}_S$ values and provide robustness checks with Wilcoxon and random-sign permutation tests. We also report the proportion of games for which the strategic AI agent is the best predictor—i.e. $\sum_{s \in S} \mathbf{1}\{\hat{\Lambda}_s > 0\}/|S|$—with its exact Clopper-Pearson 95% interval. Such intervals are valid following a nearly identical argument leading to Proposition 1.

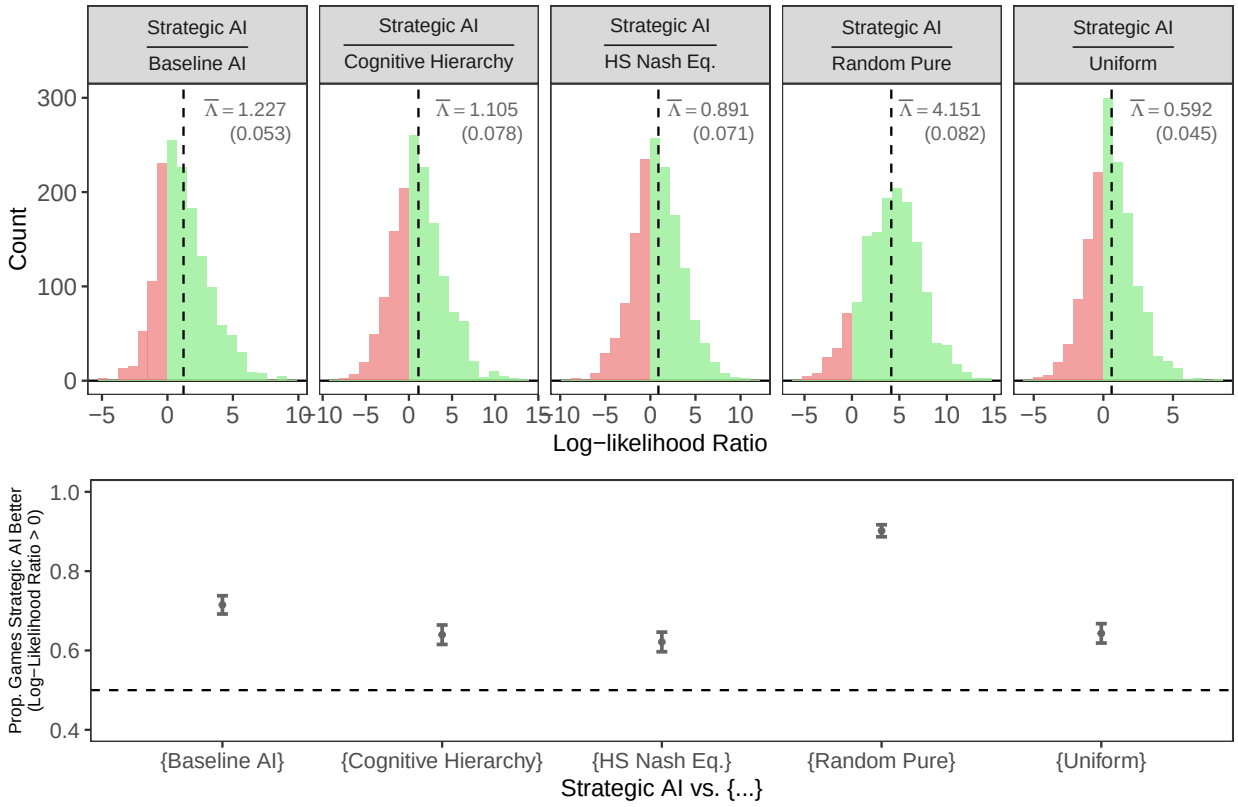
4.6 Results

The top panel of Figure 6 shows the estimation results. Each panel provides the histogram of the game-by-game log-likelihood ratios ($\hat{\Lambda}_s$) for each comparison. The vertical dashed black lines indicate the means of the log-likelihood ratios ($\bar{\Lambda}_S$), which are also given in the upper right corner of each panel. Bootstrapped standard errors are in parentheses. Green indicates ratios greater than zero, where the optimized AI has more predictive power, and red is the converse.

With $\varepsilon = 0.2$, the strategic sample of AI agent is, on average, significantly more predictive than any of the benchmarks ($p < 0.001$ for all comparisons). These differences are substantial. Starting with the leftmost panel, across all human observations in the dataset, the strategic sample of AI agents achieves an average per-observation likelihood ratio of $e^{1.23} = 3.41$ in favor of the model, relative to the baseline AI. In other words, the likelihood of the observed human data under the strategic AI agent model is, on average, 3.41 times larger per observation than under the baseline.

Moving to the right, the corresponding average likelihood ratios are $e^{1.11} = 3.02$ compared to the cognitive hierarchy model and $e^{0.89} = 2.44$ compared to the Harsanyi-Selten-selected equilibria. To be clear, this means that the strategic AI agents outperform the cognitive hierarchy model by a larger margin than they outperform the Harsanyi-Selten Nash equilibria. This is notable because throughout the literature, the cognitive hierarchy model has been a strong predictor of initial play in strategic games. It has been explicitly shown to "explain why equilibrium theory predicts behavior

Figure 6: Predictive power of strategic AI agents compared to other models ($\varepsilon = 0.2$)



Notes: The top panel shows the distribution of the log-likelihood ratios for each benchmark. The vertical black line indicates the mean, which is also provided in the upper right corner of each panel. The respective bootstrapped standard errors are in parentheses. Green indicates ratios greater than zero, where the optimized strategic AI has more predictive power, and red is the converse. The bottom panel shows the proportion of games for which the sample of strategic AI agents is the best predictor of initial play. Error bars are 95% Clopper-Pearson confidence intervals

well in some games and poorly in others" (Camerer et al., 2004). Of course, the cognitive hierarchy model might be more accurate with different values for τ, but we had no way of knowing what these were ex ante—besides those from the literature—without the data sampled from the population we wanted to predict. Finally, the strategic AI agents show large gains compared to the random pure-strategy $e^{4.15} = 63.47$ and the uniform benchmarks $e^{0.59} = 1.81$.

The bottom panel of Figure 6 shows the proportion of games for which the sample of strategic AI agents is the best predictor—i.e., has a positive log-likelihood ratio ($\sum_{s \in S} \mathbf{1}\{\hat{\Lambda}_s > 0\}/|S|$). The results are consistent with the top panel. The theory-grounded sample of strategic AI agents better predicts the human responses in more games than any other model. This proportion is large and significant for the baseline (0.715). It is smaller, although still substantial, for both the cognitive hierarchy model (0.64) and the Harsanyi-Selten Nash equilibria (0.622).

Tables in Appendix D.3 provide the same statistical analyses for $\varepsilon \in \{.05, 0.1, 0.3\}$, respectively. The above results are robust to these additional sensitivity checks. $\bar{\Lambda}_S > 0$ for all comparisons, and the proportions of games for which the optimized AI agent is the best predictor are all greater

than 50%.

Beyond relative comparisons, the strategic AI agents demonstrate impressive absolute predictive accuracy. Without any smoothing, 24% of human respondents selected the strategy for which the optimized AI agent assigns the most density. Given that the number of possible strategies per game varied evenly between 5 and 20, this is notable. Furthermore, 53% of human respondents selected one of the top three strategies for which the strategic AI agent assigns the most density. And maybe most surprisingly, for 86% of games, all human respondents selected a strategy within the support of the strategic AI agent.

4.6.1 Predictive power by bonus rule

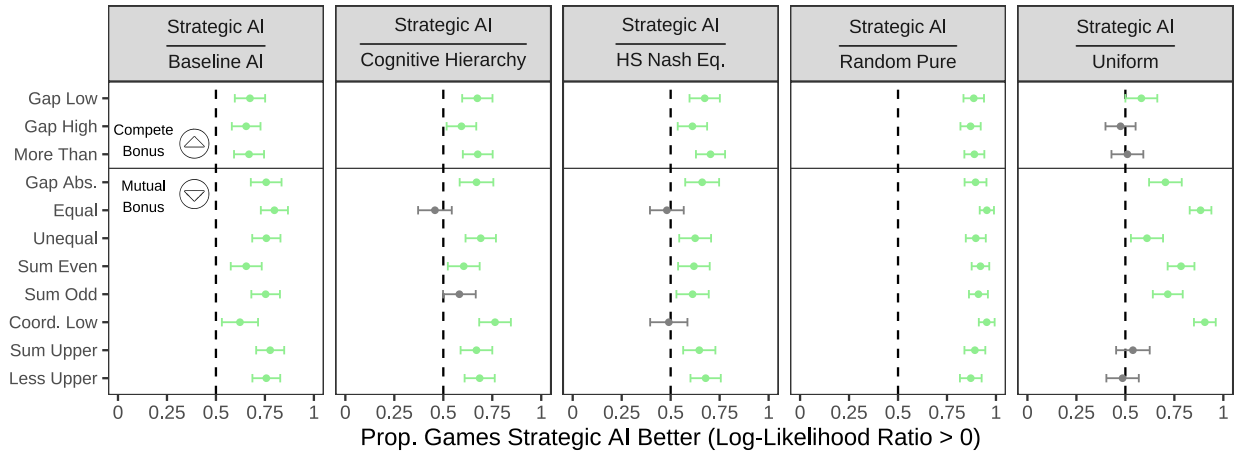
In this subsection, we briefly analyze whether the predictive power of the strategic AI agent is sensitive to the different types of games from Table 3 (and therefore different possible distributions for π). In particular, we focus on the bonus rules, although the results are similar for all dimensions of the parameter space. As a reminder, different bonus rules change the fundamental structure of the game. For some, many strategy profiles led to the bonus (e.g., "Unequal"), and others only one ("Coord. Low"). Broadly speaking, the rules can be categorized into two types: mutually achievable and competitive. In mutually achievable games, both players receive the bonus when the bonus condition is met. In competitive games, only one player can receive the bonus. This distinction roughly corresponds to the classic game-theoretic categories of coordination versus zero-sum games.

Figure 7 shows the proportion of games for which the strategic AI agent is the best predictor of initial play by bonus rule ($\sum_{s\in B} \mathbf{1}\{\hat{\Lambda}_s > 0\}/|B|$ where $B \subset S$ is the set of games with a given bonus rule). The figure is of a similar structure as the bottom panel of Figure 6, with each panel showing a different benchmark comparison with $\varepsilon = 0.2$. The y-axis shows the bonus rule, and the x-axis is the proportion of games for which the log-likelihood ratio is greater than zero. Green indicates that strategic AI agent significantly outperforms the reference model in more than 50% of games with that bonus rule, and grey indicates no significant difference. Bonus rules above the horizontal solid black line correspond to games where both players can achieve the bonus simultaneously, while those below are competitive, permitting only one player to do so.

The strategic AI agent sample weakly dominates all other benchmarks; it never significantly underperforms for any bonus rule. It outperforms the baseline and the random pure-strategy benchmark for every rule, the cognitive hierarchy model and the Harsanyi-Selten Nash equilibria for all but two rules, and the uniform benchmark for 7 of the 11 rules. There is no apparent distinction between mutually achievable and competitive bonus rules.

In Appendix D.3, we provide analogous plots for alternative values of $\varepsilon \in \{0.05, 0.1, 0.3\}$ and for the various points rules in Table 3. We also report regressions of other game parameters (lower bound, upper bound, bonus size, gap) on the log-likelihood ratio of the strategic AI agent relative to the benchmarks ($\hat{\Lambda}_s$). The results are similar to the ones presented here. The strategic AI agents are almost universally superior across different slices of the parameter space. They are therefore

Figure 7: Relative predictive power of strategic AI agents ($\varepsilon = 0.2$) by bonus rule



Notes: This figure shows the proportion of games for which the strategic AI agents is the best predictor of initial play for each bonus rule (on the y-axis). The vertical dashed line corresponds to a 50-50 split, where there is no difference between the strategic AI agent and the reference model in that panel. Bonus rules above the horizontal solid black line are mutually achievable (both players can receive the bonus), and those below are competitive (only one player can receive the bonus). Green indicates that strategic AI agent significantly outperforms the reference model in more than 50% of games with that bonus rule, and grey indicates no significant difference. Error bars show 95% Clopper-Pearson confidence intervals. The bonus rules are as follows: **Gap Low**—selecting exactly $\{gap\}$ less than the opponent; **Gap High**—selecting exactly $\{gap\}$ more than the opponent; **Gap Abs.**—when absolute difference equals $\{gap\}$; **More Than**—when difference exceeds $\{gap\}$; **Equal**—for matching opponent's number; **Unequal**—for selecting different number than opponent; **Sum Even**—when sum of both numbers is even; **Sum Odd**—when sum of both numbers is odd; **Sum Upper**—when sum equals the upper bound; **Less Upper**—when the sum is less than the upper bound; **Coord. Low**—when both players select the lower bound.

likely to generalize to alternative distributions for π across the population of games.

5 Conclusion

The great promise of AI agents lies in their potential to accurately predict human behavior in novel settings. Realizing this capability could transform social science research and public policy. It could provide the social science equivalent of a lab bench in the physical sciences: an accurate, scalable playground to test ideas before large-scale and expensive implementation with humans.[25] Yet, with the current state-of-the-art foundation models, AI agents are not yet reliable enough to be used in this way out of the box.

In this paper, we explored an approach to address this shortcoming. Our approach relies on two key principles: (i) grounding candidate AI agents in theories expected to drive human behavior in the target setting, and (ii) optimizing and then validating AI agents in distinct but related settings presumed to be well-explained by the same theory. Without theoretical grounding, optimized prompts may fail to meaningfully improve even in-sample predictions. Without validation across distinct but related datasets, optimized prompts are prone to overfit a particular

[25]This could be even more powerful given the often-observed researcher inability to predict results of their own experiments (DellaVigna et al., 2019; Milkman et al., 2021; Gandhi et al., 2023, 2024; Duckworth et al., 2025).

data-generating process. Just as economists carefully extend established theories to novel policy contexts—relying on accumulated empirical validation rather than absolute certainty—optimizing theoretically-grounded AI agents to match samples of human data, and then validating them in distinct but related settings, provides a principled foundation for predicting humans in new settings.

The improvements in predictive power yielded by this approach are substantial. In four novel and preregistered strategic games derived from Arad and Rubinstein's 11-20 money request game, theory-grounded AI agents, optimized and validated through our methodology, reduced prediction errors by approximately 53-73% compared to baseline AI predictions. Remarkably, these theory-grounded agents predicted initial play in some of the games better than the original human data from Arad and Rubinstein. Importantly, our results are not confined to games involving strategic reasoning. In Appendix A, we apply the same procedure to the allocation games from Charness and Rabin, and, using data from a preregistered experiment with entirely novel human responses, we find substantial improvements.

Although this approach provides no statistical guarantees for arbitrary novel settings—indeed, no procedure can guarantee predictive power in entirely novel domains without a fully specified and correct causal model—we demonstrated that we can make externally valid inferences within a pre-committed family of settings. Using novel data from 4,249 participants playing 1,490 games randomly sampled from a heterogeneous population of 883,320 strategic games, we found that the strategic level-k agents generalized effectively across this broad domain. In 86% of games, all human subjects chose actions within the support of the optimized AI agent responses. These general agents substantially outperformed the baseline AI off-the-shelf, a cognitive hierarchy model, and the Harsanyi-Selten equilibria.

The results were also robust to several alternative specifications and were consistent across different game structures. Because the games were randomly sampled under the assumptions stated in Section 4.5, these results are externally valid for the entire population of 883,320 games. While we can only guarantee validity within this specific population and sampling distribution, the strong performance of theoretically motivated AI agents across such a diverse set of strategic games suggests that similar approaches would likely generalize to alternative distributions and even other strategic games.

Looking ahead, researchers could leverage this approach by identifying broad domains where they need predictions and can obtain training and validation data from representative subsets—even if there is substantial structural variation within the domain. Indeed, a key advantage of AI agents is that they can make predictions in response to any setting expressed in natural language without any prior data. This is often not possible with traditional machine learning methods and economic models.

Another exciting research direction would be to automate the theory-prediction-testing loop. Such a system would start with novel human data and relevant experimental settings, iteratively generate theory-informed candidate personas, optimize their parameters, and systematically evaluate their generalizability across samples. One can also imagine just starting with a novel setting,

and then having an AI system actually search for the relevant training and validation data to execute our approach. Recent research supports the feasibility of these ideas (Xie et al., 2025; Zhu et al., 2025b). In particular, Manning et al. (2024) demonstrate AI systems capable of automating the full social scientific workflow—from hypothesis generation to experimental simulation and data analysis.

More generally, the results in this paper linking theory-grounded AI agents to robust generalizability in novel settings, and atheoretical AI agents to failure, are notable for reasons beyond prediction (Hofman et al., 2021). They suggest that the underlying LLM has correctly learned the relevant relationships between the AI agents and human responses to the given setting. This is even more notable given that it is highly unlikely that these mappings were explicitly specified during training. Such a finding aligns with recent evidence that LLMs form rich internal representations of their human-generated training corpus rather than merely memorizing it (Lindsey et al., 2025; Ameisen et al., 2025). If true for LLMs and human behavior more generally, our prompt-alignment-generalizability exercises may offer more than improved predictive power. If an LLM armed with a particular theory-grounded prompt matches human data particularly well across a wide range of related settings, it might be evidence that the theory has a lot of explanatory power for the underlying human sample. Building on a young social scientific literature (Peterson et al., 2021; Hirasawa et al., 2022; Enke and Shubatt, 2023; Si et al., 2024; Ludwig and Mullainathan, 2024; Mullainathan and Rambachan, 2024; Batista and Ross, 2024; Movva et al., 2025), this could, in turn, provide researchers with robust machine learning methods to rapidly and efficiently inform promising new hypotheses.

References

Aher, G., R. I. Arriaga, and A. T. Kalai (2023): "Using large language models to simulate multiple humans and replicate human subject studies," in *Proceedings of the 40th International Conference on Machine Learning*, JMLR.org, ICML'23.

Allcott, H. (2015): "Site Selection Bias in Program Evaluation *," *The Quarterly Journal of Economics*, 130, 1117–1165.

Almaatouq, A., T. L. Griffiths, J. W. Suchow, M. E. Whiting, J. Evans, and D. J. Watts (2024): "Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences," *Behavioral and Brain Sciences*, 47, e33.

Alsobay, M., D. G. Rand, D. J. Watts, and A. Almaatouq (2025): "Integrative Experiments Identify How Punishment Impacts Welfare in Public Goods Games," .

Ameisen, E., J. Lindsey, A. Pearce, W. Gurnee, N. L. Turner, B. Chen, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. Ben Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, and J. Batson (2025): "Circuit Tracing: Revealing Computational Graphs in Language Models," *Transformer Circuits Thread*.

Andrews, I., D. Fudenberg, L. Lei, A. Liang, and C. Wu (2025): "The Transfer Performance of Economic Models," .

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023): "Prediction-powered inference," *Science*, 382, 669–674.

Anthis, J. R., R. Liu, S. M. Richardson, A. C. Kozlowski, B. Koch, E. Brynjolfsson, J. Evans, and M. S. Bernstein (2025): "Position: LLM Social Simulations Are a Promising Research Method," in *Forty-second International Conference on Machine Learning Position Paper Track*.

Arad, A. and A. Rubinstein (2012): "The 11-20 Money Request Game: A Level-k Reasoning Study," *American Economic Review*, 102, 3561–73.

Argyle, L. P., E. C. Busby, N. Fulda, J. Gubler, C. Rytting, and D. Wingate (2022): "Out of One, Many: Using Language Models to Simulate Human Samples," *arXiv preprint arXiv:2209.06899*.

Arjovsky, M., L. Bottou, I. Gulrajani, and D. Lopez-Paz (2020): "Invariant Risk Minimization," .

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

Bai, Y., A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan (2022): "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," .

Batista, R. and J. Ross (2024): "Words that Work: Using Language to Generate Hypotheses," Available at SSRN: https://ssrn.com/abstract=4926398 or http://dx.doi.org/10.2139/ssrn.4926398.

Ben-David, S., J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan (2010): "A theory of learning from different domains," *Machine Learning*, 79.

Binz, M., E. Akata, M. Bethge, F. Brändle, F. Callaway, J. Coda-Forno, P. Dayan, C. Demircan, M. K. Eckstein, N. Éltető, T. L. Griffiths, S. Haridi, A. K. Jagadish, L. Ji-An, A. Kipnis, S. Kumar, T. Ludwig, M. Mathony, M. Mattar, A. Modirshanechi, S. S. Nath, J. C. Peterson, M. Rmus, E. M. Russek, T. Saanum, N. Scharfenberg, J. A. Schubert, L. M. S. Buschoff, N. Singhi, X. Sui, M. Thalmann, F. Theis, V. Truong, V. Udandarao, K. Voudouris, R. Wilson, K. Witte, S. Wu, D. Wulff, H. Xiong, and E. Schulz (2024): "Centaur: a foundation model of human cognition," .

Binz, M. and E. Schulz (2023): "Using cognitive psychology to understand GPT-3," *Proceedings of the National Academy of Sciences*, 120, e2218523120.

Brand, J., A. Israeli, and D. Ngwe (2023): "Using GPT for Market Research," *Harvard Business School Marketing Unit Working Paper*, 23-062.

Broska, D., M. Howes, and A. van Loon (2025): "The Mixed Subjects Design: Treating Large Language Models as Potentially Informative Observations," *Sociological Methods & Research*, 54, 1074–1109.

Bui, N., H. T. Nguyen, S. Kumar, J. Theodore, W. Qiu, V. A. Nguyen, and R. Ying (2025): "Mixture-of-Personas Language Models for Population Simulation," .

Bybee, J. L. (2025): "The Ghost in the Machine: Generating Beliefs with Large Language Models," Working paper, University of Chicago Booth School of Business.

Camerer, C. F. (2003): *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton, NJ: Princeton University Press.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004): "A Cognitive Hierarchy Model of Games*," *The Quarterly Journal of Economics*, 119, 861–898.

Capra, C. M., A. Gonzalez-Bonorino, and E. Pantoja (2024): "LLMs Model Non-WEIRD Populations: Experiments with Synthetic Cultural Agents," SSRN Electronic Journal.

Cerina, R. and R. Duch (2025): "The 2024 US Presidential Election PoSSUM Poll," *PS: Political Science & Politics*, 58, 286–297.

Chang, S., A. Chaszczewicz, E. Wang, M. Josifovska, E. Pierson, and J. Leskovec (2024): "LLMs generate structurally realistic social networks but overestimate political homophily," .

Charness, G., B. Jabarian, and J. A. List (2025): "The next generation of experimental research with LLMs," *Nature Human Behaviour*, 9, 833–835.

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

Cheng, M., T. Piccardi, and D. Yang (2023): "CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations," *ArXiv*, abs/2310.11501.

Dehejia, R., C. Pop-Eleches, and C. S. and (2019): "From Local to Global: External Validity in a Fertility Natural Experiment," *Journal of Business & Economic Statistics*, 39, 217–243.

DellaVigna, S., D. Pope, and E. Vivalt (2019): "Predict science to improve science," *Science*, 366, 428–429.

Duckworth, A. L., A. Ko, K. L. Milkman, J. S. Kay, E. Dimant, D. M. Gromet, A. Halpern, Y. Jung, M. K. Paxson, R. A. S. Zumaran, R. Berman, I. Brody, C. F. Camerer, E. A. Canning, H. Dai, M. Gallo, H. E. Hershfield, M. D. Hilchey, A. Kalil, K. M. Kroeper, A. Lyon, B. S. Manning, N. Mazar, M. Michelini, S. E. Mayer, M. C. Murphy, P. Oreopoulos, S. E. Parker, R. Rondina, D. Soman, and C. V. den Bulte (2025): "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*, 122, e2418616122.

Egami, N., M. Hinck, B. Stewart, and H. Wei (2023): "Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models," in *Advances in Neural Information Processing Systems*, ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Curran Associates, Inc., vol. 36, 68589–68601.

Enke, B. and C. Shubatt (2023): "Quantifying Lottery Choice Complexity," Working Paper 31677, National Bureau of Economic Research.

Fish, S., Y. A. Gonczarowski, and R. I. Shorrer (2025): "Algorithmic Collusion by Large Language Models," .

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

Fudenberg, D. and A. Liang (2019): "Predicting and Understanding Initial Play," *American Economic Review*, 109, 4112–41.

Gandhi, L., A. Kiyawat, C. F. Camerer, and D. J. Watts (2023): "Hypothetical Nudges Provide Misleading Estimates of Real Behavior Change," Tech. rep., University of Pennsylvania, available at OSF Preprints: https://osf.io/preprints/psyarxiv/c7mkf.

Gandhi, L., B. S. Manning, and A. L. Duckworth (2024): "Effect Size Magnification: No Variable Is as Important as the One You're Thinking About—While You're Thinking About It," *Current Directions in Psychological Science*, 33, 347–354.

Gao, Y., D. Lee, G. Burtch, and S. Fazelpour (2024): "Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina," .

Gemini Team, G. (2024): "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," .

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

Hansen, A. L., J. J. Horton, S. Kazinnik, D. Puzzello, and A. Zarifhonarvar (2024): "Simulating the Survey of Professional Forecasters," *Available at SSRN*.

Hardy, M. D., S. Zhang, J. Hullman, J. M. Hofman, and D. G. Goldstein (2025): "Improving out-of-population prediction: The complementary effects of model assistance and judgmental bootstrapping," *International Journal of Forecasting*, 41, 689–701.

Harsanyi, J. C. and R. Selten (1988): *A General Theory of Equilibrium Selection in Games*, Cambridge, MA: MIT Press.

Heinze-Deml, C., J. Peters, and N. Meinshausen (2018): "Invariant Causal Prediction for Nonlinear Models," *Journal of Causal Inference*, 6, 20170016.

Henrich, J., R. Boyd, S. Bowles, C. F. Camerer, E. Fehr, H. Gintis, and R. McElreath (2001): "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies," *The American Economic Review*, 91, 73–78.

Hewitt, L., A. Ashokkumar, I. Ghezae, and R. Willer (2024): "Predicting Results of Social Science Experiments Using Large Language Models," *Preprint*.

Hirasawa, T., M. Kandori, and A. Matsushita (2022): "Using Big Data and Machine Learning to Uncover How Players Choose Mixed Strategies," Preliminary.

Hofman, J. M., D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, A. Vespignani, and T. Yarkoni (2021): "Integrating explanation and prediction in computational social science," *Nature*, 595, 181–188.

Horton, J. J. (2023): "Large language models as simulated economic agents: What can we learn from homo silicus?" Tech. rep., National Bureau of Economic Research.

Horton, J. J. and R. Horton (2024): "EDSL: Expected Parrot Domain Specific Language for AI Powered Social Science," Whitepaper, Expected Parrot.

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

Hotz, V. J., G. W. Imbens, and J. H. Mortimer (2005): "Predicting the efficacy of future training programs using past experiences at other locations," *Journal of Econometrics*, 125, 241–270.

Jackson, M. O., Q. Mei, S. W. Wang, Y. Xie, W. Yuan, S. Benzell, E. Brynjolfsson, C. F. Camerer, J. Evans, B. Jabarian, J. Kleinberg, J. Meng, S. Mullainathan, A. Ozdaglar, T. Pfeiffer, M. Tennenholtz, R. Willer, D. Yang, and T. Ye (2025): "AI Behavioral Science," This paper grew out of a workshop of the same name held at CASBS in spring 2025.

Kagel, J. H. and A. E. Roth (1995): *The Handbook of Experimental Economics*, Princeton University Press.

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

Keynes, J. M. (1936): *The General Theory of Employment, Interest, and Money*, London: Macmillan Cambridge University Press.

Khattab, O., A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, and C. Potts (2024): "DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines," in *The Twelfth International Conference on Learning Representations*.

Kim, J., M. Kovach, K.-M. Lee, E. Shin, and H. Tzavellas (2024): "Learning to be Homo Economicus: Can an LLM Learn Preferences from Choice," .

Klivans, A., K. Stavropoulos, and A. Vasilyan (2024): "Testable Learning with Distribution Shift," in *Proceedings of Thirty Seventh Conference on Learning Theory*, ed. by S. Agrawal and A. Roth, PMLR, vol. 247 of *Proceedings of Machine Learning Research*, 2887–2943.

Leng, Y., Y. Sang, and A. Agarwal (2024): "Reduce Disparity Between LLMs and Humans: Optimal LLM Sample Calibration," Available at SSRN: https://ssrn.com/abstract=4802019 or http://dx.doi.org/10.2139/ssrn.4802019.

Li, P., N. Castelo, Z. Katona, and M. Sarvary (2024): "Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis," *Marketing Science*, 0, null.

Lindsey, J., W. Gurnee, E. Ameisen, B. Chen, A. Pearce, N. L. Turner, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. B. Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, and J. Batson (2025): "On the Biology of a Large Language Model," *Transformer Circuits Thread*.

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

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

Ludwig, J., S. Mullainathan, and A. Rambachan (2025): "Large Language Models: An Applied Econometric Framework," Working Paper 33344, National Bureau of Economic Research.

Manning, B. S., K. Zhu, and J. J. Horton (2024): "Automated Social Science: Language Models as Scientist and Subjects," Tech. rep., NBER, accessed: 2024-03-12.

McKelvey, R. D. and T. R. Palfrey (1992): "An Experimental Study of the Centipede Game," *Econometrica*, 60, 803–836.

——— (1995): "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior*, 10, 6–38, received March 18, 1994.

Mei, Q., Y. Xie, W. Yuan, and M. O. Jackson (2024): "A Turing test of whether AI chatbots are behaviorally similar to humans," *Proceedings of the National Academy of Sciences*, 121, e2313925121.

Milkman, K. L., D. Gromet, H. Ho, J. S. Kay, T. W. Lee, P. Pandiloski, Y. Park, A. Rai, M. Bazerman, J. Beshears, L. Bonacorsi, C. F. Camerer, E. Chang, G. Chapman, R. Cialdini, H. Dai, L. Eskreis-Winkler, A. Fishbach, J. J. Gross, S. Horn,

A. Hubbard, S. J. Jones, D. Karlan, T. Kautz, E. Kirgios, J. Klusowski, A. Kristal, R. Ladhania, G. Loewenstein, J. Ludwig, B. Mellers, S. Mullainathan, S. Saccardo, J. Spiess, G. Suri, J. H. Talloen, J. Taxer, Y. Trope, L. Ungar, K. G. Volpp, A. Whillans, J. Zinman, and A. L. Duckworth (2021): "Megastudies improve the impact of applied behavioural science," *Nature*, 600, 478–483.

Modarressi, I., J. Spiess, and A. Venugopal (2025): "Causal Inference on Outcomes Learned from Text," .

Movva, R., K. Peng, N. Garg, J. Kleinberg, and E. Pierson (2025): "Sparse Autoencoders for Hypothesis Generation," .

Mullainathan, S. and A. Rambachan (2024): "From Predictive Algorithms to Automatic Generation of Anomalies," Working Paper 32422, National Bureau of Economic Research.

Mullainathan, S. and J. Spiess (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106.

Nagel, R. (1995): "Unraveling in Guessing Games: An Experimental Study," *The American Economic Review*, 85, 1313–1326.

Nash, J. (1951): "Non-Cooperative Games," *Annals of Mathematics*, 54, 286–295.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe (2022): "Training Language Models to Follow Instructions with Human Feedback," in *Advances in Neural Information Processing Systems*.

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

Park, J. S., C. Q. Zou, A. Shaw, B. M. Hill, C. Cai, M. R. Morris, R. Willer, P. Liang, and M. S. Bernstein (2024): "Generative Agent Simulations of 1,000 People," .

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

Peters, J., P. Bühlmann, and N. Meinshausen (2016): "Causal Inference by using Invariant Prediction: Identification and Confidence Intervals," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78, 947–1012.

Peterson, J. C., D. D. Bourgin, M. Agrawal, D. Reichman, and T. L. Griffiths (2021): "Using large-scale experiments and machine learning to discover theories of human decision-making," *Science*, 372, 1209–1214.

Qian, C., K. Zhu, J. J. Horton, B. S. Manning, V. Tsai, J. Wexler, and N. Thain (2025): "Strategic Tradeoffs Between Humans and AI in Multi-Agent Bargaining," .

Santurkar, S., E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto (2023): "Whose opinions do language models reflect?" in *Proceedings of the 40th International Conference on Machine Learning*, JMLR.org, ICML'23.

Sarkar, S. and K. Vafa (2024): "Lookahead Bias in Pretrained Language Models," *SSRN Electronic Journal*.

Savani, R. and T. L. Turocy (2025): "Gambit: The package for computation in game theory," Version 16.3.0.

Schelling, T. C. (1960): *The Strategy of Conflict*, Cambridge, MA: Harvard University Press.

Shah, A., K. Zhu, Y. Jiang, J. G. Wang, A. K. Dayi, J. J. Horton, and D. C. Parkes (2025): "Learning from Synthetic Labs: Language Models as Auction Participants," .

Si, C., D. Yang, and T. Hashimoto (2024): "Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers," .

Stahl, D. O. and P. W. Wilson (1994): "Experimental Evidence on Players' Models of Other Players," *Journal of Economic Behavior & Organization*, 25, 309–327.

——— (1995): "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10, 218–254.

Suh, J., E. Jahanparast, S. Moon, M. Kang, and S. Chang (2025): "Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions," .

Tadelis, S. (2013): *Game Theory: An Introduction*, Princeton, New Jersey: Princeton University Press.

Tranchero, M., C.-F. Brenninkmeijer, A. Murugan, and A. Nagaraj (2024): "Theorizing with Large Language Models," Working Paper 33033, National Bureau of Economic Research.

Vapnik, V. N. (1998): *Statistical Learning Theory*, John Wiley & Sons.

Wang, M., D. J. Zhang, and H. Zhang (2025): "Large Language Models for Market Research: A Data-augmentation Approach," .

Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou (2024): "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., NIPS '22.

Xie, Y., Q. Mei, W. Yuan, and M. O. Jackson (2025): "Using Language Models to Decipher the Motivation Behind Human Behaviors," .

Zhu, J., J. C. Peterson, B. Enke, et al. (2025a): "Capturing the complexity of human strategic decision-making with machine learning," *Nature Human Behaviour*.

Zhu, J.-Q., H. Xie, D. Arumugam, R. C. Wilson, and T. L. Griffiths (2025b): "Using Reinforcement Learning to Train Large Language Models to Explain Human Decisions," .

A Predicting behavior in novel allocation games

We further test the robustness of our approach by predicting human behavior on a set of novel allocation games. Unlike the strategic reasoning games studied in Sections 3 and 4, these games—adapted from Charness and Rabin (2002)'s (CR) experiments on social preferences—require individuals to balance their own monetary payoffs against those of others. They offer a distinct theoretical and empirical context for validating the generalizability of prompts identified using our approach. In addition to this new context, there are two key technical differences from the previous section. First, samples of agents are optimized over several training settings simultaneously. This further decreases the likelihood of overfitting on idiosyncratic features of any particular setting. Second, we employ the novel construction method and parameterize a prompt template to optimize the agents in-sample.

We follow the same analytical structure and use similar notation as in Section 3. We first briefly describe the dictator settings originally explored by CR, as these form our training dataset. Next, we detail our procedure for optimizing samples of AI agents, where the motivating theory is drawn directly from the social-preference models in CR. We then validate these subjects on a distinct set of two-player games studied by CR and demonstrate the pitfalls of optimizing over atheoretical prompts. Finally, we introduce a new series of structurally distinct three-player allocation games and use them to illustrate the empirical efficacy of our approach in novel games that were not in the LLM's training corpus.

A.1 Charness and Rabin (2002)'s unilateral dictator games

CR study a set of simple allocation decisions in which one player (the dictator) chooses between two ways of splitting money with a passive recipient. In one version of these games, for example, the dictator (Person B) unilaterally decides between the options "Left" and "Right":

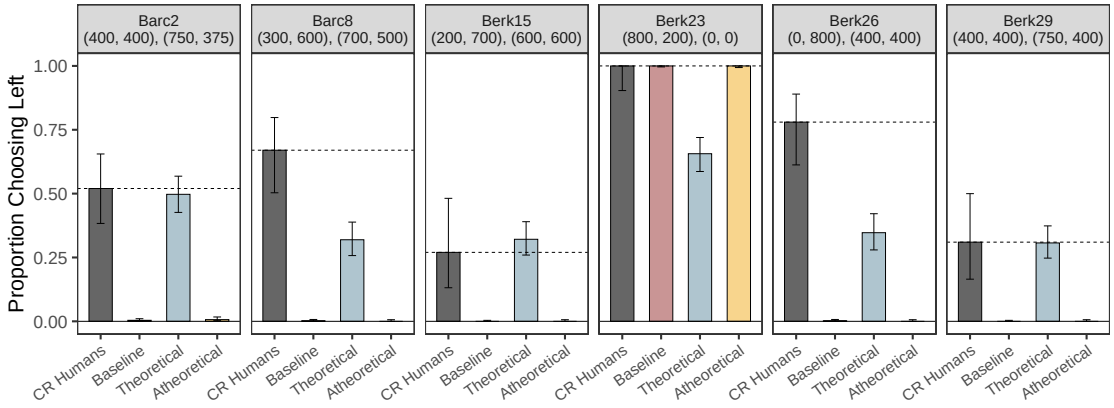
$$\underbrace{(\underbrace{400}_{\text{To A}}, \underbrace{600}_{\text{To B}})}_{\text{"Left"}} \quad \text{vs.} \quad \underbrace{(\underbrace{700}_{\text{To A}}, \underbrace{300}_{\text{To B}})}_{\text{"Right"}}.$$

CR collected human responses for six variations of this basic dictator setting, each featuring different payoff distributions. These six settings constitute our training dataset from which we derive the joint empirical distribution P of choosing Left.

Figure A1 shows the original results from CR. The columns represent different settings and show the payoffs for each player depending on the dictator's choice of "Left" or "Right". The y-axis shows the proportion of the sample that chose "Left" for each setting, and the black bars correspond to the distribution of human responses from CR. Besides the Pareto-dominated Berk23 setting, where everyone chooses "Right," the human data is balanced across the two options.

To establish the baseline ($\hat{P}_0$), we elicited 1,000 responses per setting from GPT-4o, without

Figure A1: Distribution of responses for the single-stage training dictator games



Notes: This figure reports the results of replications of the unilateral dictator games from Charness and Rabin (2002). Columns each represent a different game, and the x-axis corresponds to different samples of subjects playing each game. The y-axis shows the proportion of that sample choosing the option "Left." The black bars (and the dashed black lines) are the human responses from the original paper, red is the baseline AI agents, blue are the agents optimized using efficiency, self-interest, inequity aversion as parameters, and the yellow are atheoretical agents with preferences for the TV show new girl, taxidermy, and swimming. Error bars report 95% Wilson confidence intervals.

any additional instructions.[26] The red bars in Figure A1 represent these baseline AI responses. Notably, the baseline AI strongly favors choosing "Right" in nearly every setting, diverging sharply from the balanced human distributions. Quantitatively, this mismatch is substantial: using mean absolute error (MAE) as our distance metric, we find $\frac{1}{6}\sum_{s\in S} d(P_s, \hat{P}_0) = 0.42$. Given that the maximum possible MAE is 1, this is poor baseline predictive accuracy.

A.2 Constructing the sample of AI agents

CR hypothesize that a combination of efficiency concerns, inequity aversion, and self-interest is a key determinant of dictators' choices. To construct the sample of AI agents to better match the human data from these six settings simultaneously, we build a prompt template that incorporates these three traits as our theoretical motivation for the agents. Specifically, we parameterize each trait in the following prompt:

> $\theta(\phi_{eff}, \phi_{self}, \phi_{ineq})$ = *On a scale from 1 to 10, your efficiency level is:* $\{\phi_{eff}\}$. *10 means you strongly prioritize maximizing combined payoffs, and 1 means you don't care. On a scale from 1 to 10, your self-interest level is:* $\{\phi_{self}\}$. *10 means you strongly prioritize your own payoffs, and 1 means you don't care. On a scale from 1 to 10, your inequity aversion level is:* $\{\phi_{ineq}\}$. *10 means you strongly prioritize fairness between players, and 1 means you don't care.*

Our goal is to identify the parameter vector (or combination of vectors) that generates AI response distributions closely matching the observed human data. To do this, we create sets of $k = 3$ agents, each with a distinct parameter vector $\boldsymbol{\phi}$. Thus, each agent's prompt is $\theta(\boldsymbol{\phi})$,

[26] Horton (2023) also explore the baseline for the same games. Although their goal is to provide an early demonstration of AI simulation more generally.

where $\boldsymbol{\phi} = (\phi_{eff}, \phi_{self}, \phi_{ineq})$. We begin by randomly sampling 5 triples from the feasible space $\Phi = \{1, \ldots, 10\}^3$. For each sampled combination, we query the model 30 times per agent, producing the empirical distribution of responses P.

We then employ Bayesian optimization to iteratively search the parameter space, evaluating an additional 15 sets of parameter combinations (for a total of 20). Using mean absolute error to measure divergence from human data, this optimization identifies the optimal parameter vectors as: $\left(\boldsymbol{\phi}_1^*, \boldsymbol{\phi}_2^*, \boldsymbol{\phi}_3^*\right) = \big((7, 10, 10), (3, 1, 3), (1, 10, 2)\big)$. Assigning these parameters to three AI agents forms the optimized sample $\boldsymbol{\theta}^*$. As shown by the blue bars in Figure A1, the resulting distribution aligns much closer with the human responses: $\frac{1}{6}\sum_{s \in S} d(P_s, \hat{P}_{\boldsymbol{\theta}^*}) = 0.2$. This divergence represents a significant improvement, more than halving the baseline AI's error (MAE = 0.42).

A.3 Validation using two-stage games from Charness and Rabin (2002)

To validate whether $\boldsymbol{\theta}^*$ generalizes to new games, we apply the same prompt template and the values to a new set of more complicated sequential two-stage games from CR—the test set. Like the validation variants from AR (costless and cycle), these games are plausibly driven by similar underlying mechanisms as the training games, but are still different enough to provide a nontrivial test of generalization. In the first stage, Person A chooses either a given allocation or lets Person B choose one of two other known allocations. Person B chooses an allocation but is not informed of Person A's choice—until the payoffs are realized. For example, in one game, players are shown the following options:

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

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

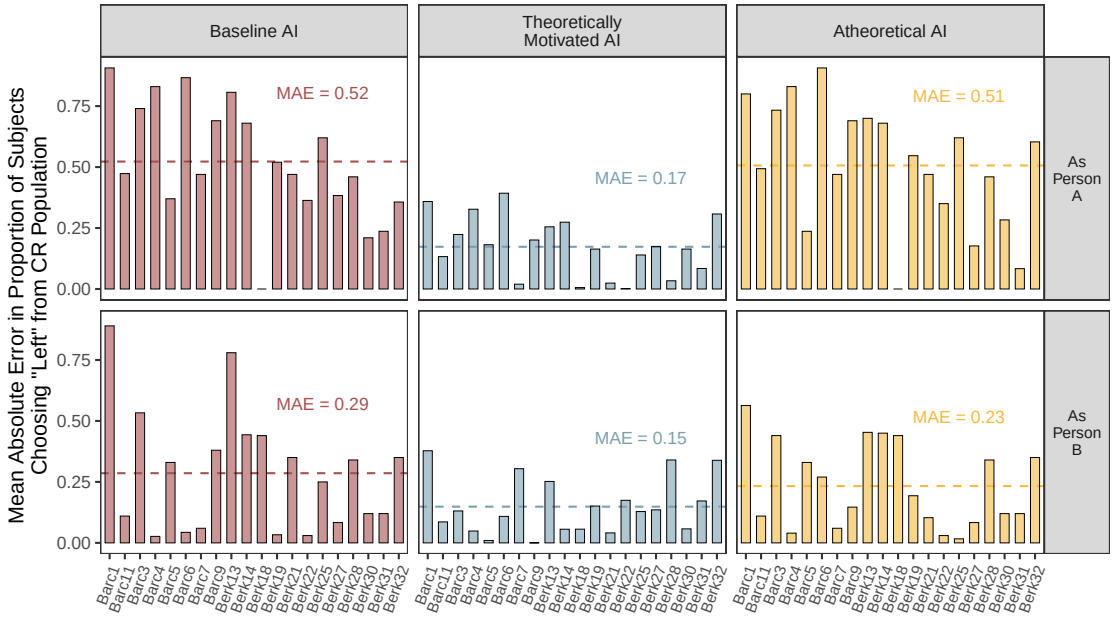
Table A1 in Appendix D provides all 20 versions of these two-stage games (each with a different set of payoffs), along with the human results from CR.

As a baseline, we elicit GPT-4o's responses to these 20 games 150 times each with the temperature set to 1. We then do the same for the theory-grounded sample $\boldsymbol{\theta}^*$—each of the three agents in the mixture plays each game 50 times.

Figure A2 shows the results. The top row shows the responses for the AI agents as Person A, and the bottom row for Person B. Each column corresponds to a different sample of subjects. The x-axis shows the setting name and the y-axis shows the mean absolute difference between the fraction of AI agents choosing "Left" and the fraction of human subjects choosing "Left" in Charness and Rabin. The difference between the baseline (red) and selected AI agents (blue) is substantial. The MAE between the baseline and the human subjects as Player A (0.52) is three

times larger than that for the optimized agents relative to the humans (0.17). The difference in MAE is twice as large for Player B (0.29 vs. 0.15).

Figure A2: Distances between human and AI agents for the two-stage dictator games



Notes: This figure reports the results of replications of the sequential two-player games from Charness and Rabin (2002) with AI agents. Each row shows responses from either Person A (left) or Person B (right), while each column corresponds to a different set of subjects. The x-axis shows the game, and the y-axis shows the mean absolute difference between the fraction of AI agents choosing "Left" and the fraction of human subjects choosing "Left" in Charness and Rabin. The left column displays the baseline AI agents (red), the middle column is the selected AI agents (blue), and the right column shows the atheoretical AI agents (yellow). The horizontal dashed lines show the mean absolute error in each pane.

This predictive improvement is robust across settings. In 31 of the 40 total decisions (20 settings, each played as Person A and Person B), the optimized theory-grounded agents more accurately matched human behavior than the baseline AI. Never was the absolute error in a game larger than 0.50 for the optimized theory-grounded agents, which is less than the MAE for the baseline AI as player A.

A.4 Optimizing among atheoretical prompts

Similarly to our study of AR, grounding a prompt template in theory is important for generalization. We do this via negative example. Specifically, without a theoretical grounding, the optimization procedure may fail to find a prompt that even fits in-sample.

Unlike in the Section 3 with the Always Pick 'N' agents, we do not offer an analogous overfitting example. Because human data from multiple games is used for optimization, finding a sample of AI agents which overfits requires finding prompts which overfit to all six settings. This is a much more difficult task than finding a prompt which overfits to a single game. Indeed, this is an attractive feature of using multiple training samples to construct and validate agent samples.

To generate samples of arbitrary agents, we repeat the entire process from Section A.2 but replace the theory-grounded attributes (i.e., efficiency, inequity aversion, and self-interest) with wholly unscientific ones: a self-reported fondness for the TV show *New Girl*, an enthusiasm for taxidermy, and swimming ability. This new prompt template is:

> $\theta(\phi_{ng}, \phi_{tax}, \phi_{swim})$ = *On a scale from 1 to 10, you think the show New Girl is:* $\{\phi_{ng}\}$. *10 means you love New Girl, and 1 means you hate it. On a scale from 1 to 10, your passion for taxidermy is:* $\{\phi_{tax}\}$. *10 means you love taxidermy, and 1 means you hate it. On a scale from 1 to 10, your ability to swim is:* $\{\phi_{swim}\}$. *10 means you are a great swimmer, and 1 means you can't swim.*

Using identical hyperparameters and the same Bayesian optimization procedure, we search over this atheoretical space to see if any combination of $(\phi_{ng}, \phi_{tax}, \phi_{swim})$ (each a sample of 3 agents with their own parameter vector) could even match the original single-stage dictator games in-sample. The resulting atheoretical parameter vector was: $\left(\boldsymbol{\phi}^*_{ath-1}, \boldsymbol{\phi}^*_{ath-2}, \boldsymbol{\phi}^*_{ath-3}\right) = \big((5,\ 7,\ 1),\ (9,\ 9,\ 5),\ (7,\ 6,\ 8)\big)$. As shown in Figure A1 (yellow), $\boldsymbol{\theta}^*_{ath}$ constructed using these parameters and template failed to beat even the baseline AI agents' performance. In fact, throughout the search, no parameter combination for "loving *New Girl*," "passion for taxidermy," or "swimming skill" ever produced a distribution of choices that aligned more closely with real humans. This result demonstrates the importance of grounding AI agents in theoretical constructs.

This lack of improvement persisted in the two-stage validation games as well (Figure A2; rightmost column). The atheoretical AI agents and the baseline were effectively indistinguishable in their distribution of responses as Player A, and the atheoretical subjects were only a little better as Player B. Overall, the atheoretical AI agents were far less aligned than the theory-grounded AI agents. They were closer to the human data than the baseline in 32.5% of the settings, worse than the baseline in 22.5% of the settings, and identical in the remaining games. The only way these arbitrary prompts generalize is that their poor performance is consistent across settings.

A.5 Predicting the novel three-player games

We conclude this section by introducing a set of 8 novel three-player allocation games to evaluate $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^*_{ath}$ in new settings with a new participant pool. We recruited $n = 494$ participants from Prolific to make three allocation decisions drawn from eight distinct settings, each involving a choice between two monetary allocations. Participants were paid \$1.00 and could earn a bonus of up to an additional \$1.00, depending on their own or others' choices. A representative setting is:

$$\textbf{Option A:} \begin{cases} \$1.00 & \text{To Selected} \\ \$0.75 & \text{To Each Other Player} \end{cases} \qquad \textbf{Option B:} \begin{cases} \$0.50 & \text{To Selected} \\ \$1.00 & \text{To Each Other Player} \end{cases}$$

After completion, one of the three decisions was randomly selected for payment. Participants were then randomly grouped into triads, with one member randomly chosen as the Selected Player. All three members received bonuses according to the allocation chosen by their group's Selected

Player.[27] Multiple attention checks confirmed participants understood the instructions and the payoffs. Participants' decisions only determined payments if they were the Selected Player. Uncertainty over selection aimed to ensure that choices reflected genuine social preferences.

The entire experimental design—including settings, procedures, and the optimized AI agent parameters—was preregistered prior to data collection with human participants. Figure B1 in Appendix B shows the full instructions for an example setting. To the best of our knowledge, games with these exact payoffs have never been used in an experiment with publicly available data.[28]

Figure A3a shows the results for all four subject samples: human participants (black), baseline AI agents (red), theory-grounded AI agents ($\boldsymbol{\theta}^*$ in blue), and atheoretical AI agents ($\boldsymbol{\theta}^*_{ath}$ in yellow). Each column corresponds to a setting with the relevant options indicated, with the y-axis indicating the proportion of subjects choosing Option A.

Human responses generally reflect a fairly even split between options, except for the extreme setting 8, where participants unanimously select Option A. The baseline AI consistently diverges from human behavior, disproportionately favoring Option B in nearly every setting (MAE = 0.259). Atheoretical AI agents offer no relative improvement, with a slightly worse fit (MAE = 0.264).

On average, $\boldsymbol{\theta}^*$ better approximates human choices across settings (MAE = 0.206)—about 21% less than the baseline. This improvement is emphasized in Figure A3b, showing the per-game absolute error along with the MAE. Importantly, this performance improvement is not driven by a few outliers: the theoretically motivated sample matches or exceeds both baseline and atheoretical AI agents in five settings. And in the games where the baseline AI and atheoretical AI agents are better, the difference is not large.

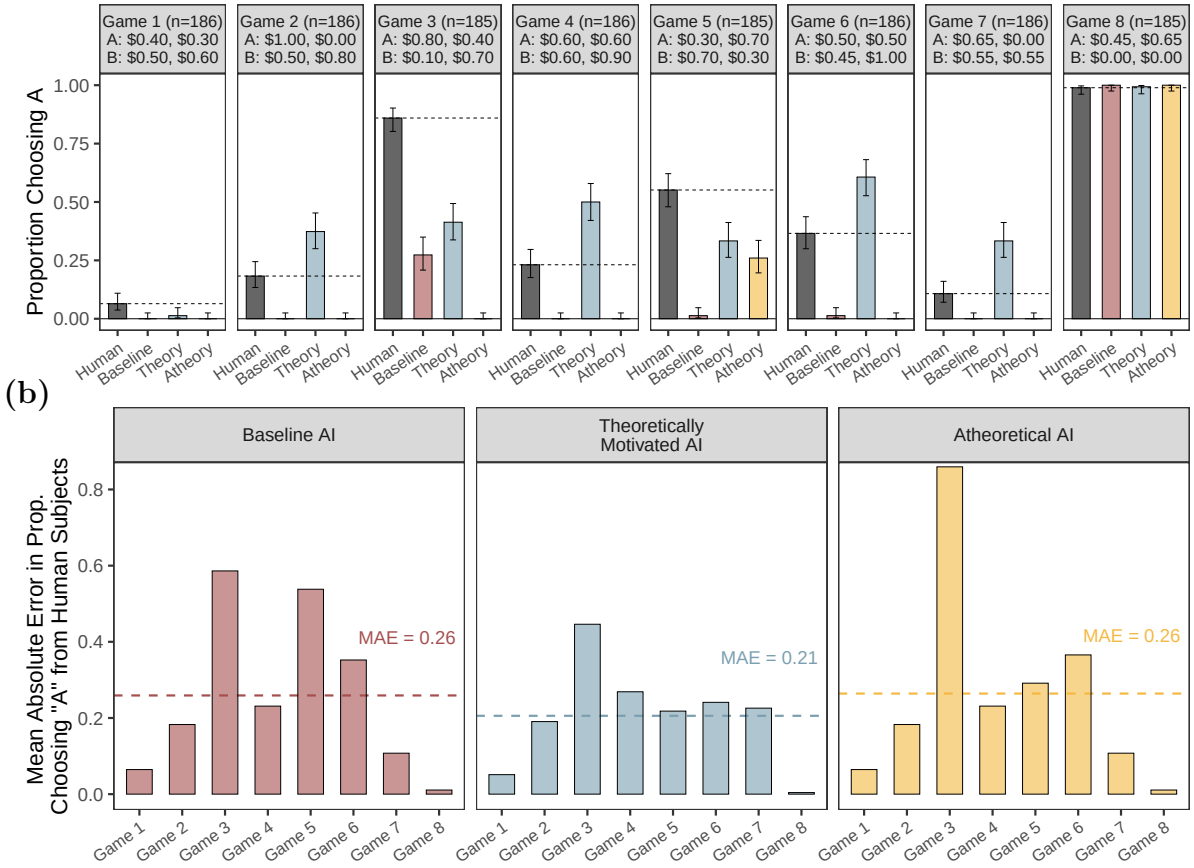
As with the results in Section 3, these findings demonstrate that theoretically grounded AI agents, optimized and then validated on data from related but distinct data-generating processes, can significantly improve the predictive power of AI simulations in novel settings.

[27] Suppose you, the reader, are completing this task and choose option A in the above setting. If after the survey is completed, the decision above is selected for payment and you are randomly chosen as the Selected Player, you will receive a $1.00 bonus, and the other two players each receive an extra $0.75. However, if another player was chosen as the Selected Player and they had picked Option A, then you would receive a $0.75 bonus payment.

[28] CR tested some 3-player games, but these had different payoffs, bonus rules, and involved imperfect information.

(a)

Figure A3: Results from the novel three-player allocation games



Notes: Panel (a) shows the proportion of choices for Option A across eight novel three-player allocation settings. Human responses are depicted in black (with the dashed lines), baseline AI in red, theoretically motivated AI in blue, and atheoretical AI in yellow. Error bars indicate 95% Wilson confidence intervals. Panel (b) presents the absolute error between human and AI choices across the settings, with dashed lines marking the MAE.

Table A1: Human subjects results for two-person response games in Charness and Rabin (2002)

		Human Subject Responses			
Game	Description	Out	Enter	Left	Right
Panel A: B's payoffs identical					
Barc7	A chooses (750,0) or lets B choose (400,400) vs. (750,400)	.47	.53	.06	.94
Barc5	A chooses (550,550) or lets B choose (400,400) vs. (750,400)	.39	.61	.33	.67
Berk28	A chooses (100,1000) or lets B choose (75,125) vs. (125,125)	.50	.50	.34	.66
Berk32	A chooses (450,900) or lets B choose (200,400) vs. (400,400)	.85	.15	.35	.65
Panel B: B's sacrifice helps A					
Barc3	A chooses (725,0) or lets B choose (400,400) vs. (750,375)	.74	.26	.62	.38
Barc4	A chooses (800,0) or lets B choose (400,400) vs. (750,375)	.83	.17	.62	.38
Berk21	A chooses (750,0) or lets B choose (400,400) vs. (750,375)	.47	.53	.61	.39
Barc6	A chooses (750,100) or lets B choose (300,600) vs. (700,500)	.92	.08	.75	.25
Barc9	A chooses (450,0) or lets B choose (350,450) vs. (450,350)	.69	.31	.94	.06
Berk25	A chooses (450,0) or lets B choose (350,450) vs. (450,350)	.62	.38	.81	.19
Berk19	A chooses (700,200) or lets B choose (200,700) vs. (600,600)	.56	.44	.22	.78
Berk14	A chooses (800,0) or lets B choose (0,800) vs. (400,400)	.68	.32	.45	.55
Barc1	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	.96	.04	.93	.07
Berk13	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	.86	.14	.82	.18
Berk18	A chooses (0,800) or lets B choose (0,800) vs. (400,400)	.00	1.00	.44	.56
Panel C: B's sacrifice hurts A					
Barc11	A chooses (375,1000) or lets B choose (400,400) vs. (350,350)	.54	.46	.89	.11
Berk22	A chooses (375,1000) or lets B choose (400,400) vs. (250,350)	.39	.61	.97	.03
Berk27	A chooses (500,500) or lets B choose (800,200) vs. (0,0)	.41	.59	.91	.09
Berk31	A chooses (750,750) or lets B choose (800,200) vs. (0,0)	.73	.27	.88	.12
Berk30	A chooses (400,1200) or lets B choose (400,200) vs. (0,0)	.77	.23	.88	.12

Notes: This table presents the complete set of two-person response games from Charness and Rabin along with human subject responses. This figure is identical to the one they show in the original paper. For each game, we show the proportion of subjects choosing each option. "Out" and "Enter" refer to Person A's initial choice, while "Left" and "Right" refer to Person B's choice if given the opportunity. All payoff values are in experimental currency units.

B All game instructions

Basic 11-20 Game

You and another player are playing a game in which each player requests an amount of money. The amount must be (an integer) between 11 and 20 shekels. Each player will receive the amount he requests. A player will receive an additional amount of 20 shekels if he asks for exactly one shekel less than the other player. What amount of money would you request?

Cycle 11-20 Game

You and another player are playing a game in which each player requests an amount of money. The amount must be (an integer) between 11 and 20 shekels. Each player will receive the amount of money he requests. A player will receive an additional amount of 20 shekels if: (i) he asks for exactly one shekel less than the other player or (ii) he asks for 20 shekels and the other player asks for 11 shekels. What amount of money would you request?

Costless 11-20 Game

You and another player are playing a game in which each player chooses an integer in the range 11-20. A player who chooses 20 will receive 20 shekels (regardless of the other player's choice). A player who chooses any other number in this range will receive three shekels less than in the case where he chooses 20. However, he will receive an additional amount of 20 shekels if he chooses a number that is one less than that chosen by the other player. Which number would you choose?

Basic 1-10 Game

You are going to play a game where you must select a whole number between 1 and 10. You will receive a number of points equivalent to that number. For example, if you select 3, you will get 3 points. If you select 7, you will get 7 points, etc. After you tell us your number, we will randomly pair you with another Prolific worker who is also playing this game. They will also have chosen a number between 1 and 10. If either of you select a number exactly one less than the other player's number, than the player with the lower number will receive an additional 10 points. Please choose a number between 1 and 10.

Cycle 1-10 Game

You are going to play a game where you must select a whole number between 1 and 10. You will receive a number of points equivalent to that number. For example, if you select 3, you will get 3 points. If you select 7, you will get 7 points, etc. After you tell us your number, we will randomly pair you with another Prolific worker who is also playing this

game. They will also have chosen a number between 1 and 10. There are 2 ways to win an additional 10 points based on both yours and the other player's choice: 1. If either of you select a number exactly one less than the other player's number, then the player with the lower number will receive an additional 10 points. 2. If either of you select 10 and the other selects 1, then the player who chose 10 will receive an additional 10 points. Please choose a number between 1 and 10.

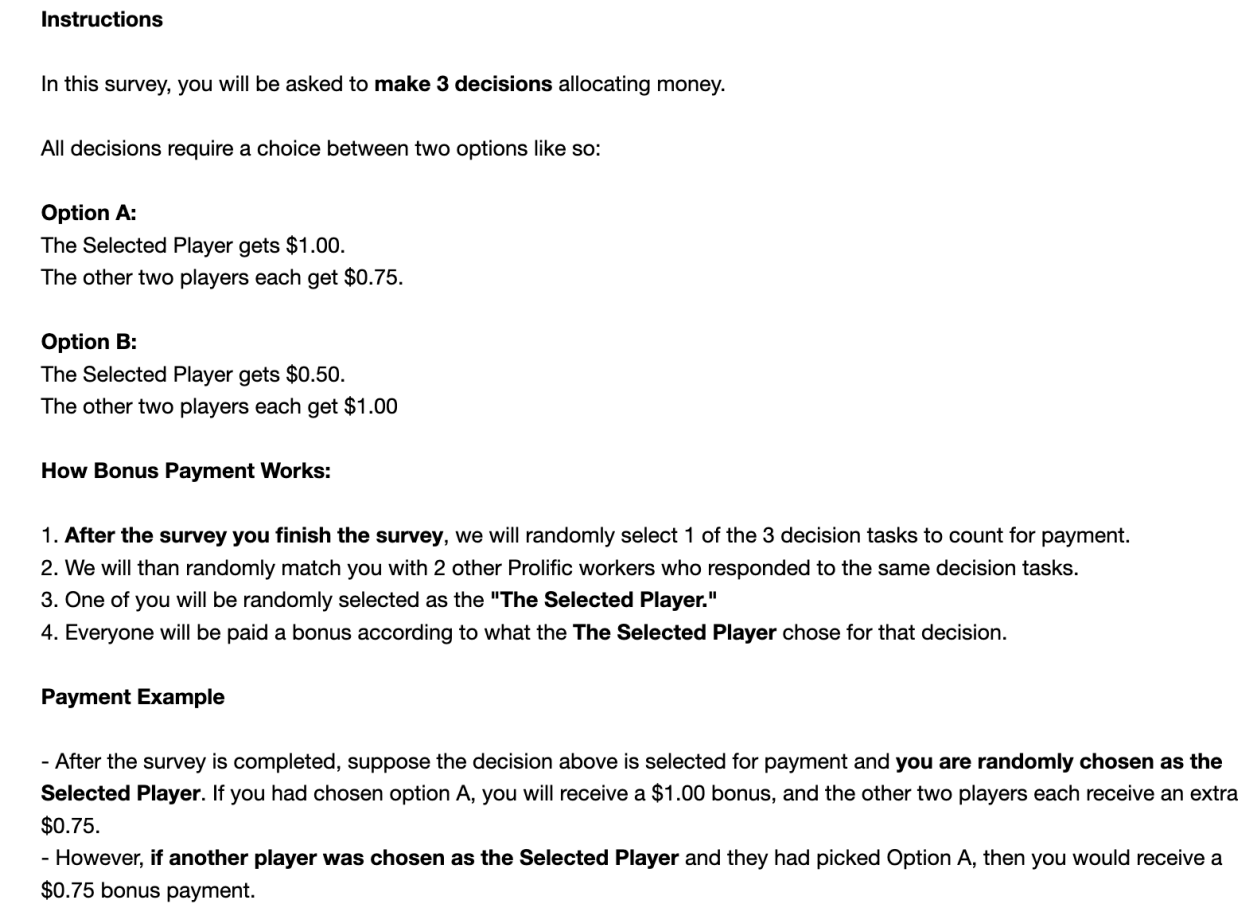
Costless 1-10 Game

You are going to play a game where you must select a whole number between 1 and 10. You will receive 10 points if you select the number 10 and you will receive 7 points for selecting any other number. After you tell us your number, we will randomly pair you with another Prolific worker who is also playing this game. They will also have chosen a number between 1 and 10. If either of you select a number exactly one less than the other player's number, than the player with the lower number will receive an additional 10 points. Please choose a number between 1 and 10.

1-7 Game

You are going to play a game where you must select a whole number between 1 and 7. You will receive a number of points equivalent to that number. For example, if you select 3, you will get 3 points. If you select 6, you will get 6 points, etc. After you tell us your number, we will randomly pair you with another Prolific worker who is also playing this game. They will also have chosen a number between 1 and 7. If either of you select a number exactly one less than the other player's number, than the player with the lower number will receive an additional 10 points. Please choose a number between 1 and 7.

Figure B1: Screenshot of the three-player game instructions



Instructions

In this survey, you will be asked to **make 3 decisions** allocating money.

All decisions require a choice between two options like so:

Option A:
The Selected Player gets $1.00.
The other two players each get $0.75.

Option B:
The Selected Player gets $0.50.
The other two players each get $1.00

How Bonus Payment Works:

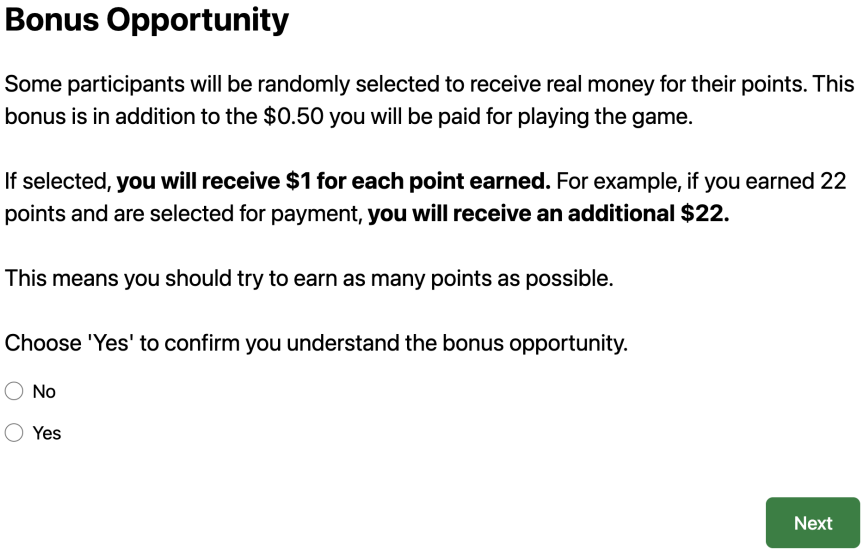
1. **After the survey you finish the survey**, we will randomly select 1 of the 3 decision tasks to count for payment.
2. We will than randomly match you with 2 other Prolific workers who responded to the same decision tasks.
3. One of you will be randomly selected as the **"The Selected Player."**
4. Everyone will be paid a bonus according to what the **The Selected Player** chose for that decision.

Payment Example

- After the survey is completed, suppose the decision above is selected for payment and **you are randomly chosen as the Selected Player**. If you had chosen option A, you will receive a $1.00 bonus, and the other two players each receive an extra $0.75.
- However, **if another player was chosen as the Selected Player** and they had picked Option A, then you would receive a $0.75 bonus payment.

Notes: This figure shows the instructions for the novel three-player allocation game presented to participants.

Figure B2: Bonus opportunity for the games in Section 4



Bonus Opportunity

Some participants will be randomly selected to receive real money for their points. This bonus is in addition to the $0.50 you will be paid for playing the game.

If selected, **you will receive $1 for each point earned.** For example, if you earned 22 points and are selected for payment, **you will receive an additional $22.**

This means you should try to earn as many points as possible.

Choose 'Yes' to confirm you understand the bonus opportunity.

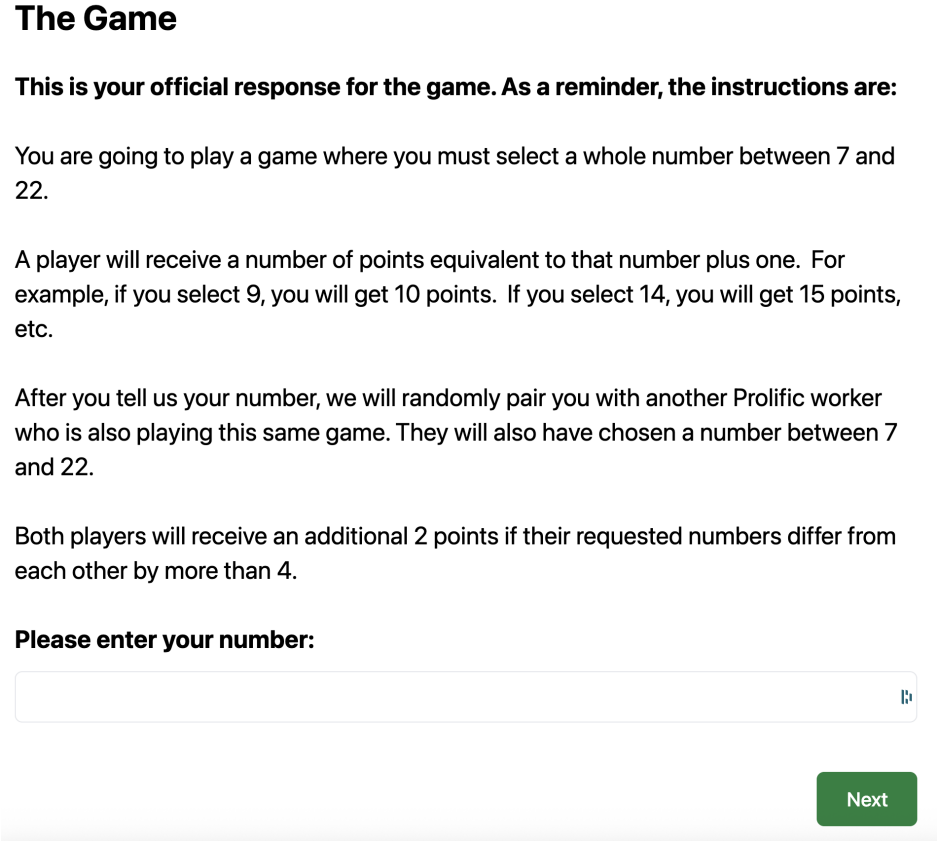
○ No

○ Yes

Next

Notes: This shows the instructions for the bonus opportunity presented to participants for the novel sample of 1,500 games.

Figure B3: Choosing a number for the assigned game in Section 4



Notes: This figure shows an example screenshot of participants selecting their number for their assigned game from the set of 1,500 games.

C Harsanyi-Selten selector implementation details

Let E be the finite set of Nash equilibria of a simultaneous, two-player normal-form game that is symmetric, so that the row player's payoff matrix is U and the column player's payoff matrix is $U^\top$. Harsanyi and Selten (1988)'s four-stage procedure deterministically selects a single equilibrium. For the vast majority of games in our setting, an equilibrium is selected in one of the first three steps. The fourth is barely used and is more a formality.

Our implementation follows that blueprint with two minimal deviations that (1) protect symmetric components in the Pareto filter and (2) enforce symmetry in the reported profile after tracing. The procedure deterministically returns a single selection; in symmetric games, and when the tracing routine returns normally, the selected profile is symmetric. The code is or will soon be available at `https://benjaminmanning.io/`. The following pseudocode broadly outlines the procedure.

Step 1 (component decomposition). Two equilibria $e = (\sigma^r, \sigma^c)$ and $e' = (\tau^r, \tau^c)$ are adjacent when they differ in exactly one player's strategy. The connected components of the resulting

graph—call them $C_1, \ldots, C_K$—are the equilibrium components.

Step 2 (Pareto filter with symmetry safeguard). For each component C_k compute its security vector $v(C_k) = \big(\min_{e \in C_k} u_1(e), \min_{e \in C_k} u_2(e)\big)$, where $u_1(e) = \sigma^{r\top} U \sigma^c$ and $u_2(e) = \sigma^{r\top} U^\top \sigma^c$. Delete C_k if some C_ℓ is strictly better in both coordinates.

Step 3 (symmetry filter and risk dominance). Discard all remaining components that contain no symmetric equilibrium. If multiple components survive, choose the one whose representative symmetric equilibrium (the first symmetric equilibrium encountered when iterating the component) minimizes the risk-dominance index

$$R(\sigma) = \sum_{i \neq j} \sigma_i \sigma_j \big[U_{ii} - U_{ji}\big]\big[U_{ii} - U_{ij}\big].$$

If two or more symmetric components attain exactly the same minimal value, we keep the first one encountered in iteration order. If no symmetric components remain after the symmetry filter, we select the first Pareto-surviving component as a fallback before Step 4.

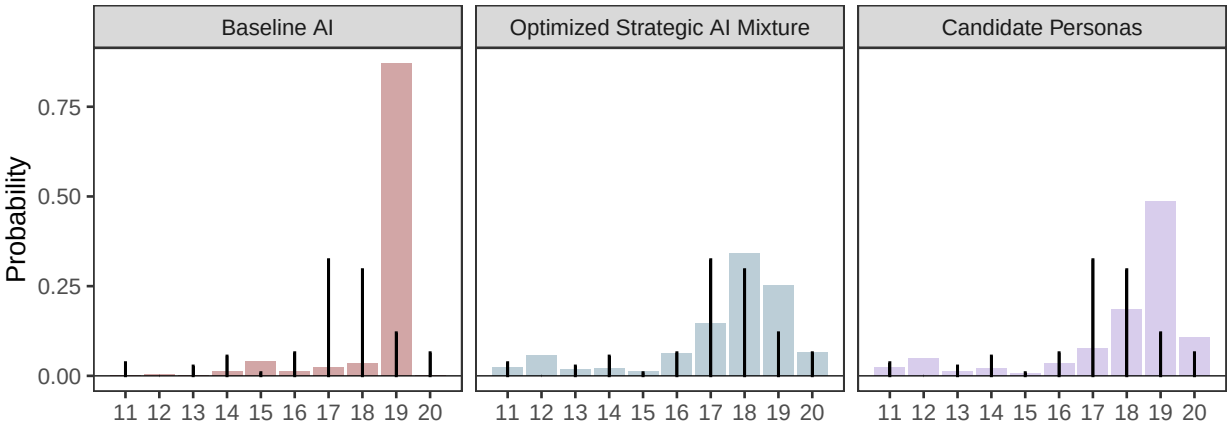
Step 4 (alpha-tracing). Let the winning component be the one selected in Step 3 (or the first Pareto-surviving component if no symmetric component remains).

- If the winning component is a singleton that already contains a symmetric equilibrium, we return it directly (no tracing).
- Otherwise, we run Gambit's logit α-tracing procedure on the full game—not restricted to the winning component—starting from the uniform prior. We follow the path to $\alpha = 1$ and take the resulting profile as the candidate equilibrium. To guard against numerical asymmetries, we then enforce symmetry in the reported profile by setting $\sigma^r = \sigma^c$ equal to the traced row strategy. Because the prior and the game are symmetric, the traced profile is generically symmetric; the coercion is a safeguard.
- *Deviation 2 (singleton asymmetric case).* If the winning component is a singleton asymmetric equilibrium, we run the same α-tracing procedure and then report the coerced symmetric profile as above. If the tracing routine raises an exception, the unique equilibrium is returned unchanged.

D Additional Tables and Figures

D.1 Permutations of the 11-20 game in Section 3

Figure D1: Response Distributions for the Basic Version for the 11-20 Game with raw candidate responses



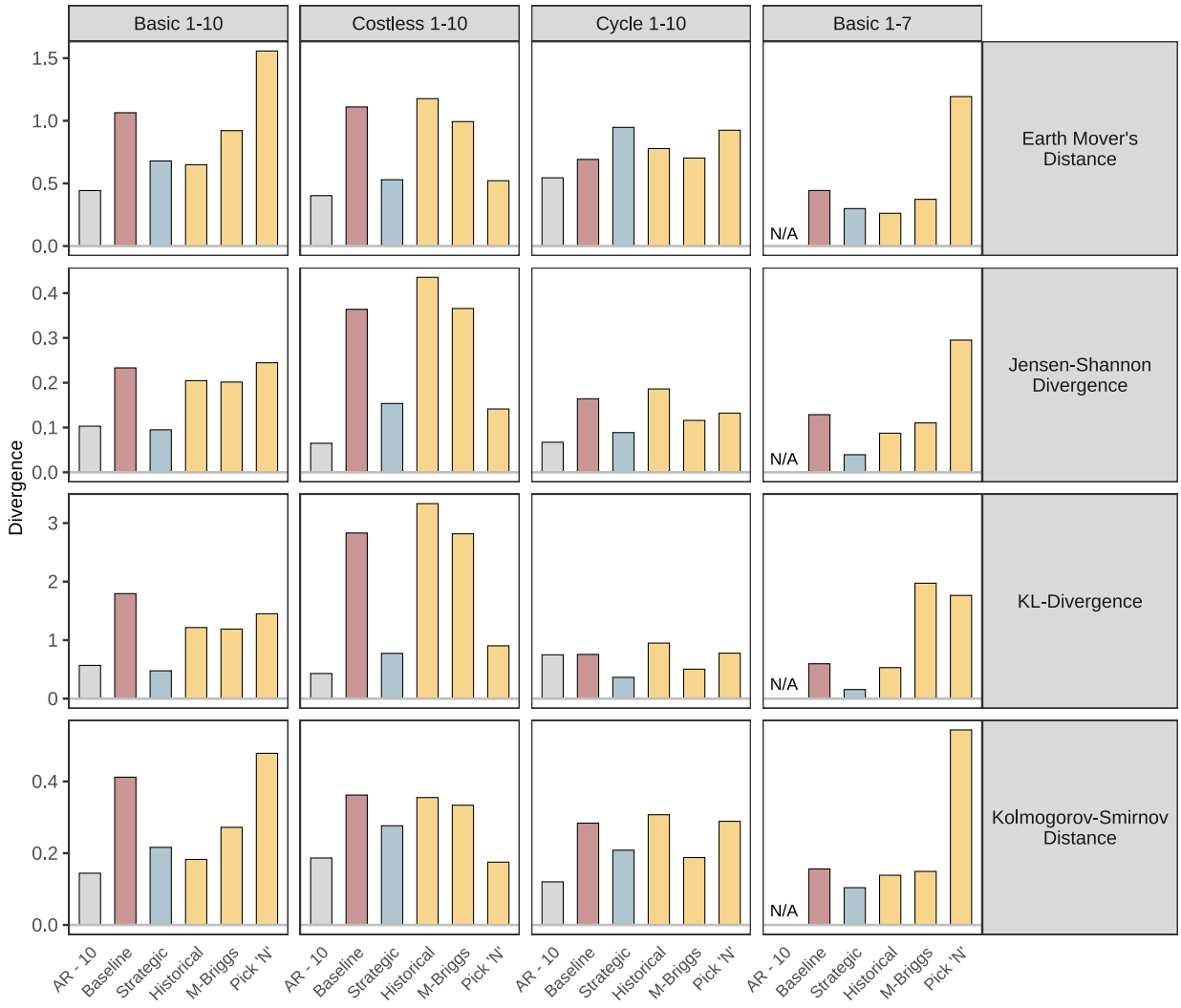
Notes: This figure displays empirical PMFs for three samples playing the basic 11-20 money request game: human subjects from Arad and Rubinstein (left panel), the naive baseline (center-left panel), responses from our selected AI agents based on the weights in Table 2 (center-right panel), and responses based on the unweighted and evenly distributed prompts in Table 2 (right panel).

Table D1: Atheoretical AI agents and resulting mixture weights

Historical Figures			
Persona	Weight	Persona	Weight
Cleopatra	0.000	Genghis Khan	0.000
Julius Caesar	0.891	Mother Teresa	0.000
Confucius	0.109	Martin Luther King	0.000
Joan of Arc	0.000	Frida Kahlo	0.000
Nelson Mandela	0.000	George Washington	0.000
Mahatma Gandhi	0.000	Winston Churchill	0.000
Harriet Tubman	0.000	Mansa Musa	0.000
Leonardo da Vinci	0.000	Sacagawea	0.000
Albert Einstein	0.000	Emmeline Pankhurst	0.000
Marie Curie	0.000	Socrates	0.000
MBTI Types			
Type	Weight	Type	Weight
You are an ESTJ	0.000	You are an ISTJ	0.000
You are an ESTP	0.000	You are an ISTP	0.000
You are an ESFJ	0.000	You are an ISFJ	0.000
You are an ESFP	0.000	You are an ISFP	0.000
You are an ENTJ	0.000	You are an INTJ	0.000
You are an ENTP	0.000	You are an INTP	0.000
You are an ENFJ	0.000	You are an INFJ	0.000
You are an ENFP	1.000	You are an INFP	0.000
Always Pick 'N'			
Number	Weight	Number	Weight
You always like to pick 11	0.037	You always like to pick 16	0.065
You always like to pick 12	0.000	You always like to pick 17	0.324
You always like to pick 13	0.028	You always like to pick 18	0.296
You always like to pick 14	0.056	You always like to pick 19	0.120
You always like to pick 15	0.009	You always like to pick 20	0.065

Notes: This table displays three sets of arbitrary prompts—each a different Θ. The weights columns display the optimized weights $\boldsymbol{w}^*$ when performing the selection method on the basic version of the 11-20 game. Weights sum to 1 within each set. For the historical figures, each prompt is told *"You are X"* where X is a historical figure. For the Myers-Briggs set, each prompt is also told that the four letters are in reference to the Myers-Briggs personality type indicator.

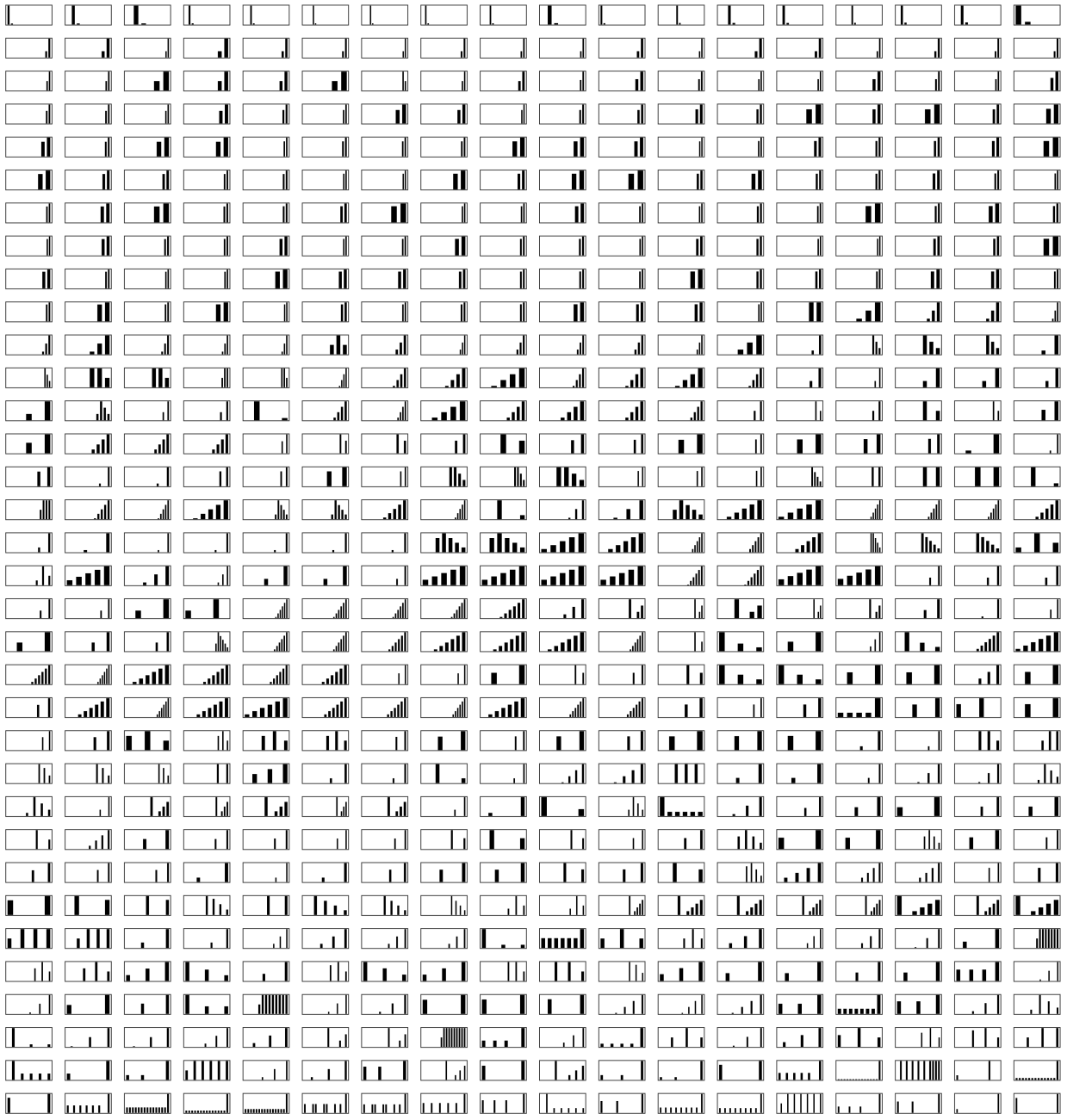
Figure D2: Comparison of novel 1-10 for alternative distance metrics



Notes: Reports the divergence between human and each AI distribution for the novel games across various additional distance metrics. For three of the four metrics, the optimized strategic agents outperform the baseline. Only in the costless version of the game, with the Earth Mover's distance as the metric, is the baseline slightly better.

D.2 Additional Tables and Figures for Section 4

Figure D3: Equilibrium PMFs across all games in S with mixed strategy Harsanyi-Selten solutions



Notes: Each panel shows the PMF of a mixed strategy Harsanyi-Selten equilibrium for a game in S. Panels are ordered by the variance of the equilibrium distribution (left to right, top to bottom). The x-axes are scaled freely, so games with different action spaces (e.g., 5, 11, or 20 options) span the same width. All y-axes use probability units but are scaled freely in height to better visualize different distributions. 612 games with mixed strategy equilibria are shown. The rest of S admit pure-strategy equilibria, collapsing to a single vertical bar (not shown).

D.3 Robustness checks for the family of pre-committed games in Section 4

Table D2: Summary statistics for absolute predictive accuracy of different models ($\varepsilon = 0$).

	Strategic AI	Baseline AI	HS Eq.	Cog. Hierarchy
% Humans Choose Max Prob. Strategy	24.3	16.8	30.4	28.5
% Humans Choose Top 3 Prob. Strategy	52.9	39.1	49.6	49.5
% Humans Choose Pos. Prob. Strategy	94.3	81.9	46.4	100.0
% Games Any Human Chooses Pos. Prob. Strategy	99.3	93.7	74.7	100.0
% Games All Humans Choose Pos. Prob. Strategy	86.3	65.3	17.7	100.0

Notes: This table reports summary statistics for the absolute predictive accuracy of different models in Section 4. These are the raw results without any smoothing. The first row corresponds to the proportion of human subjects who chose the most likely strategy for the given model in the columns. The second row corresponds to the proportion of human subjects who chose the top 3 most likely strategies for the given model in the columns. The third is the proportion of human subjects who chose a strategy with a positive probability for the given model in the columns. The fourth is the proportion of games in which any human chose a strategy with a positive probability for the given model in the columns. The fifth is the proportion of games in which all humans chose a strategy with a positive probability for the given model in the columns. Note that the cognitive hierarchy model, by design, provides positive probability for all strategies (even without smoothing) because the level-0 player chooses uniformly at random. None of the other models has this feature.

Table D3: Statistical tests comparing strategic AI agents vs other models ($\varepsilon = 0.05$)

Comparison (n Games)	$\bar{\Lambda}_S$	**Wilcoxon**	**Permutation Test**	$\sum_{s\in S} \mathbf{1}\{\hat{\Lambda}_s > 0\}/\lvert S\rvert$
Baseline AI	1.903***	$p < .001$***	$p < .001$***	0.726***
	(0.076)			(0.012)
Cognitive Hierarchy	0.848***	$p < .001$***	$p < .001$***	0.596***
	(0.082)			(0.013)
Harsanyi-Selten Nash	2.587***	$p < .001$***	$p < .001$***	0.730***
	(0.105)			(0.012)
Mixed	2.302***	$p < .001$***	$p < .001$***	0.732***
	(0.136)			(0.018)
Pure	2.788***	$p < .001$***	$p < .001$***	0.728***
	(0.149)			(0.015)
Random Pure Strategy	7.413***	$p < .001$***	$p < .001$***	0.942***
	(0.122)			(0.006)
Uniform	0.335***	$p < .001$***	$p < .001$***	0.598***
	(0.058)			(0.013)

Notes: This table shows the results of the statistical tests comparing the strategic AI agents to the other models for $\varepsilon = 0.05$. No smoothing is applied to the cognitive hierarchy model. The first column shows the comparison model. The second presents $e^{\bar{\Lambda}_S}$ with bootstrap confidence intervals comparing the strategic AI agents to the other models. The third and fourth columns present p-values for the Wilcoxon signed-rank test and random-sign permutation test, respectively. The fifth column presents the proportion of games for which the strategic AI agents is the best predictor with its 95% Clopper-Pearson interval. Significance Indicator: ***p<0.001, **p<0.01, *p<0.05.

Table D4: Statistical tests comparing optimized vs other models ($\varepsilon = 0.1$)

Comparison (n Games)	$\bar{\Lambda}_S$	**Wilcoxon**	**Permutation Test**	$\sum_{s\in S} \mathbf{1}\{\hat{\Lambda}_s > 0\}/\lvert S\rvert$
Baseline AI	1.588***	$p < .001$***	$p < .001$***	0.724***
	(0.063)			(0.012)
Cognitive Hierarchy	0.981***	$p < .001$***	$p < .001$***	0.619***
	(0.080)			(0.013)
Harsanyi-Selten Nash	1.713***	$p < .001$***	$p < .001$***	0.679***
	(0.089)			(0.012)
Mixed	1.562***	$p < .001$***	$p < .001$***	0.691***
	(0.111)			(0.019)
Pure	1.819***	$p < .001$***	$p < .001$***	0.671***
	(0.127)			(0.016)
Random Pure Strategy	5.779***	$p < .001$***	$p < .001$***	0.928***
	(0.103)			(0.007)
Uniform	0.468***	$p < .001$***	$p < .001$***	0.611***
	(0.051)			(0.013)

Notes: This table shows the results of the statistical tests comparing the strategic AI agents to the other models for $\varepsilon = 0.1$. No smoothing is applied to the cognitive hierarchy model. The first column shows the comparison model. The second presents $e^{\bar{\Lambda}_S}$ with bootstrap confidence intervals comparing the strategic AI agents to the other models. The third and fourth columns present p-values for the Wilcoxon signed-rank test and random-sign permutation test, respectively. The fifth column presents the proportion of games for which the strategic AI agents is the best predictor with its 95% Clopper-Pearson interval. Significance Indicator: ***p<0.001, **p<0.01, *p<0.05.

Table D5: Statistical tests comparing optimized vs other models ($\varepsilon = 0.2$)

Comparison (n Games)	$\bar{\Lambda}_S$	**Wilcoxon**	**Permutation Test**	$\sum_{s\in S} \mathbf{1}\{\hat{\Lambda}_s > 0\}/\lvert S\rvert$
Baseline AI	1.227***	$p < .001$***	$p < .001$***	0.715***
	(0.050)			(0.012)
Cognitive Hierarchy	1.105***	$p < .001$***	$p < .001$***	0.640***
	(0.078)			(0.012)
Harsanyi-Selten Nash	0.891***	$p < .001$***	$p < .001$***	0.622***
	(0.072)			(0.013)
Mixed	0.878***	$p < .001$***	$p < .001$***	0.647***
	(0.087)			(0.019)
Pure	0.899***	$p < .001$***	$p < .001$***	0.604***
	(0.104)			(0.017)
Random Pure Strategy	4.151***	$p < .001$***	$p < .001$***	0.902***
	(0.083)			(0.008)
Uniform	0.592***	$p < .001$***	$p < .001$***	0.643***
	(0.044)			(0.012)

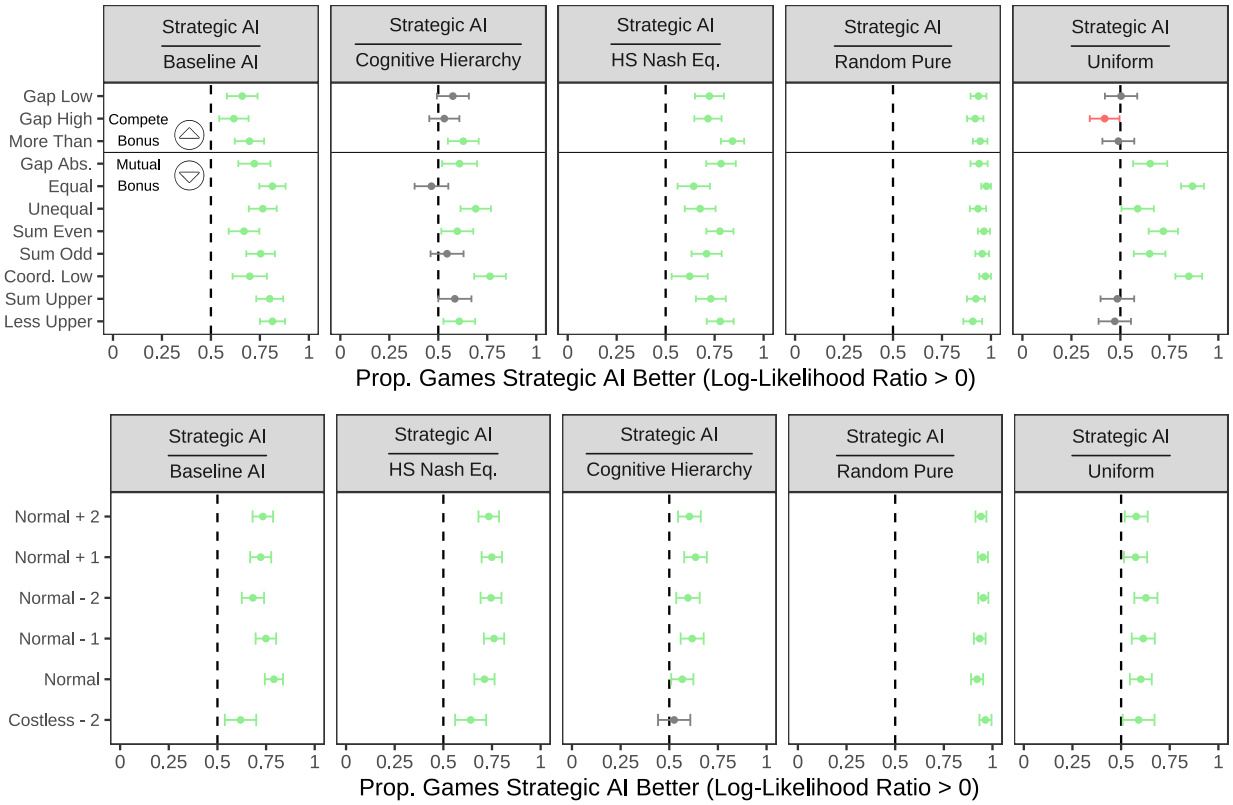
Notes: This table shows the results of the statistical tests comparing the strategic AI agents to the other models for $\varepsilon = 0.2$. No smoothing is applied to the cognitive hierarchy model. The first column shows the comparison model. The second presents $e^{\bar{\Lambda}_S}$ with bootstrap confidence intervals comparing the strategic AI agents to the other models. The third and fourth columns present p-values for the Wilcoxon signed-rank test and random-sign permutation test, respectively. The fifth column presents the proportion of games for which the strategic AI agents is the best predictor with its 95% Clopper-Pearson interval. Significance Indicator: ***p<0.001, **p<0.01, *p<0.05.

Table D6: Statistical tests comparing optimized vs other models ($\varepsilon = 0.3$)

Comparison (n Games)	$\bar{\Lambda}_{\mathbf{S}}$	**Wilcoxon**	**Permutation Test**	$\sum_{s\in S} \mathbf{1}\{\hat{\Lambda}_s > 0\}/\lvert S\rvert$
Baseline AI	0.985***	$p < .001$***	$p < .001$***	0.705***
	(0.042)			(0.012)
Cognitive Hierarchy	1.155***	$p < .001$***	$p < .001$***	0.647***
	(0.078)			(0.012)
Harsanyi-Selten Nash	0.446***	$p < .001$***	$p < .001$***	0.575***
	(0.062)			(0.013)
Mixed	0.517***	$p < .001$***	$p < .001$***	0.611***
	(0.073)			(0.020)
Pure	0.396***	$p < .001$***	$p < .001$***	0.551**
	(0.090)			(0.017)
Random Pure Strategy	3.191***	$p < .001$***	$p < .001$***	0.884***
	(0.070)			(0.008)
Uniform	0.642***	$p < .001$***	$p < .001$***	0.666***
	(0.039)			(0.012)

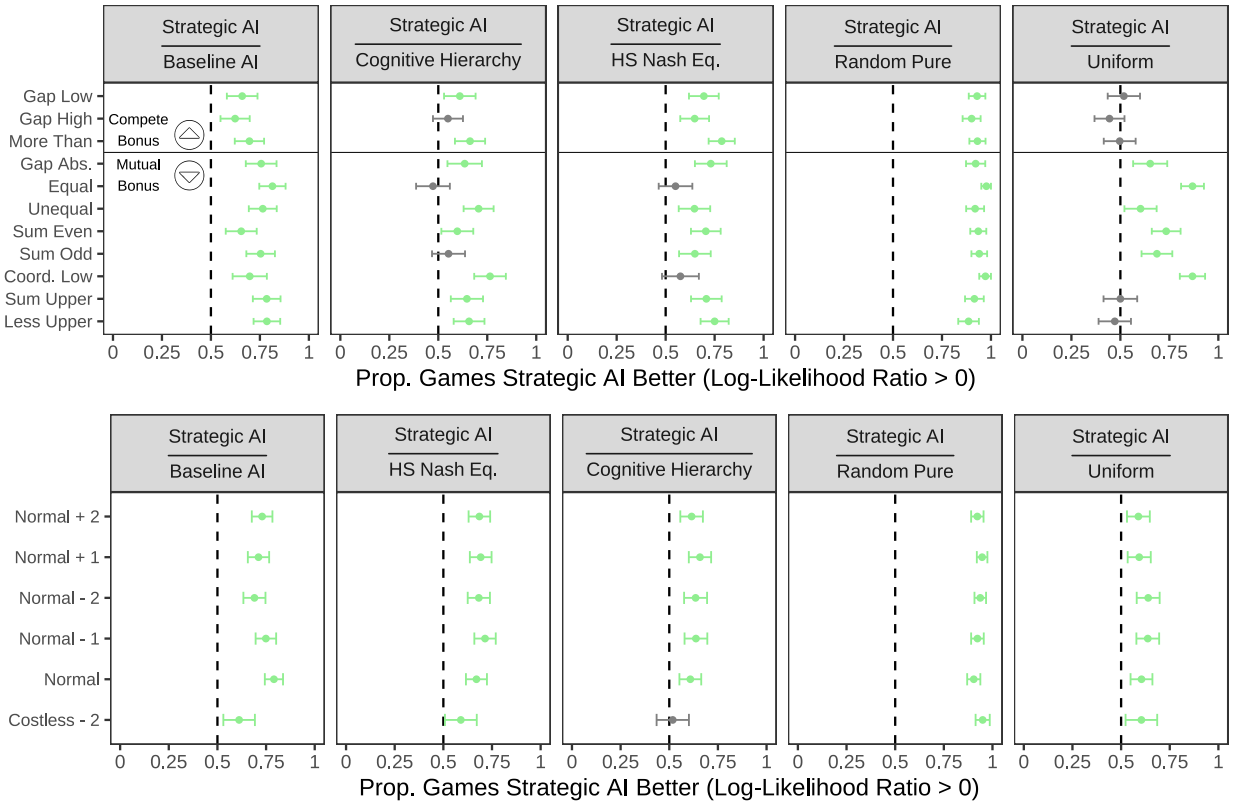
Notes: This table shows the results of the statistical tests comparing the strategic AI agents to the other models for $\varepsilon = 0.3$. No smoothing is applied to the cognitive hierarchy model. The first column shows the comparison model. The second presents $e^{\bar{\Lambda}_S}$ with bootstrap confidence intervals comparing the strategic AI agents to the other models. The third and fourth columns present p-values for the Wilcoxon signed-rank test and random-sign permutation test, respectively. The fifth column presents the proportion of games for which the strategic AI agents is the best predictor with its 95% Clopper-Pearson interval. Significance Indicator: ***p<0.001, **p<0.01, *p<0.05.

Figure D4: Relative predictive accuracy of the strategic AI agent vs other models for different game types ($\varepsilon = 0.05$)



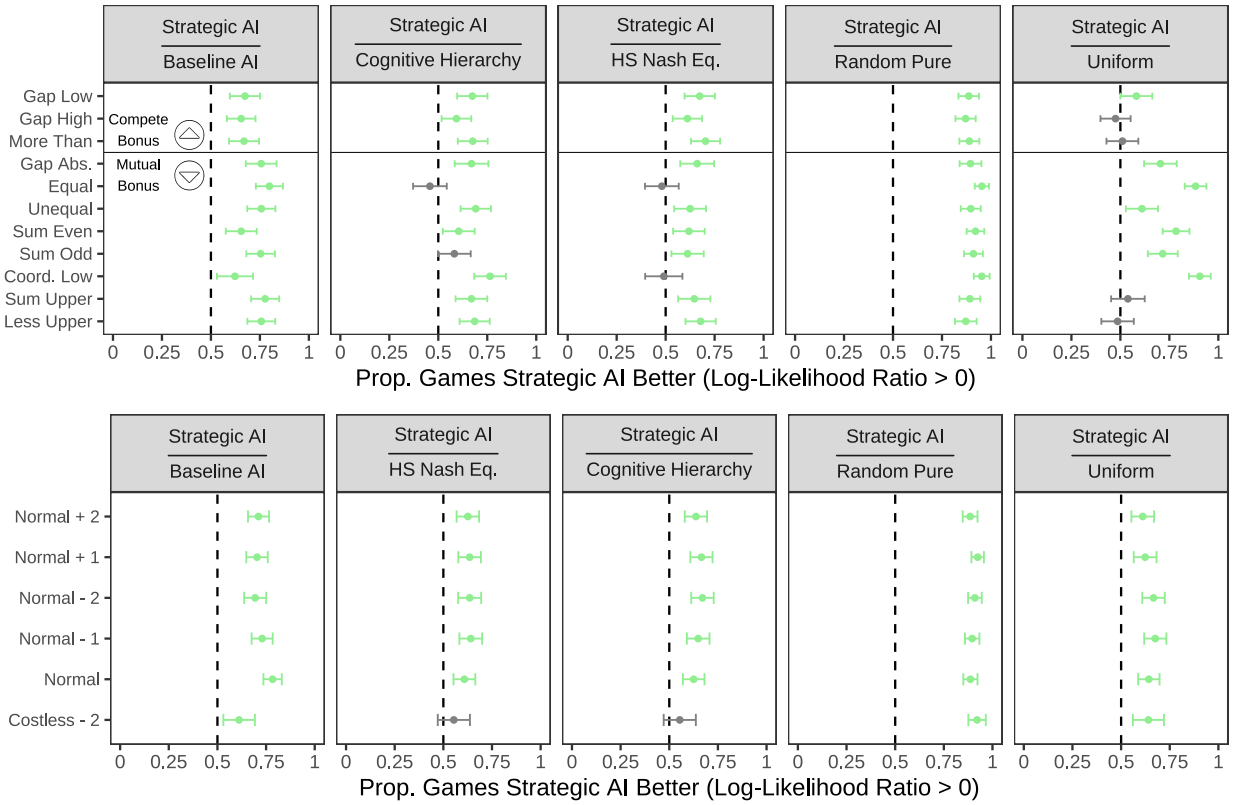
Notes: This figure shows the proportion of games for which strategic AI agents are the best predictor of initial play, separated by bonus rule (top panel) and points rule (bottom panel). All distributions except the cognitive hierarchy model are smoothed with $\varepsilon = 0.05$. The vertical dashed line represents equal performance (50-50 split). Green indicates that strategic AI agent significantly outperforms the reference model in more than 50% of games, red indicates significantly worse performance in more than 50% of games, and grey indicates no significant difference. Error bars show 95% Clopper-Pearson confidence intervals.

Figure D5: Relative predictive accuracy of the strategic AI agent vs other models for different game types ($\varepsilon = 0.1$)



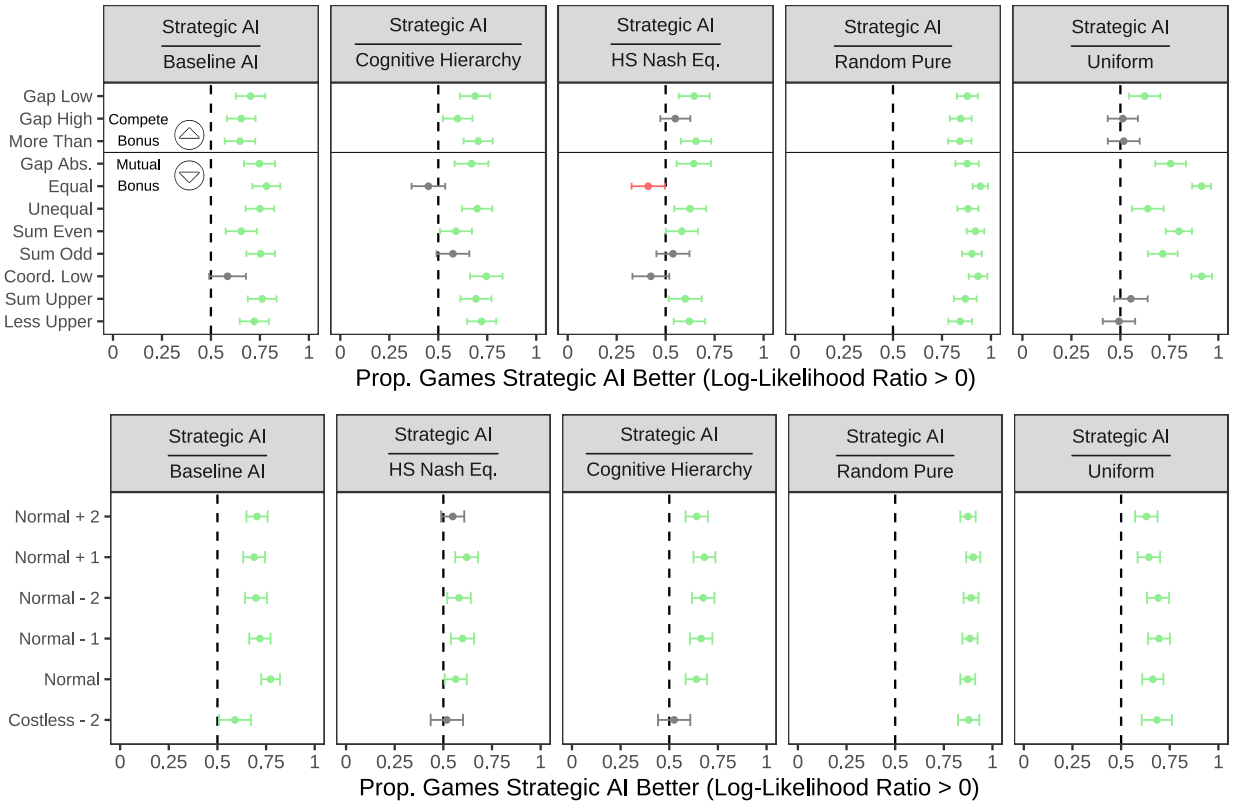
Notes: This figure shows the proportion of games for which strategic AI agents are the best predictor of initial play, separated by bonus rule (top panel) and points rule (bottom panel). All distributions except the cognitive hierarchy model are smoothed with $\varepsilon = 0.1$. The vertical dashed line represents equal performance (50-50 split). Green indicates that strategic AI agent significantly outperforms the reference model in more than 50% of games, red indicates significantly worse performance in more than 50% of games, and grey indicates no significant difference. Error bars show 95% Clopper-Pearson confidence intervals.

Figure D6: Relative predictive accuracy of the strategic AI agent vs other models for different game types ($\varepsilon = 0.2$)



Notes: This figure shows the proportion of games for which strategic AI agents are the best predictor of initial play, separated by bonus rule (top panel) and points rule (bottom panel). All distributions except the cognitive hierarchy model are smoothed with $\varepsilon = 0.2$. The vertical dashed line represents equal performance (50-50 split). Green indicates that strategic AI agent significantly outperforms the reference model in more than 50% of games, red indicates significantly worse performance in more than 50% of games, and grey indicates no significant difference. Error bars show 95% Clopper-Pearson confidence intervals.

Figure D7: Relative predictive accuracy of the strategic AI agent vs other models for different game types ($\varepsilon = 0.3$)



Notes: This figure shows the proportion of games for which strategic AI agents are the best predictor of initial play, separated by bonus rule (top panel) and points rule (bottom panel). All distributions except the cognitive hierarchy model are smoothed with $\varepsilon = 0.3$. The vertical dashed line represents equal performance (50-50 split). Green indicates that strategic AI agent significantly outperforms the reference model in more than 50% of games, red indicates significantly worse performance in more than 50% of games, and grey indicates no significant difference. Error bars show 95% Clopper-Pearson confidence intervals.

Table D7: Log-likelihood ratio regressions across game types ($\varepsilon = 0.05$)

	Log-Likelihood Ratio (Strategic AI / {...})				
	{Baseline AI}	{HS Nash Eq.}	{Cognitive Hierarchy}	{Random Pure}	{Uniform}
	(1)	(2)	(3)	(4)	(5)
Bonus Size	0.022	−0.046**	0.033*	0.042*	0.028**
	(0.013)	(0.017)	(0.014)	(0.020)	(0.010)
Lower Bound	0.035*	−0.020	−0.033*	−0.012	−0.019
	(0.014)	(0.018)	(0.015)	(0.021)	(0.011)
Action Space Size	0.098***	0.064**	−0.009	0.165***	0.035**
	(0.017)	(0.022)	(0.017)	(0.025)	(0.012)
Gap	−0.012	0.211*	0.078	−0.126	−0.052
	(0.070)	(0.092)	(0.076)	(0.106)	(0.052)
Constant	0.139	1.960***	0.782*	5.406***	−0.048
	(0.334)	(0.441)	(0.353)	(0.510)	(0.250)
Observations	1,477	1,477	1,477	1,477	1,477
R^2	0.029	0.015	0.008	0.031	0.013

Notes: Each column reports regression estimates where the dependent variable is the log-likelihood ratio of the strategic AI agent to the other models. The independent variables are the game features from Table 3 (excluding the bonus rule and points rule). We report heteroskedasticity-robust standard errors. Significance Indicator: ***p<0.001, **p<0.01, *p<0.05.

Table D8: Log-likelihood ratio regressions across game types ($\varepsilon = 0.1$)

	Log-Likelihood Ratio (Strategic AI / {...})				
	{Baseline AI}	{HS Nash Eq.}	{Cognitive Hierarchy}	{Random Pure}	{Uniform}
	(1)	(2)	(3)	(4)	(5)
Bonus Size	0.020	−0.032*	0.031*	0.039*	0.026**
	(0.011)	(0.014)	(0.014)	(0.017)	(0.009)
Lower Bound	0.028*	−0.018	−0.031*	−0.011	−0.017
	(0.011)	(0.015)	(0.014)	(0.018)	(0.009)
Action Space Size	0.088***	0.051**	0.001	0.150***	0.045***
	(0.014)	(0.019)	(0.017)	(0.021)	(0.011)
Gap	−0.011	0.172*	0.089	−0.103	−0.041
	(0.058)	(0.078)	(0.074)	(0.089)	(0.047)
Constant	0.038	1.176**	0.767*	3.917***	−0.063
	(0.279)	(0.370)	(0.345)	(0.428)	(0.223)
Observations	1,477	1,477	1,477	1,477	1,477
R^2	0.032	0.013	0.008	0.036	0.019

Notes: Each column reports regression estimates where the dependent variable is the log-likelihood ratio of the strategic AI agent to the other models. The independent variables are the game features from Table 3 (excluding the bonus rule and points rule). We report heteroskedasticity-robust standard errors. Significance Indicator: $^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$.

Table D9: Log-likelihood ratio regressions across game types ($\varepsilon = 0.2$)

	Log-Likelihood Ratio (Strategic AI / {...})				
	{Baseline AI}	{HS Nash Eq.}	{Cognitive Hierarchy}	{Random Pure}	{Uniform}
	(1)	(2)	(3)	(4)	(5)
Bonus Size	0.017	−0.019	0.029*	0.034*	0.024**
	(0.009)	(0.012)	(0.013)	(0.014)	(0.008)
Lower Bound	0.020*	−0.015	−0.029*	−0.009	−0.015
	(0.009)	(0.012)	(0.014)	(0.014)	(0.008)
Action Space Size	0.075***	0.036*	0.007	0.131***	0.051***
	(0.011)	(0.015)	(0.017)	(0.017)	(0.009)
Gap	−0.009	0.133*	0.100	−0.079	−0.030
	(0.047)	(0.063)	(0.073)	(0.072)	(0.040)
Constant	−0.073	0.478	0.785*	2.492***	−0.045
	(0.221)	(0.299)	(0.338)	(0.343)	(0.191)
Observations	1,477	1,477	1,477	1,477	1,477
R^2	0.036	0.009	0.008	0.042	0.028

Notes: Each column reports regression estimates where the dependent variable is the log-likelihood ratio of the strategic AI agent to the other models. The independent variables are the game features from Table 3 (excluding the bonus rule and points rule). We report heteroskedasticity-robust standard errors. Significance Indicator: $^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$.

Table D10: Log-likelihood ratio regressions across game types ($\varepsilon = 0.3$)

	Log-Likelihood Ratio (Strategic AI / {...})				
	{Baseline AI}	{HS Nash Eq.}	{Cognitive Hierarchy}	{Random Pure}	{Uniform}
	(1)	(2)	(3)	(4)	(5)
Bonus Size	0.015*	−0.012	0.027*	0.031**	0.022**
	(0.007)	(0.010)	(0.013)	(0.012)	(0.007)
Lower Bound	0.016*	−0.013	−0.027*	−0.008	−0.013
	(0.008)	(0.011)	(0.014)	(0.012)	(0.007)
Action Space Size	0.066***	0.024	0.008	0.115***	0.052***
	(0.009)	(0.013)	(0.017)	(0.014)	(0.008)
Gap	−0.007	0.108*	0.107	−0.063	−0.023
	(0.039)	(0.054)	(0.072)	(0.060)	(0.035)
Constant	−0.132	0.134	0.809*	1.702***	−0.021
	(0.184)	(0.256)	(0.334)	(0.290)	(0.167)
Observations	1,477	1,477	1,477	1,477	1,477
R^2	0.038	0.007	0.007	0.046	0.034

Notes: Each column reports regression estimates where the dependent variable is the log-likelihood ratio of the strategic AI agent to the other models. The independent variables are the game features from Table 3 (excluding the bonus rule and points rule). We report heteroskedasticity-robust standard errors. Significance Indicator: ***p<0.001, **p<0.01, *p<0.05.