The Mixed Subjects Design: Treating Generative Artificial Intelligence as (Potentially) Informative Observations in Experiments

David Broska, Michael Howes, and Austin van Loon

Draft. Please do not circulate.

Feedback welcome: dbroska@stanford.edu

Abstract: Large Language Models (LLMs) promise to transform the social sciences through affordable and near-instantly available predictions of human behavior. While LLMs also misrepresent human behavior, current approaches to studying causal effects with LLMs require researchers to assume that predicted and observed behavior are *interchangeable*. We argue that data on human subjects should be used to correct misrepresentations through LLMs in a *mixed-subjects design*. This paradigm offers valid and more precise estimates of causal effects at a lower cost than experiments conducted with human subjects only. We demonstrate – and extend – Prediction-Powered Inference, a recent statistical method that instantiates the mixed subjects design. Our innovation is a power analysis that allows for optimally choosing between recruiting human subjects or obtaining more affordable predictions to achieve a desired level of statistical power or maximize power under a budget constraint. We show empirically that the mixed subjects design increases the statistical precision and reliably reproduces treatment effects in an experiment with complex design — even though the predicted behavior on its own does not.

Keywords: Generative AI, Large Language Model, Experiment, Prediction-Powered Inference, Mixed-Subject Design, Power Analysis

1 Introduction

Large language models (LLMs)—a class of generative artificial intelligence trained on human-generated content—have been shown to mimic how humans respond to surveys and experimental treatments in a variety of settings. Accurately predicting rather than observing human behavior could serve as a cost-effective and near-instantly available alternative to observing humans' behavior with the potential to transform the social sciences. This approach to learning about social phenomena, known as "silicon sampling" (Argyle et al., 2023), might accelerate scientific progress, reduce inequities in access to costly evidence on hypotheses and research questions, and protect human subjects from deception and other risks associated with experimentation.

However, there is scant guidance on how to leverage LLMs for conducting scientifically valid research. Researchers who currently use LLMs to predict human behavior have to rely implicitly or explicitly on what we term the *interchangeability assumption*, a general premise that researchers adopt when drawing conclusions from predictive models of social processes rather than observations thereof (e.g. Friedman, 1953). When predicting human behavior, this assumption implies that the data extracted from LLMs closely correspond to human behavior or the responses given in a survey. The underlying logic of this approach is to treat silicon subjects *as if* they were human participants. This assumption leads to valid inference only if the predicted responses approximate—at least on average—the same parameter estimate as from the human subjects.

Unfortunately, there is growing evidence that LLMs inaccurately portray human behavior (Bisbee et al., 2024; Park et al., 2024; Takemoto, 2024). Even in settings where LLMs happen to realistically predict human behavior, there is a lack of generalizable procedures, metrics, and conventions to assess when this approximation is sufficiently accurate to be used in traditional null hypothesis testing. Currently, the interchangeability of predicted and observed behavior is assessed empirically on a case-by-case basis. This leads silicon sampling to be of minimal benefit in practice, since human subjects data must be collected alongside LLM predictions at a scale sufficient to adjudicate the veracity of the interchangeability assumption. Some have therefore suggested that simulations of human behavior should be confined to exploratory stages of research, such as LLM-powered pilot studies for anticipating effect sizes (Grossmann et al., 2023).

To address this gap, we propose a "mixed subjects" approach to designing research with LLMs. Rather than outright rejecting the assumption that human behavior and LLM predictions are interchangeable a priori, we argue that data collected from human subjects should inform inferences drawn from LLMs about human behavior in a coherent statistical framework. We propose that observations from human subjects can be leveraged to benchmark and correct for misrepresentations in LLM predictions. We

demonstrate how to implement this approach with Prediction-Powered Inference (PPI), a recent statistical method that instantiates the mixed subjects approach. PPI allows researchers to combine observed human behavior with behavior predicted by LLMs and other algorithms. So long as there is some correspondence between predicted and observed human behavior, PPI produces consistent point estimates with narrower confidence intervals compared to those derived solely from human subject samples.

To enhance the practicality of PPI for applied researchers, we derive a power analysis that allows researchers to assess the increases in statistical precision they can expect from including LLM predictions of human behavior in their study and optimally choose between recruiting human subjects and obtaining predictions of human behavior. Statistical software published alongside this article allows researchers to identify the cheapest combination of sample sizes of human subjects and LLM predictions for a desired level of power. Alternatively, the software allows researchers to allocate a fixed budget most effectively to a combination of silicon and human subjects that maximizes statistical power. With the extension of a power analysis, PPI becomes a fully usable methodology for conducting mixed-subject studies.

This article demonstrates how researchers can use LLM predictions not only in exploratory but also in confirmatory research. The mixed subjects design with PPI ensures the validity of point estimates while allowing researchers to achieve higher precision at a lower cost than with human subjects alone. Therefore, the mixed subjects design may provide many of the potential benefits of silicon sampling while avoiding its potential drawbacks.

2 The Silicon Subjects Design

Experiments in surveys, labs, and the field significantly enhanced our understanding of causal processes in the social sciences. However, experiments also face limitations, including the costs of conducting research, recruitment of harder-to-reach participants, and issues of measurement and generalizability. Below we outline how the silicon subjects approach promises to address these issues.

2.1 Promises of Silicon Sampling

The silicon subjects approach claims that LLMs can mimic the behavior of human participants in empirical studies based on a prompt given by the researcher. In the context of experiments, the prompt contains the experimental manipulation, with silicon subjects being randomly assigned to a condition. The prompt may also include a profile of study participants with demographics, attitudes, and other information. While such a profile is necessarily an incomplete representation of a participant, it allows researchers to create a "silicon sample" (Argyle et al., 2023) that matches the demographic makeup of the population of interest. The additional context provided to the LLM may increase the diversity of

responses one would expect in a human population or even increase predictive accuracy for treatment effects (Gui and Toubia, 2023). Based on studies that sought to reproduce findings of canonical experiments, some scholars concluded that LLM predictions were interchangeable with human behavior under certain conditions:

> *These findings could indicate that—at least in some instances—GPT-3 is not just a stochastic parrot and could pass as a valid subject for some of the experiments we have administered.* (Binz and Schulz, 2023)

> *Practically speaking, LLMs may be most useful as participants when studying specific topics, when using specific tasks, at specific research stages, and when simulating specific samples.* (Dillion et al., 2023)

If silicon subjects can substitute human participants, LLMs may help overcome the limitations of experiments that exclusively draw on responses from human subjects. A first set of issues relate to the cost of conducting experiments with human subjects. Experiments, particularly those that systematically assess a broad range of hypothesis (Almaatouq et al., 2024) and those aimed at testing heterogeneous treatment effects, require large number of participants to precisely identify an effect. The costs recruiting a sufficient number of human subjects may be prohibitive for researchers with more limited budgets. Silicon sampling offers a cost-effective alternative to human respondents since predicting the responses of human participants is less expensive than recruiting them to participate in a study.

A second set of issues relates to challenges in finding suitable participants for a study. While researchers often go to significant lengths to create a sample representative of a target population, certain participants remain hard to reach. For instance, typical online panels for survey research consist of more liberal than conservative respondents. Collecting a sample representative on this and other dimensions such as age, gender, income, and education may be infeasible since certain characteristics—and combinations thereof—are rare and not present among the participants available on a survey platform. If accurate, silicon sampling allows researchers to collect more observations on these otherwise hard-to-reach populations, with observations being nearly instantly available. Silicon subjects may even serve as alternative study populations if ethical concerns and risks limit the number of participants that can be recruited for experiments.

Finally, experimental research, like other quantitative scholarship, is only as good as the quality of measurements taken. Limited attention spans, distractions, and varying compliance with experimental protocols inhibit precise measurement of concepts and may lead to participant attrition during the study. While these features characterize the *typical* participant, silicon sampling envisions the *ideal* participant—a prediction algorithm that exhibits human-like behavior but which allows researchers to

control how much the responses randomly vary from one prompt to another, explicitly defining how erratic silicon subjects should behave. It may be unrealistic but advantageous to prompt LLMs to be inhumanely consistent in its responses and strictly abide by the researchers' directions. Proponents of the silicon sampling approach could even argue that unrealistic distributions of LLM-predictions are helpful in estimating parameters. While predictions of human responses exhibit less variability than human responses (Bisbee et al., 2024; Mei et al., 2024), it may be precisely this misrepresentation that allows for more precise measurement of central tendencies such as the mean. This property does not imply that researchers obtain an accurate parameter estimate (e.g. the conditional mean), but that this estimate exhibits less statistical uncertainty than an estimate obtained from a sample of human subjects.

2.2 Perils of Silicon Sampling

Empirical studies question whether silicon subjects alone will be sufficient to draw valid conclusions about human behavior. Predictions of human behavior have been shown to systematically diverge from observed behavior. For example, Atari et al. (2023) find that LLMs respond to various tasks more like those from western, educated, industrialized democracies than those from other parts of the world. Alvero et al. (2024) find that LLMs, when compared to actual college applicants, write college admissions essays most similarly to those who were male and from neighborhoods with high socio-economic status. When researchers are interested in populations or tasks that LLMs are less able to mimic, silicon sampling may lead them astray.

Sources of prediction error are manifold. LLMs may inaccurately predict outcomes of certain groups of individuals because these have been misrepresented or underrepresented in the models' training data (Crockett and Messeri, 2023). While improving the representativeness and overall data quality may enhance prediction accuracy of LLMs, fixing the inputs to LLMs may not be sufficient to rule out error stemming from the prediction algorithm itself. LLMs have been shown to respond differently depending on the order a question or give the same answer consistently (Park et al., 2024). Errors may also arise from the complexity of the research design. While predicting responses to survey questions question may yield a satisfactory level of accuracy, prompting the model to predict responses to experimental stimuli may be too complex. At a more fundamental level, it remains unclear whether LLM predictions can be trusted without validation against human respondents. A divergence between an LLM-derived treatment effect from previous findings could be a statistical fluke or a trend that would replicate with human participants (Harding et al., 2023).

Beyond the risk of drawing incorrect conclusions from inaccurate predictions aspects of conducting mixed subjects experiments, LLMs also pose a more fundamental question of what it means to perform hypothesis tests. The availability of predictions about human behavior at low cost allows researchers to achieve

extreme levels of statistical precision. By generating a large number of predicted values and regressing these on an independent variable, researchers can achieve highly precise estimates simply because standard errors shrink with sample size. It is commonly assumed that gathering more observations is desirable because a larger sample has the double advantage of increasing power to detect a true effect (avoiding false negatives) and decreasing the risk of observing a significant result when there is no effect (avoiding false positives). However, if LLMs inaccurately predict human behavior, the point estimates misrepresent the relationship between Y and X that would be observed in a sample of human subjects. The narrow confidence intervals around the point estimates thus create a false sense of precision. Similar to analyses of Big Data on non-representative samples, researchers risk being "precisely inaccurate" (McFarland and McFarland, 2015). As a result, silicon sampling could further amplify doubts about the replicability of findings from experimental social science (see Freese and Peterson, 2018), not because studies lack sufficient statistical power to discern true effects from false positives, but because they are sufficiently powered to detect *any* effect.

3 The Mixed Subjects Design

We propose the mixed subjects design, an alternative to silicon sampling and an umbrella term for statistical methods that provide valid inferences about human behavior while maintaining the benefits of employing silicon subjects. The mixed subjects approach treats silicon subjects as *potentially* informative of human behavior, relying on the interchangeability assumption to an intermediate degree and in a way that is subject to disconfirmation via empirical evidence. Human respondents count as *ground truth* to correct potentially flawed predictions from LLMs. The goal is to build confidence in LLMs as a research tool by combining human and silicon subjects with statistical methods that produce valid parameter estimates while maintaining the benefits of low costs of LLM predictions and increased statistical power to detect treatment effects. The mixed subjects approach aims to integrate LLM predictions of human behavior into the research process beyond pilot studies and other exploratory types of analyses (e.g. Grossmann et al., 2023). The reduction of costs, coupled with valid inferences about parameters, enables researchers to systematically explore a large number of hypotheses (Almaatouq et al., 2024). In the following, we present prediction-powered inference (Angelopoulos et al., 2023), a recent statistical framework that instantiates the mixed subjects approach.

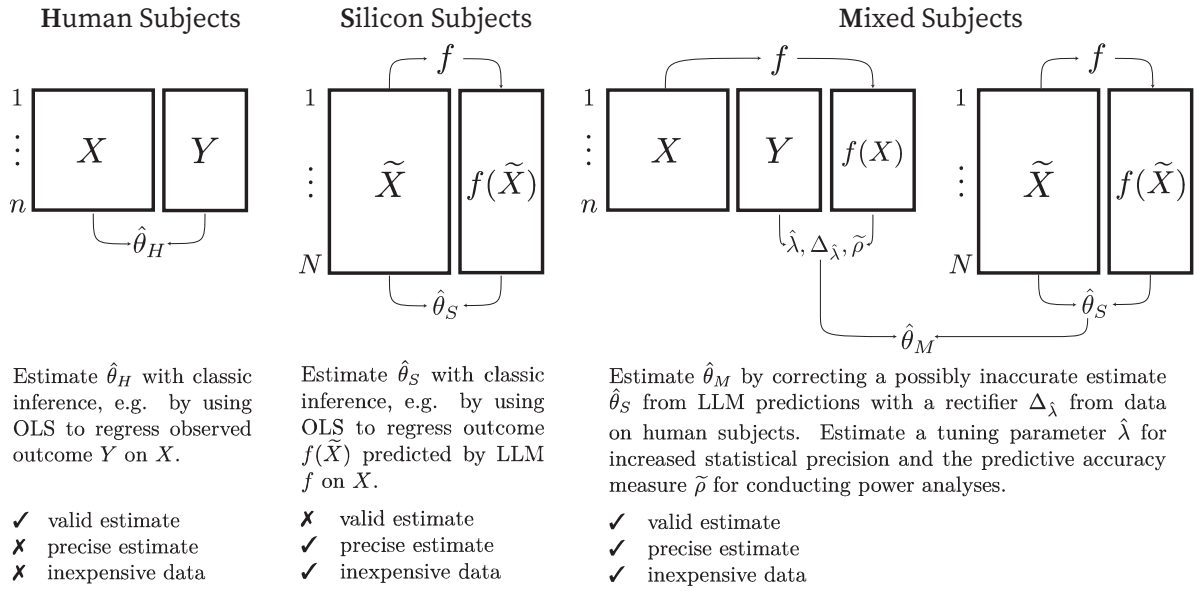

Figure 1: Comparison of experiments with human, silicon, and mixed subjects

3.1 Mixed Subjects Studies with PPI

Prediction-powered inference (PPI) is a statistical method that combines a dataset of "gold-standard" observations with predictions from a machine learning algorithm to estimate a broad class of estimands, including population means, regression coefficients, and quantiles (Angelopoulos et al., 2023, 2024). PPI does not make assumptions about the accuracy of the machine learning algorithm used to predict the dependent variable. Instead, the predictions are treated as informative but imperfect proxies. PPI uses the gold-standard observations to estimate prediction error and adjust parameter estimates accordingly. These corrected estimates are close to those obtained from classical inference with gold-standard observations alone (e.g. a regression coefficient estimated with a sample of responses from human subjects). Yet the PPI estimates are also more precise since increasing sample size with machine learning predictions leads to narrower confidence intervals. In the PPI framework, predictions and gold-standard observations thus complement each other in obtaining valid and precise point estimates.

To explain PPI in more detail we will follow the notation in Angelopoulos et al. (2023, 2024). To estimate a parameter θ, PPI requires three things—a gold-standard (or labeledO dataset $\{(X_i, Y_i)\}_{i=1}^n$, an unlabeled dataset $\{\widetilde{X}_i\}_{i=1}^N$, and a machine learning algorithm f that maps X to a prediction of Y. PPI applies the machine learning algorithm to both datasets—this gives $\{(X_i, Y_i, f(X_i))\}_{i=1}^n$ and $\{(\widetilde{X}_i, f(\widetilde{X}_i))\}_{i=1}^N$. To understand how PPI estimates a parameter θ with these two datasets, it is instructive to consider how PPI estimates the simple population mean, e.g. the average student test score, loan amount, or time spent on social media (Equation 1, see also Figure 1). The estimand $\hat{\theta}^{\text{PP}}$ comprises two

parts: the estimand based on the algorithm's predictions $\hat{\theta}_S$ and a rectifier $\hat{\Delta}_{\hat{\lambda}}$. The rectifier quantifies the difference between the predicted and observed values from the gold-standard dataset and uses this information to adjust the estimate obtained from the prediction algorithm. If the algorithm is very accurate, the rectifier will be close to zero and the estimate is largely based on predictions. To optimize statistical precision, PPI estimates λ, an additional tuning parameter ranging from 0 to 1. $\hat{\lambda} \approx 1$ implies that full weight is given to the predictions whereas $\hat{\lambda} \approx 0$ means that the mean estimate is mostly based on the gold-standard observations.

$$\hat{\theta}^{\text{PP}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \hat{\lambda} f(\widetilde{X}_i)}_{\hat{\lambda}\hat{\theta}_S} - \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\lambda} f(X_i) - \frac{1}{n}\sum_{i=1}^{n} Y_i\right)}_{\hat{\Delta}_{\hat{\lambda}} = \hat{\lambda}\hat{\theta}_{f(X_i)} - \hat{\theta}_H} \tag{1}$$

To apply PPI in a mixed subjects experiment, the human subjects represent the gold standard dataset $\{(X_i, Y_i)\}_{i=1}^n$. The variable X_i encodes the demographic covariates and the treatment assignment of the ith human subject. The variable Y_i is the response from the ith human subject. For the unlabeled dataset $\{\widetilde{X}_i\}_{i=1}^N$ we create N silicon subjects, for example by obtaining a representative sample of the target population (Argyle et al., 2023). For a silicon sample with covariates $\widetilde{X}$, we turn the information in $\widetilde{X}$ into a prompt for an LLM, resulting in a predicted survey response $f(\widetilde{X})$. Prompting the LLM for both the human and silicon samples gives the datasets $\{(X_i, Y_i, f(X_i))\}_{i=1}^n$ and $\{(\widetilde{X}_i, f(\widetilde{X}_i))\}_{i=1}^N$. Researchers can then compute point estimates and confidence intervals with these two datasets with the software provided by Angelopoulos et al. (2023, 2024).

To apply PPI to a mixed subjects experiment, we need to verify a number of assumptions. First, PPI requires the classical assumption that $\{(X_i, Y_i)\}_{i=1}^n$ are independent and identically distributed (i.i.d). In addition, PPI requires that $\{\widetilde{X}_i\}_{i=1}^N$ are i.i.d. and that $\{\widetilde{X}_i\}_{i=1}^N$ are drawn from the same distribution as $\{X_i\}_{i=1}^n$. That is, the hypothetical demographics of the silicon subject population must match the demographics of the human subject population. Likewise, the treatment assignment mechanism must be the same for both groups. Ideally, each silicon subject should correspond to a human subject who would have been surveyed had the sample size n been larger. Second, PPI requires that the training of the machine learning algorithm f is independent of both datasets. This assumption may be violated when the data from human subjects has been previously published and included in the LLM's training data. Finally, the procedure of prompting the LLM must be the same for the gold standard and the partially observed dataset. This means that the same parameters and model should be used on both datasets. Likewise, the method used to turn the demographic and treatment information into a prompt must be the same for both datasets.

3.2 PPI Power Analysis

Power analyses allow researchers to determine the necessary sample size for a desired level of power—i.e., the probability of rejecting the null hypothesis of no effect when there is an effect (Cohen, 1988). A power analysis not only addresses the practical question of how many resources researchers need to invest to find a significant effect but is also instrumental in advancing science. True treatment effects, particularly small ones, may go unnoticed if the sample size is too small. Reporting false negatives impedes researchers in discerning sound from flawed explanations for social phenomena, thwarts the accumulation of knowledge, and may explain the existence of inconsistent findings on core concepts in the social sciences (Thye, 2000; Stadtfeld et al., 2020). Power analyses are therefore crucial for testing and advancing theory. Yet no such method has been developed for PPI. To address this gap, we derive a power analysis, completing the toolkit necessary for conducting mixed subjects experiments.

In section 2.2, we argued that using classical inference methods such as regression to estimate parameters based on large numbers of predictions of human behavior gives a false sense of precision. Even if LLMs inaccurately portray human behavior, point estimates derived from their predictions exhibit too little uncertainty. Therefore, any power analysis that treats LLM predictions not as interchangeable with human subjects must account for how closely LLMs predict observed behavior. While a classical power analysis mainly depends on the effect size and standard error (Rice, 2007: 433-5), we show that a PPI power analysis for mixed subjects study also depends on $\widetilde{\rho} \in [-1, 1]$, a measure of the correlation between the classical estimator $\hat{\theta}_H$ based on human subjects and the estimator $\hat{\theta}_S$ based on the LLM predictions. The number $\widetilde{\rho}$ quantifies the accuracy of LLMs in predicting human responses. As shown in Supporting Information SI 1, the standard error of $\hat{\theta}^{\text{PP}}$ can be written as

$$\text{SE}(\hat{\theta}^{\text{PP}}) = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{N}{N+n} \tilde{\rho}^2}, \tag{2}$$

where $\sigma/\sqrt{n}$ is the standard error of a classical estimator $\hat{\theta}^{\text{classical}}$. As $\tilde{\rho}$ is always between -1 and 1, the standard error of $\hat{\theta}^{\text{PP}}$ is always less than the classical standard error $\sigma/\sqrt{n}$. This implies that a mixed subjects design with PPI will produce narrower confidence intervals and higher statistical power than classical inference. Figure 2 illustrates how achieving this higher statistical precision depends on $\widetilde{\rho}$ and on the ratio N/n. When $\widetilde{\rho}$ is close to 1, the benefit of including LLM predictions is higher. For example, suppose that $N/n = 5$ so that for every human sample there are 5 silicon subjects. If $\tilde{\rho} = 0.5$, then $\sqrt{1 - (N/(N+n))\tilde{\rho}^2} \approx 0.89$ and the PPI standard error will be approximately 11% smaller than the classical standard error. If $\tilde{\rho}$ increased to 0.75 and the same sample sizes were used, the PPI standard error would be approximately 27% smaller than the classic standard error.

The standard error is the key metric for statistical inference on a parameter because it directly influences

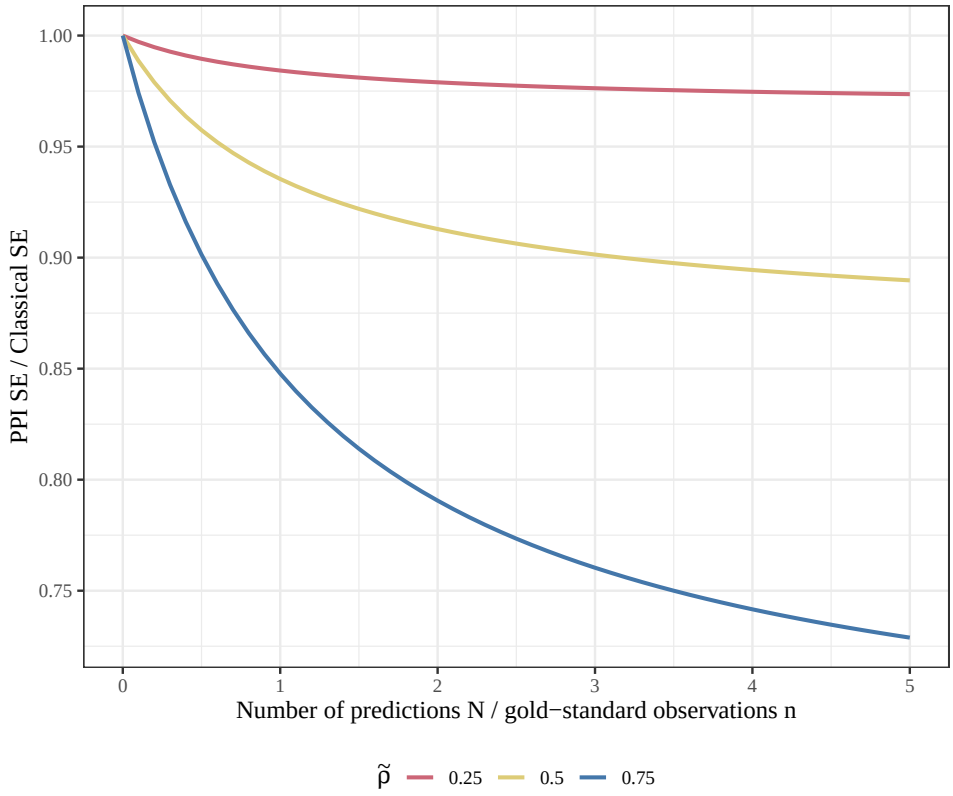

Figure 2: The x-axis shows the ratio N/n of samples sizes with N predictions $f(X_i)$ and n gold standard observations Y_i. The y-axis shows the ratio of the PPI standard error to the classical standard error, defined by $\sqrt{1-(N/(N+n))\tilde{\rho}^2}$. The PPI standard error becomes narrower than the classical standard error for larger N relative to n, and this increase in statistical precision is greatest for stronger associations $\tilde{\rho}$ between the predicted values $f(X_i)$ and observed Y_i. This pattern also applies to confidence intervals since the ratio of standard errors shown is equivalent to the ratio of the width of confidence intervals.

the width of confidence intervals

$$\text{CI}_\alpha^{\text{PP}} = \hat{\theta}^{\text{PP}} \pm z_{1-\alpha/2}\widehat{\text{SE}}(\hat{\theta}^{\text{PP}}), \tag{3}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and $\widehat{\text{SE}}(\hat{\theta}^{\text{PP}})$ is an estimate of the standard error from equation (2). Moreover, the standard error determines the statistical power for a hypothesis test since

$$\text{Power} = 1 - \Phi\left(z_{1-\alpha/2} - \frac{\Delta}{\text{SE}(\hat{\theta}^{\text{PP}})}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\Delta}{\text{SE}(\hat{\theta}^{\text{PP}})}\right) \tag{4}$$

where $\Delta = \theta_{H_a} - \theta_{H_0}$ is the hypothesized effect size and Φ is the cumulative standard normal distribution. Equations (3) and (4) show that decreasing the standard error is the most immediate way to narrow the width of confidence intervals and to increase statistical power because the other quantities are constants (i.e. the quantiles) or mostly beyond the control of the researcher (i.e. the effect size). The PPI

standard error directly depends on $\widetilde{\rho}$, the ratio of silicon subjects to the total sample size $N/(N+n)$, and the standard error of the classical estimator $\sigma/\sqrt{n}$. Researchers could increase $\widetilde{\rho}$ by choosing more accurate prediction algorithms, such as using LLMs trained with more parameters on higher-quality data or by fine-tuning a model for the prediction task. Prompt engineering—such as providing more context, examples, or specific instructions—may also enhance prediction accuracy. As shown in Figure 2, including more silicon subjects is most effective in reducing standard errors when $\widetilde{\rho}$ is closer to 1. Hence, researchers can maximize returns on predictions by using algorithms that accurately portray human behavior. If no such informative algorithms exist, researchers also could increase precision by recruiting more human subjects like in classical inference. Yet there is a trade-off. Obtaining silicon samples is much more affordable than recruiting human subjects. However, silicon subjects are not as informative as human subjects for estimating a parameter. There is an optimal choice for combining a sample size of a human subjects n with N silicon subjects, and this combination depends on the prediction accuracy $\widetilde{\rho}$, a hypothesized effect size, a desired level of power, the cost of recruiting human and silicon subjects, and the research budget. Given these parameters, our power analysis allows researchers to optimally decide between recruiting *costly but informative* human subjects or *cheap but less informative* silicon subjects. This multidimensional choice problem can be solved with constraint optimization (Apostol, 1969), allowing researchers to answer the following two questions.

First, which pair of sample sizes (n, N) yields the highest power given a fixed research budget B, the ratio of the cost of silicon subjects to human subjects γ, and the accuracy of LLMs in predicting human behavior $\widetilde{\rho}$? Researchers may be particularly interested in finding the most powerful pair if a limited research budget is the main constraint and resources should be allocated most effectively to maximize power. Figure 3a illustrates this optimization problem. Finding the most powerful pair involves selecting combinations of n and N that satisfy the budget constraint and identifying the point where statistical power is highest. Second, which pair of sample sizes N, n is the cheapest to sample given a desired level of power, the cost ratio γ, and predictive accuracy $\widetilde{\rho}$? Researchers might be more interested in this question if budget constraints are less salient but resource allocation should still be as efficient as possible. Identifying the cheapest pair means selecting the combination (n, N) that gives a desired level of power and identifying the point where the cost is lowest (Figure 3b).

PPI produces smaller narrower confidence intervals and higher statistical power than classical inference. Whether PPI is also more cost-effective depends on the accuracy of predictions $\widetilde{\rho}$ and the ratio of the costs of surveying silicon and human subjects γ—i.e. the cost of surveying a silicon subject divided by the cost of surveying a human subject. Specifically, PPI is more cost-effective than classic inference with human subjects if and only if

$$\widetilde{\rho} > \frac{2\sqrt{\gamma}}{1+\gamma}. \tag{5}$$

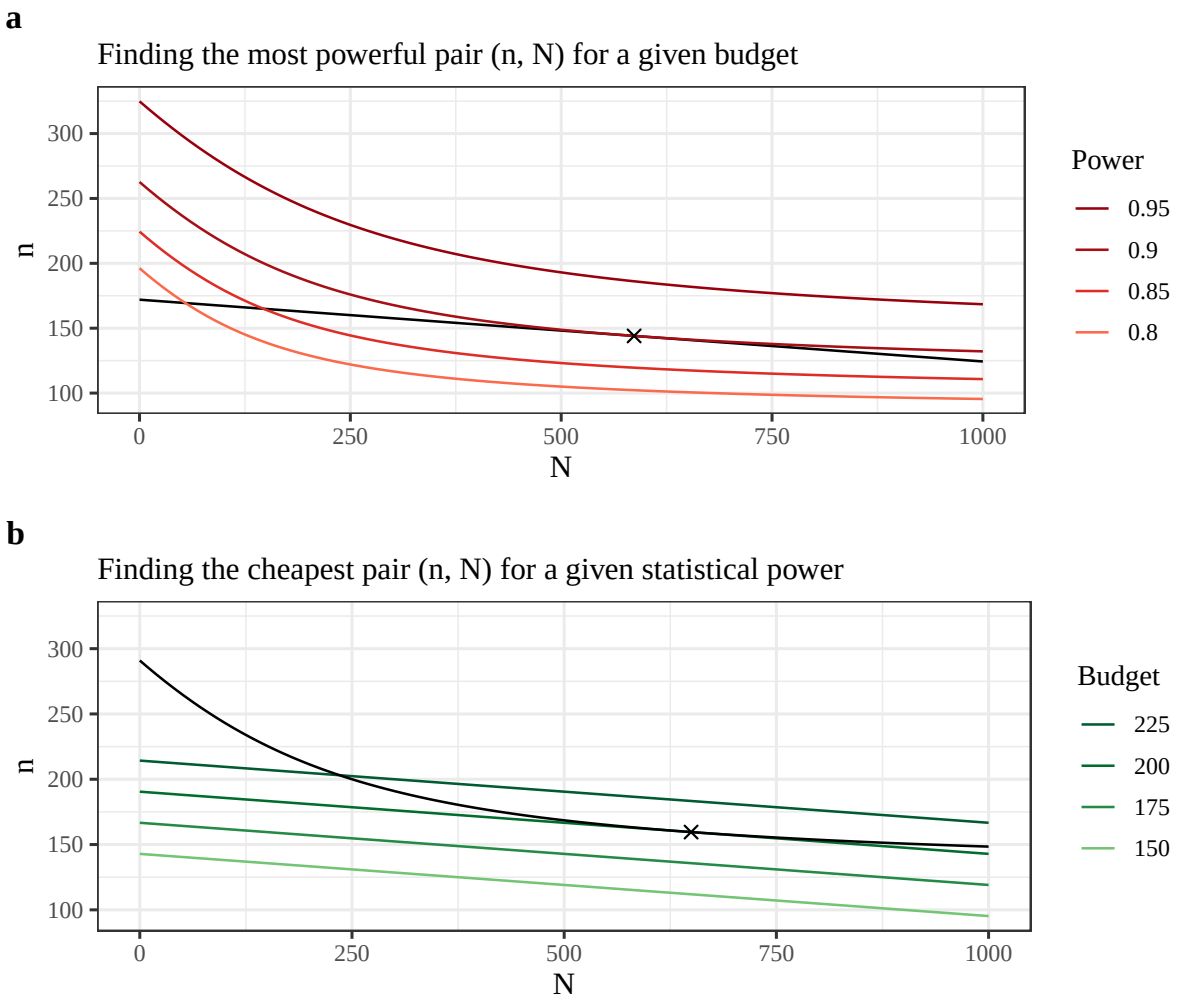

Figure 3: Illustration of the constraint optimization with $\widetilde{\rho} = 0.75$, effect size $\Delta = 0.2$, classical standard error $\hat{\sigma}/\sqrt{n} = 1$, and a ratio of cost of sampling silicon subjects to human subjects of $\gamma = 0.05$ (**a**) Given a fixed budget, the power analysis identifies the combination of sample sizes (n, N) that with the highest statistical power. (**b**) Given a desired level of statistical power for detecting an effect, the power analysis identifies the combination of sample sizes (n, N) that minimizes budget expenditure.

For example, we could assume a cost of \$3.20 for a participant to fill out a 12 minute survey. Based on the costs of \$0.003 for prompting LLMs for this study, we calculated that any $\widetilde{\rho} > 0.06$ would be sufficient for PPI to save costs when compared to classical inference with human subjects only (see SI 1 for details). More generally, if condition (5) is satisfied, the percentage cost savings of using PPI with the optimal sample size is given by $\widetilde{\rho}^2(1-\gamma) - 2\sqrt{\gamma\widetilde{\rho}^2(1-\widetilde{\rho}^2)}$. Figure 4 shows that larger values for $\widetilde{\rho}$ imply more savings, and that $\widetilde{\rho}$ has to be larger if obtaining silicon subjects is more expensive relative to recruiting human subjects.

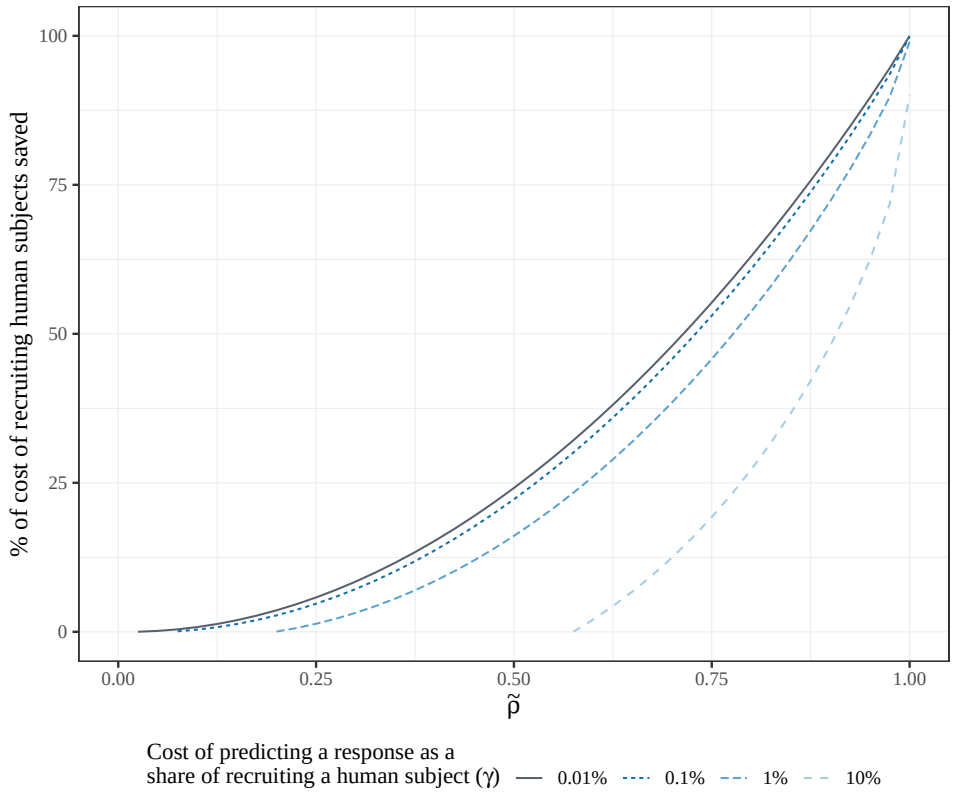

Figure 4: Percent of cost of recruiting human subjects saved when using PPI for different levels of $\widetilde{\rho}$ and ratio of costs γ

Statistical software published alongside this article offers user-friendly tools for conducting power analyses in mixed subjects studies with PPI. This power analysis is *data-driven* in that $\widetilde{\rho}$ needs to be estimated, e.g. from a pilot study. The software takes into account $\widetilde{\rho}$, a hypothesized effect size Δ, a desired level of power $1-\beta$, a research budget B, and the cost of recruiting human subjects c_Y, the cost of generating silicon subjects c_X, and the cost of running the prediction algorithm c_f (see SI 1 for details).

4 Application to the Moral Machine Experiment

The Moral Machine Experiment (Awad et al., 2018) sought to better understand the factors that influence people's decisions when faced with moral dilemmas that self-driving cars might encounter on the road.

In this conjoint experiment, participants were presented with hypothetical scenarios involving harm to either passengers or pedestrians due to a sudden break failure in the vehicle. Participants could only save one group, pedestrians or passengers, considering factors such as age, gender, social status, and the number of individuals involved (Figure 6 in SI 2). We chose the Moral Machine Experiment to demonstrate the mixed subjects design because we believe that LLMs will be particularly valuable in data-intensive contexts (e.g. multi-condition experiments). In these settings, the cost of conducting experiments solely with human subjects may be prohibitive since accurately estimating effects requires large numbers of observations.

4.1 Methods

We exemplify a mixed-subject study with a representative sample of 2,097 Americans who evaluated a total of $22,315$ moral dilemmas. This sample was drawn from the participants of the Moral Machine experiment with quotas on age, education, gender, and income from the 2016 American Community Survey (Ruggles et al., 2024). Our sample closely resembles the American population, except for older individuals, who could not be sufficiently sampled due to their minimal presence in the Moral Machine experiment (Figure 7 in SI 2).

Next, we used LLMs to predict the decisions of these survey respondents to the moral dilemmas. We prompted GPT4 Turbo (gpt-4-turbo), GPT4o (gpt-4o), and GPT3.5 Turbo (gpt-3.5-turbo-0125) to evaluate a moral dilemma based on text describing of the two possible outcomes (see Table 1 in SI for details). We created these descriptions from the data recording the experimental conditions that respondents were exposed to. We automatically generated the prompts describing the dilemmas to the LLMs by adjusting code from a related study (Takemoto, 2024).

To evaluate the causal impact of the presence or absence of specific scenario features such as young versus old, we estimate the Average Marginal Component Effect (AMCE). This estimand is implemented as a weighted simple linear regression (Hainmueller et al., 2014), and can be estimated with PPI. Our main interest in this empirical study is to assess whether (i) PPI yields accurate point estimates of the AMCE, (ii) the extent to which including LLM predictions increases statistical precision—i.e. narrows confidence intervals, and (iii) compare these results against estimates derived from human and LLM-predictions alone.

4.2 Results

Prompting LLMs to predict 22,315 survey responses resulted in modest correlations, ranging from $r = 0.36$ for GPT4 Turbo to $r = 0.11$ for GPT3.5 Turbo. We conducted two supplemental analyses to assess the possibility of increasing predictive accuracy (see Table 2 in SI 2 for details). First, we prompted

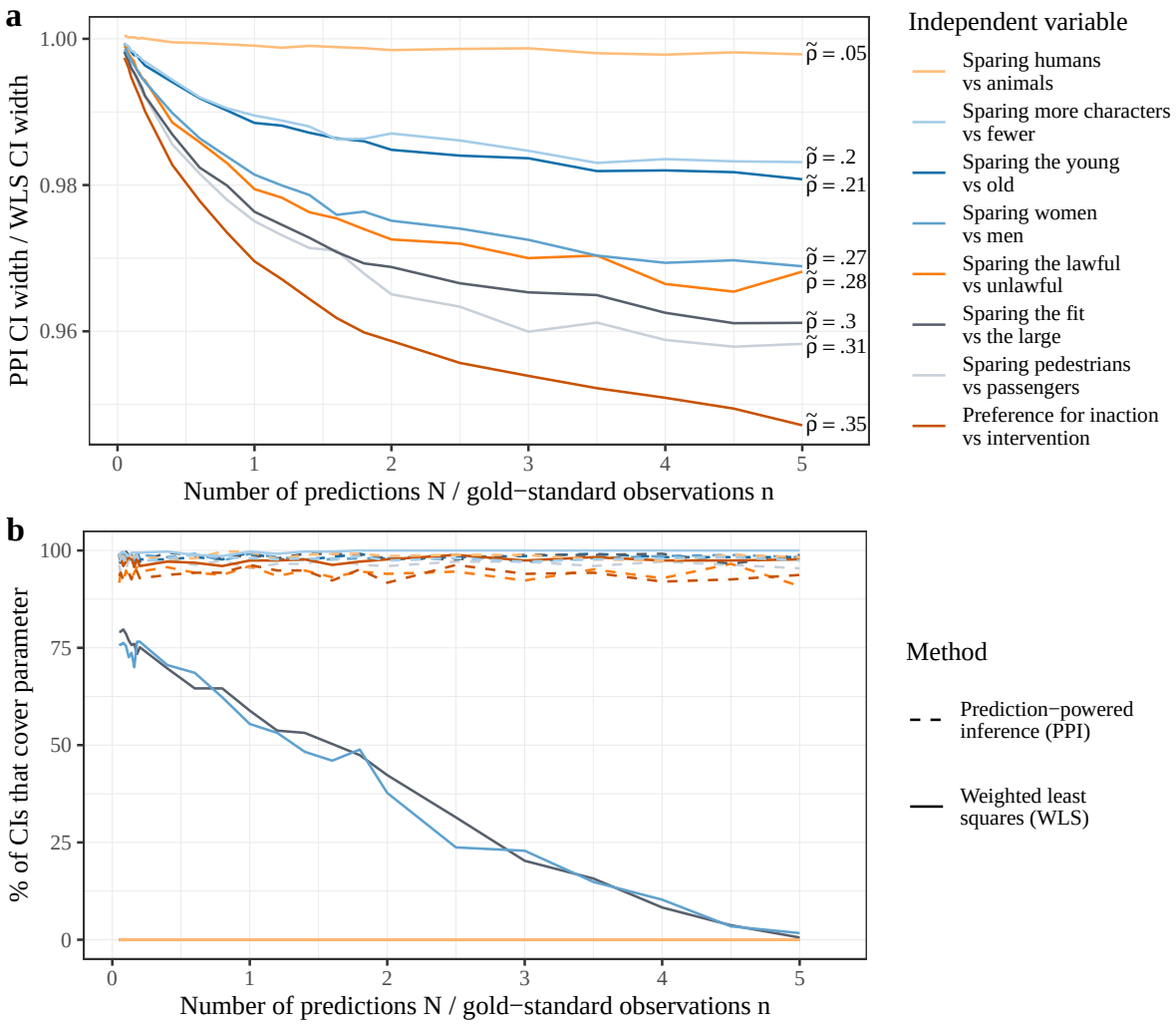

Figure 5: The x-axes show the ratio of samples sizes N/n with N predictions $f(X_i)$ and n gold standard observations Y_i. (**a**) The y-axis shows the width of the prediction-powered inference (PPI) confidence interval (CI) as a share of the CI from weighted linear regression (WLS), with values smaller than 1 showing the percent reduction in PPI CI width relative to WLS CI (cf. Figure 2). (**b**) The y-axis shows the percent of CIs calculated with PPI and WLS that cover the AMCE estimate from the representative sample of 2,097 Americans; the estimate from this sample is used as the best available approximation of the AMCE parameter.

each of the LLMs twice to give 5,000 additional predictions for identical prompts. We then created a composite with the mode of three predictions. If anything, taking the modal prediction only minimally increases the correlation. Second, for a separate set of 5,000 predictions, we omit the demographic persona from the prompt. Excluding the persona minimally decreases the correlation, except for GPT3.5 Turbo where the correlation increases from $r = 0.11$ to $r = 0.17$. Overall, the predictions yield very similar correlations across these cases. All analyses reported in this section are based on the 22,315 predicted survey responses by GPT4 Turbo (see SI 2 for the prompt).

Figure 5 compares statistical precision (i.e., width of confidence intervals) and validity (i.e. percent of confidence intervals that cover the true parameter) in the mixed subjects approach with PPI versus the silicon subjects approach for sample size $n = (50, 100, 500)$ and predictions $N = (25, ..., 2500)$. For each of these combinations of sample sizes (n, N), we drew 350 repeated random samples and calculated the mean width and coverage of the confidence intervals. Figure 5a shows that adding an increasing number of LLM predictions reduces the width of confidence intervals more strongly for larger values of $\widetilde{\rho}$. Figure 5b shows that the percent of PPI confidence intervals that cover the best parameter estimate remains stable at high level levels. In contrast, the coverage of the classical intervals computed on the pooled sample cover the true parameter remains stable only for some independent variables. In sum, this analysis illustrates that PPI that produces valid point estimates, with increases in statistical precision depending on the accuracy of the LLM in predicting human decisions. The silicon sampling approach may also produce valid point estimates but this is impossible ascertain without validation on human subjects. PPI automatically handles this validation by introduction a statistical correction to the predictions from LLMs.

5 Conclusion

Large language models, and generative AI more generally, may transform the study of human behavior. However, like novel data sources that have come before (Lazer et al., 2009), computational social scientists must carefully consider the limitations of LLMs and develop methods to ensure sound conclusions from this new data source. Researchers voiced concerns about recently developed approaches that treat LLM as interchangeable with human behavior. In some cases, the scientific costs of accepting this assumption may be worth the practical benefits gained—in other cases, the trade-off may not be clear. We propose an integrated approach to incorporating LLM simulations in social science research, the mixed subjects approach, which provides compromise solutions and expands the set of trade-offs researchers can make. LLM simulations should be integrated with, rather than replace, human subjects in a hybrid study design. Our approach maintains scientific credibility while leveraging information contained within cost-effective LLM-based simulations.

We propose implementing the mixed subjects design with PPI, a statistical method that uses data from human subjects to correct possibly inaccurate predictions from LLMs and other prediction algorithms. By combining these two data sources, PPI allows researchers to obtain valid point estimates at lower cost than classical inference applied to data from human subjects only. Additionally, PPI produces narrower confidence intervals for parameters and higher statistical power than classical inference since PPI standard errors will always be smaller than classic standard errors. To complete the toolkit necessary to conduct mixed subjects studies, we derive a power analysis for PPI. Using constraint optimization, the power analysis allows researchers to optimally choose between recruiting *informative but costly human subjects* and *less informative but cheap silicon subjects*.

We believe that mixed subject studies with LLMs and human subjects will be most useful when researchers operate under budget constraints, for example when systematically assessing a large set of hypotheses such as the Moral Machine Experiment. The cost of recruiting sufficient human subjects may prevent researchers from obtaining precise estimates or conducting such experiments altogether. By reducing the cost of data collection, coupled with valid inferences about parameters, the mixed subjects design could increase scientific productivity and reduce inequality in access to otherwise costly evidence on research questions and hypotheses. Freed-up financial resources could also be allocated to other research projects or used to pay higher wages to survey participants.

References

Almaatouq, Abdullah, Thomas L. Griffiths, Jordan W. Suchow, Mark E. Whiting, James Evans, and Duncan J. Watts. 2024. "Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences." *Behavioral and Brain Sciences* 47:e33.

Alvero, AJ, Jinsook Lee, Alejandra Regla-Vargas, Rene Kizilec, Thorsten Joachims, and Anthony Lising Antonio. 2024. "Large Language Models, Social Demography, and Hegemony: Comparing Authorship in Human and Synthetic Text." *Preprint* pp. 1–25.

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. "Prediction-powered inference." *Science* 382:669–674. Publisher: American Association for the Advancement of Science.

Angelopoulos, Anastasios N., John C. Duchi, and Tijana Zrnic. 2024. "PPI++: Efficient Prediction-Powered Inference."

Apostol, Tom. 1969. *Calculus Vol. II*. New York: Wiley & Sons, 2nd edition.

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

Atari, Mohammad, Mona J Xue, Peter S Park, Damián E Blasi, and Joseph Henrich. 2023. "Which Humans?"

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. "The Moral Machine experiment." *Nature* 563:59–64.

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

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis* p. 1–16.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, N.J.: L. Erlbaum Associates, 2nd edition.

Crockett, Molly and Lisa Messeri. 2023. "Should large language models replace human participants?"

Dillion, Danica, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. "Can AI language models replace human participants?" *Trends in Cognitive Sciences* .

Freese, Jeremy and David Peterson. 2018. "The Emergence of Statistical Objectivity: Changing Ideas of Epistemic Vice and Virtue in Science." *Sociological Theory* 36.

Friedman, Milton. 1953. "The Methodology of Positive Economics." In *Essays in Positive Economics*, pp. 3–43. University of Chicago Press.

Grossmann, Igor, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. 2023. "AI and the transformation of social science research." *Science* 380:1108–1109. Publisher: American Association for the Advancement of Science.

Gui, George and Olivier Toubia. 2023. "The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective." *arXiv preprint arXiv:2312.15524* .

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22:1–30.

Harding, Jacqueline, William D'Alessandro, NG Laskowski, and Robert Long. 2023. "AI language models cannot replace human research participants." *Ai & Society* pp. 1–3.

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. 2009. "Computational Social Science." *Science* 323:721–723.

McFarland, Daniel A and H Richard McFarland. 2015. "Big Data and the danger of being precisely inaccurate." *Big Data & Society* 2:2053951715602495.

Mei, Qiaozhu, Yutong Xie, Walter Yuan, and Matthew O. Jackson. 2024. "A Turing test of whether AI chatbots are behaviorally similar to humans." *Proceedings of the National Academy of Sciences* 121:e2313925121.

Park, Peter S., Philipp Schoenegger, and Chongyang Zhu. 2024. "Diminished diversity-of-thought in a standard large language model." *Behavior Research Methods* .

Rice, John A. 2007. *Mathematical statistics and data analysis*. Duxbury advanced series. Belmont, CA: Thomson/Brooks/Cole, 3rd edition.

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. 2024. "2012-2016 American Community Survey." IPUMS USA 10.18128/D010.V15.0 (Accessed June 7, 2024).

Stadtfeld, Christoph, Tom A. B. Snijders, Christian Steglich, and Marijtje van Duijn. 2020. "Statistical Power in Longitudinal Network Studies." *Sociological Methods & Research* 49:1103–1132.

Takemoto, Kazuhiro. 2024. "The moral machine experiment on large language models." *Royal Society Open Science* 11:231393.

Thye, Shane R. 2000. "Reliability in Experimental Sociology." *Social Forces* 78:1277–1309.

Supporting Information (SI)

1 PPI Power Analysis

Let $\{(X_i, Y_i)\}_{i=1}^n$ and $\{\widetilde{X}_i\}_{i=1}^N$ be the labeled and unlabeled datasets as in Section 3.1. Let f be the machine learning algorithm that predicts Y from X. Finally, let $\ell_\theta(x, y)$ be a loss function with $\theta \in \mathbb{R}^d$. The loss function ℓ_θ defines an estimand $\theta^\star$ by

$$\theta^\star = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathbb{E}[\ell_\theta(X, Y)].$$

For example, if $\ell_\theta(x, y) = \frac{1}{2}(x^\top \theta - y)^2$, then $\theta^\star$ is the vector of ordinary least squares coefficients for regressing Y on X.

Angelopoulos et al. (2024) introduce a family of estimators $\hat{\theta}^{\text{PP}}_\lambda$ for $\theta^\star$. These estimators depend on a tuning parameter $\lambda \in \mathbb{R}$ and are given by

$$\hat{\theta}^{\text{PP}}_\lambda = \underset{\theta \in \mathbb{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^n \ell_\theta(X_i, Y_i) + \lambda \left(\frac{1}{N} \sum_{i=1}^N \ell_\theta(\widetilde{X}_i, f(\widetilde{X}_i)) - \frac{1}{n} \sum_{i=1}^n \ell_\theta(X_i, f(X_i)) \right).$$

Taking $\lambda = 0$ corresponds to the classical M-estimator for $\theta^\star$

$$\hat{\theta}^{\text{classical}} = \hat{\theta}^{\text{PP}}_0 = \underset{\theta \in \mathbb{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^n \ell_\theta(X_i, Y_i).$$

In Theorem 1 of Angelopoulos et al. (2024), the authors show that, under certain assumptions of the loss ℓ_θ, the PPI estimator $\hat{\theta}^{\text{PP}}_\lambda$ satisfies a central limit theorem. Specifically, if $n, N \to \infty$ with $n/N \to r$, then

$$\sqrt{n} \left(\hat{\theta}^{\text{PP}}_\lambda - \theta^\star \right) \xrightarrow{d} \mathcal{N}(0, \Sigma^\lambda), \tag{6}$$

where $\mathcal{N}(\mu, \Sigma)$ denoted the d-dimensional Guassian distribution with mean μ and covariance matrix Σ. The asymptotic covariance matrix Σ^λ has the following "sandwich" form

$$\begin{aligned} \Sigma^\lambda =& H_{\theta^\star}^{-1} \operatorname{Cov}(\nabla \ell_{\theta^\star}) H_{\theta^\star}^{-1} + \lambda^2 (1 + r) H_{\theta^\star}^{-1} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}) H_{\theta^\star}^{-1} \\ & - \lambda H_{\theta^\star}^{-1} \left(\operatorname{Cov}(\nabla \ell^f_{\theta^\star}, \nabla \ell_{\theta^\star}) + \operatorname{Cov}(\nabla \ell_{\theta^\star}, \nabla \ell^f_{\theta^\star}) \right) H_{\theta^\star}^{-1}, \end{aligned} \tag{7}$$

where $\nabla \ell_{\theta^\star}$ is the gradient of $\ell_\theta(X, Y)$ with respect to θ evaluated at $\theta^\star$. Likewise $\nabla \ell^f_{\theta^\star}$ is the gradient of $\ell_\theta(X, f(X))$ evaluated at $\theta^\star$ and $H^\star_\theta = \mathbb{E}[\nabla^2 \ell_{\theta^\star}(X, Y)]$.

The central limit theorem in (6) is used to compute PPI confidence intervals. The confidence interval for the jth coordinate of $\theta^\star$ is given by

$$\text{CI}^{\text{PP}}_{\alpha,\lambda}(j) = \hat{\theta}^{\text{PP}}_{\lambda,j} \pm z_{1-\alpha/2} \sqrt{\Sigma^\lambda_{j,j}/n}.$$

The parameter λ can be chosen by power tuning (Angelopoulos et al., 2024, Section 6). Power tuning chooses the value of λ that minimizes the expected width of the above confidence interval. This is

equivalent to choosing λ to minimize $\Sigma^{\lambda}_{j,j}$. From (7), we have

$$\begin{aligned}\Sigma^{\lambda}_{j,j} &= \left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j} + \lambda^2(1+r)\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j} \\ &\quad - \lambda\left[H^{-1}_{\theta^\star}\left(\operatorname{Cov}(\nabla \ell^f_{\theta^\star}, \nabla \ell_{\theta^\star}) + \operatorname{Cov}(\nabla \ell_{\theta^\star}, \nabla \ell^f_{\theta^\star})\right) H^{-1}_{\theta^\star}\right]_{j,j} \\ &= \left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j} + \lambda^2(1+r)\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j} \\ &\quad - 2\lambda\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}, \nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j}.\end{aligned}$$

To get the final expression, we have used that $\operatorname{Cov}(\nabla \ell^f_{\theta^\star}, \nabla \ell_{\theta^\star})^\top = \operatorname{Cov}(\nabla \ell_{\theta^\star}, \nabla \ell^f_{\theta^\star})$ and that $H^{-1}_{\theta^*}$ is symmetric. The function $\lambda \mapsto \Sigma^{\lambda}_{j,j}$ is quadratic in λ and its minimum occurs at

$$\lambda^\star_j = \frac{1}{1+r} \frac{\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}, \nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j}}{\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j}}.$$

Furthermore, when $\lambda = \lambda^\star_j$, we have

$$\begin{aligned}\Sigma^{\lambda^\star_j}_{j,j} &= \left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j} \\ &\quad \times \left(1 - \frac{1}{1+r} \frac{\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}, \nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]^2_{j,j}}{\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j} \left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j}}\right) \\ &= \sigma^2_j \left(1 - \frac{1}{1+r} \tilde{\rho}^2_j\right), \end{aligned} \tag{8}$$

where we have defined $\sigma^2_j = \left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j}$ and

$$\tilde{\rho}^2_j = \frac{\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}, \nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]^2_{j,j}}{\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j} \left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j}}.$$

The quantity σ^2_j is the asymptotic variance of the classical estimator $\hat{\theta}^{\text{classical}}$ and $\tilde{\rho}_j$ measures the correlation between $\nabla \ell^f_{\theta^\star,j}$ and $\nabla \ell_{\theta^\star,j}$. That is, $\tilde{\rho}_j$ relates to the correlation between $\hat{\theta}^{\text{classical}}$ and an estimator based on replacing Y with $f(X)$.

In a sample of n labeled data points and N unlabeled data points, the standard error of $\hat{\theta^{\text{PP}}}_{\lambda^\star_j,j}$ is

$$\sqrt{\Sigma^{\lambda^\star_j}_{j,j}/n} = \frac{\sigma_j}{\sqrt{n}} \sqrt{1 - \frac{N}{N+n} \tilde{\rho}^2_j}.$$

In practice, $\lambda^\star_j$ has to be estimated. Angelopoulos et al. (2024) provide a consistent estimator $\hat{\lambda}_j$ for $\lambda^\star_j$ and show that $\hat{\theta}^{\text{PP}}_{\hat{\lambda}_j,j}$ achieves the same asymptotic variance as $\hat{\theta}^{\text{PP}}_{\lambda^\star_j,j}$. We will write $\hat{\theta}^{\text{PP}}$ for $\hat{\theta}^{\text{PP}}_{\hat{\lambda}_j,j}$. Likewise we will write σ instead of σ_j and $\widetilde{\rho}$ instead of $\widetilde{\rho}_j$. With this notation we have

$$\text{SE}(\hat{\theta}^{\text{PP}}) = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{N}{N+n} \widetilde{\rho}^2},$$

as claimed in equation (2).

PPI can achieve the same level of statistical precision as classical inference while using fewer observations from human subjects (Angelopoulos et al., 2024). Below we quantify this increase in precision and identify the conditions under which this benefit is greatest.

Let $S(n, N)$ be the PPI standard error for the estimate $\hat{\theta}^{\text{PP}}$ when using n human samples and N silicon samples. That is,

$$S(n, N) = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{N}{n+N} \widetilde{\rho}^2}.$$

To quantify the increase in statistical precision in PPI relative to classical inference, we let n_0 be a baseline sample size of an experiment with only human subjects. The standard error of using n_0 human subjects and no silicon subject is

$$S(n_0, 0) = \frac{\sigma}{\sqrt{n_0}}.$$

For any n satisfying $n_0(1 - \widetilde{\rho}^2) < n \leq n_0$, if $N = \frac{n(n-n_0)}{n - n_0(1-\widetilde{\rho}^2)}$, then

$$S(n, N) = S(n_0, 0).$$

Equivalently, the level curves of the function $S(n, N)$ are given by

$$N = \frac{n(n_0 - n)}{n - n_0(1 - \widetilde{\rho}^2)}, \quad n_0(1 - \widetilde{\rho}^2) < n \leq n_0. \tag{9}$$

There are three distinct costs involved in performing PPI: the cost c_X of collecting an unlabeled sample X, the cost c_Y of collecting the label Y, and the cost c_f of running the prediction algorithm to compute $f(X)$. For a pair of sample sizes (n, N), the total cost is

$$c(n, N) = c_X(n + N) + c_Y n + c_f(n + N) = c_Y n + (c_X + c_f)(n + N).$$

In a mixed subject experiment, we will assume that $c_X = 0$. That is, the cost of human subject is exactly equal to c_Y and the cost of prompting the LLM is c_f. Typically, we will also have $c_f \ll c_Y$. That is, the cost of collecting human responses Y far outweighs the cost of generating a silicon sample $f(X)$.

Let $\gamma = \frac{c_f}{c_Y}$ be the ratio of the cost of a silicon subjects to the cost of a human subjects. Then, the cost of performing PPI in a mixed subject experiment is

$$c(n, N) = c_Y \left(n + \gamma(n + N)\right). \tag{10}$$

This is because we must prompt the LLM for every human subject and every silicon subject.

By substituting (9) into (10), we can minimize over n to identify the most cost-effective way of using PPI to obtain a point estimate with a standard error of size $\sigma/\sqrt{n_0}$. i.e., a level statistical precision on par with classical inference with n_0 human subjects.

The minimum cost occurs at

$$n^\star = n_0 \left(1 - \widetilde{\rho}^2 + \sqrt{\gamma \widetilde{\rho}^2 (1 - \widetilde{\rho}^2)}\right), \quad N^\star = n^\star \cdot \frac{n_0 - n^\star}{n^\star - (1 - \widetilde{\rho}^2) n_0}. \tag{11}$$

The minimal cost of using the sample size $(n^\star, N^\star)$ is

$$c(n^\star, N^\star) = c_Y n_0 \left(1 - \widetilde{\rho}^2 + \gamma \widetilde{\rho}^2 + 2\sqrt{\gamma \widetilde{\rho}^2 (1 - \widetilde{\rho}^2)}\right). \tag{12}$$

The cost of performing classical inference is simply $c_Y n_0$. Thus, PPI is more cost-effective than classical inference when $\widetilde{\rho}^2 - \gamma\widetilde{\rho}^2 - 2\sqrt{\gamma\widetilde{\rho}^2(1-\widetilde{\rho}^2)} > 0$. This is equivalent to

$$\widetilde{\rho} > \frac{2\sqrt{\gamma}}{1+\gamma}.$$

That is, the correlation $\widetilde{\rho}$ must be larger than $2\sqrt{\gamma}/(1+\gamma)$ to improve precision over classic inference. Beyond this minimal condition, we can quantify the cost reduction when using PPI. The *absolute* cost reduction of PPI at the optimal budget $(n^\star, N^\star)$ compared to classical inference are

$$c_Y n_0 - c(n^\star, N^\star) = c_Y n_0 \left(\widetilde{\rho}^2 - \gamma\widetilde{\rho}^2 - 2\sqrt{\gamma\widetilde{\rho}^2(1-\widetilde{\rho}^2)}\right).$$

The *percent* cost reduction of PPI using the optimal sample sizes is thus

$$\frac{c_Y n_0 - c(n^\star, N^\star)}{c(n^\star, N^\star)} = \widetilde{\rho}^2 - \gamma\widetilde{\rho}^2 - 2\sqrt{\gamma\widetilde{\rho}^2(1-\widetilde{\rho}^2)}.$$

For example, we could take $c_Y = \$3.20$ which corresponds to paying respondents a California minimum wage for filling out a 12 minute survey. If we let $c_f = 0.003$ be the cost of simulating a survey session, we have $\gamma = \frac{0.003}{3.20} = 0.0009375$. Thus, for PPI to outperform classical inference, it is necessary that

$$\widetilde{\rho} > \frac{2\sqrt{0.0009375}}{1+0.0009375} \approx 0.061.$$

In the moral machine dataset, the largest value of $\widetilde{\rho}$ was 0.35. For this value, $\widetilde{\rho}$, the savings are approximately $0.102 \cdot c_Y n_0$, so that PPI can save 10.2% of the costs compared to classical inference.

Equation (11) can be used to find the most powerful pair $(n^\star, N^\star)$ subject to a budget constraint $c(n, N) \leq B$. This is done by setting $c(n^\star, N^\star) = B$ where $c(n^\star, N^\star)$ is as in equation (12). This gives $n_0 = \frac{B}{c_Y(1-\tilde{\rho}^2+\gamma\tilde{\rho}^2+2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)})}$. This value of n_0 can then be plugged into equation (11). Likewise, equation (11) can be used to choose the pair (n, N) that minimizes $c(n, N)$ subject to PPI having power $1-\beta$. If n_0 is the minimum sample size needed for classical inference to have power $1-\beta$, then equation (11) gives a pair $(n^\star, N^\star)$ with a lower cost and the same level of power.

2 Details on the Moral Machine Experiment

2.1 Example of a Moral Dilemma

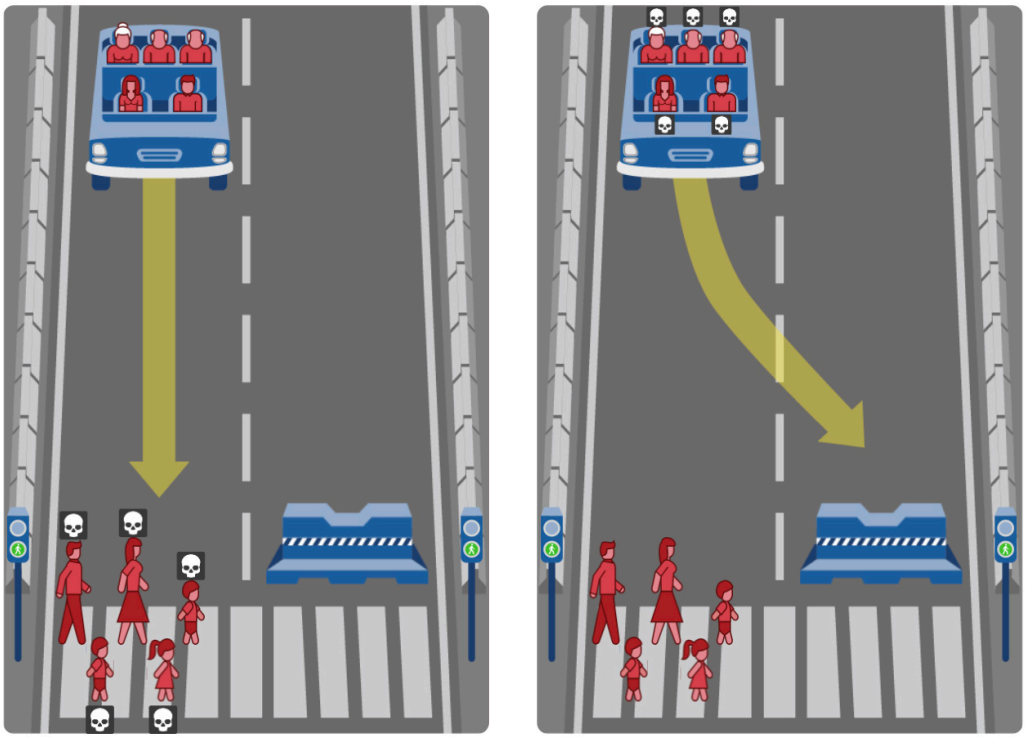

Figure 6: Example dilemma presented to a survey respondent in the Moral Machine Experiment. The image was created on the *moralmachine.net* website with information on the dilemma *2224g4ytARX4QT5rB* from the replication data for Awad et al. (2018).

2.2 Example prompt

The LLM was given the following prompt to predict the survey respondent's decision to dilemma *2224g4ytARX4QT5rB*. The order of the demographic information was randomized.

Carefully consider the following information and think step by step. You are an American with the following characteristics:

- You earn an annual income of under $5,000 US dollars.
- Your highest level of education is a high school diploma.
- You are 35 years old.
- You are a man.
- On a scale from 0 (Not Religious) to 100 (Very Religious), you report a score of 0 for your religious views.
- On a scale from 0 (Conservative) to 100 (Progressive), you report a score of 100 for your political views.

Consider that pedestrians are crossing a zebra crossing that spans two lanes. Please indicate which case is better for autonomous driving by choosing either 'Case 1' or 'Case 2'. Respond with one word only.

Case 1. A self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in the death of 1 man, 1 girl, 2 boys, and 1 woman, who were abiding by the law by crossing on the green signal. Meanwhile, it will spare 1 man, 1 elderly woman, 2 elderly men, and 1 woman, who were inside the car.

Case 2. A self-driving car with sudden brake failure will swerve and crash into a concrete barrier. This will result in the death of 1 man, 1 elderly woman, 2 elderly men, and 1 woman, who were inside the car. Meanwhile, it will spare 1 man, 1 girl, 2 boys, and 1 woman, who were abiding by the law by crossing on the green signal.

2.3 Summary Statistics on Sampling

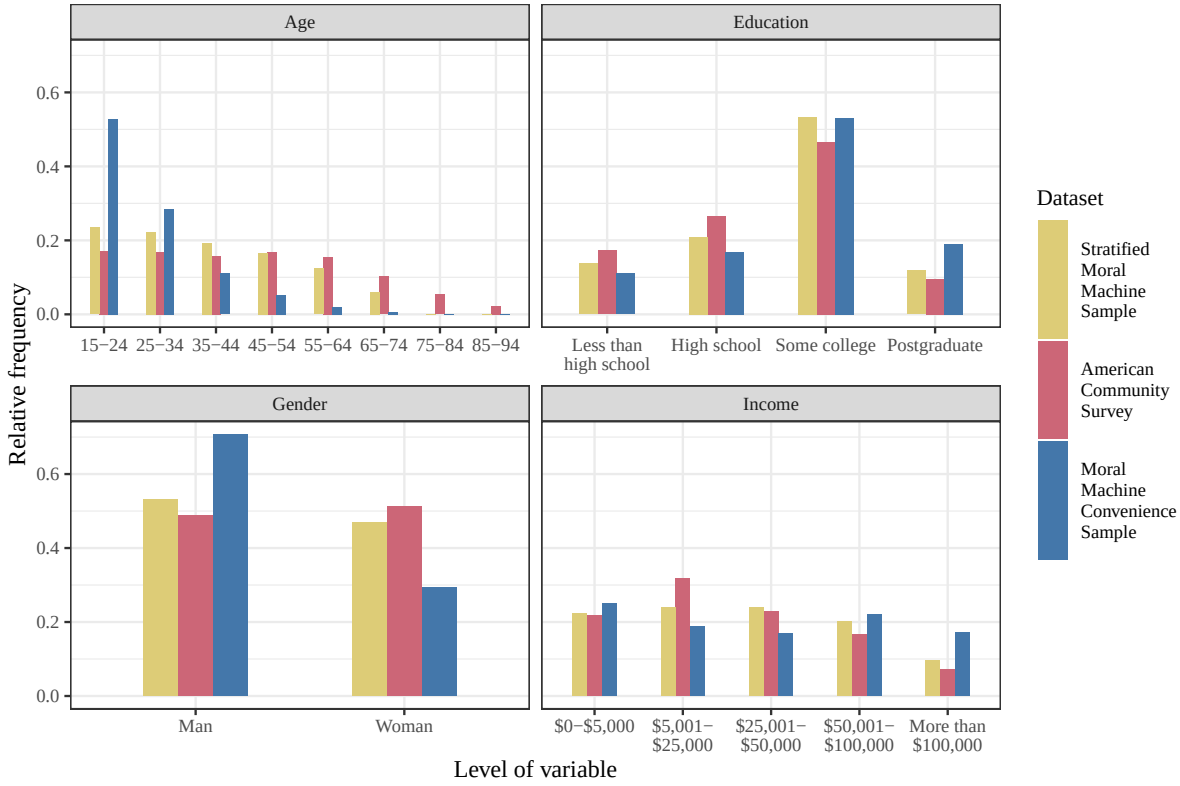

Figure 7: Comparison of demographic distributions of the convenience sample reported in the Moral Machine experiment (Awad et al., 2018), the 2016 American Community Survey, and the stratified sample used to demonstrate PPI in the present article. The stratified sample (yellow) is representative of the American population (red) on age, education, gender, and income.

2.4 Replication of AMCE estimates

Figure 8 compares the estimated causal effects of scenario attributes on the decision to save characters across samples. Compared to the representative sample of Americans from the Moral Machine experiment (yellow), the LLMs often fail to accurately predict how survey respondents' decisions depend on the scenario's characteristics (blue). We argue that the inaccuracies in LLM predictions render methods such as prediction-powered inference necessary when the goal is not only to obtain precise but also valid point estimates of causal effects.

For completeness, we also show the estimated AMCEs reported in Awad et al. (2018). While these estimates from their cross-country convenience sample do not have to align with the ones obtained from the stratified sample of Americans, there is not much difference. Note also that Awad et al. (2018) do not report confidence intervals due to their negligible size resulting from the large number of observations.

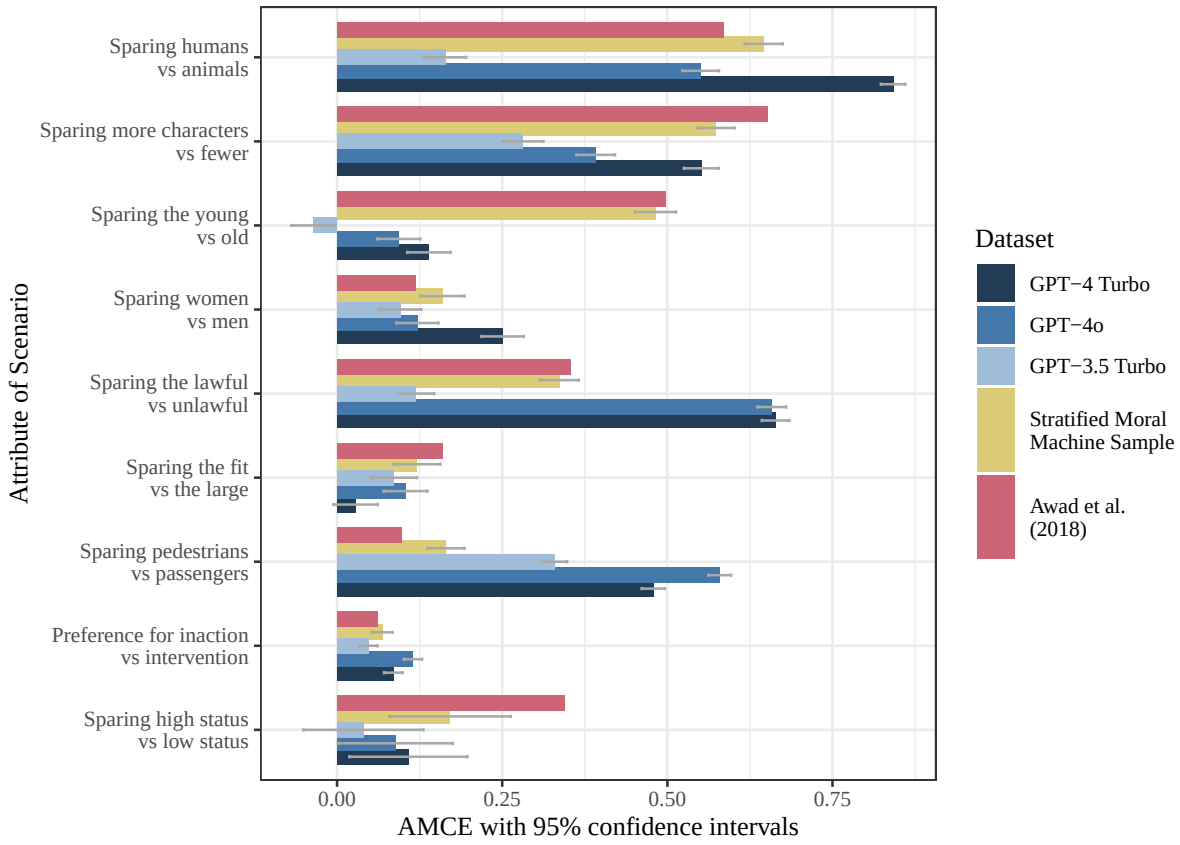

Figure 8: Comparison of AMCE estimates from human subjects against silicon subjects.

2.5 Prompting LLMs to Predict Decisions

Language model	**Context window**	**Training data**	**Input cost 1K tokens**	**Output cost 1K tokens**
gpt-4-turbo	128,000 tokens	Up to Dec 2023	$0.010	$0.030
gpt-4o	128,000 tokens	Up to Oct 2023	$0.005	$0.015
gpt-3.5-turbo-0125	16,385 tokens	Up to Sep 2021	$0.0005	$0.0015

Table 1: Details on LLMs used to predict survey responses

Model	Type	Correlation	N
gpt4turbo	Prediction	0.361	22315
	Replicate 1	0.346	5000
	Replicate 2	0.344	5000
	Mode across prediction and replicates	0.347	5000
	Without persona	0.337	4989
gpt4o	Prediction	0.311	22312
	Replicate 1	0.325	5000
	Replicate 2	0.304	5000
	Mode across prediction and replicates	0.317	5000
	Without persona	0.293	4974
gpt35turbo0125	Prediction	0.113	22314
	Replicate 1	0.112	4999
	Replicate 2	0.129	4999
	Mode across prediction and replicates	0.144	4998
	Without persona	0.174	5000

Table 2: Pearson correlation of survey respondents' decision for a moral dilemma with the LLM predicted decision. In addition to the 22,315 predictions, we assess how the correlation varies by prompting the LLM to give 5,000 additional predictions. We form a composite from the modal prediction of three identical prompts. For a separate set of predictions, we omit the demographic persona from the prompt.