Sociological Methods and Research

The Mixed Subjects Design: Treating Large Language Models as (Potentially) Informative Observations

Journal:	*Sociological Methods and Research*
Manuscript ID	Draft
Manuscript Type:	Original Article
Keywords:	Mixed Subjects Design, Prediction-Powered Inference, PPI Correlation, Experiments, Power Analysis, Machine Learning, Large Language Models, Moral Machine experiment, Computational Social Science

Abstract:	Large Language Models (LLMs) promise to transform the social sciences through cost-effective predictions of human behavior. However, despite growing evidence that LLMs can misrepresent such behavior, current approaches to studying causal effects with LLMs require researchers to assume that predicted and observed behavior are *interchangeable*. Instead, we argue that human subjects should serve as a gold standard to correct misrepresentations within a *mixed subjects design*. This paradigm offers valid and more precise estimates of causal effects at a lower cost than experiments relying solely on human subjects. We demonstrate—and extend—prediction-powered inference, a statistical method that instantiates the mixed subjects design. Our innovation is a power analysis for optimally choosing between *informative but costly* human subjects and *less informative but cheap* predictions of human behavior. Mixed subjects designs could enhance scientific productivity and reduce inequality in access to costly evidence on research questions by offering valid, precise, and cost-effective inferences on causal effects and other parameters.

The Mixed Subjects Design: Treating Large Language Models as (Potentially) Informative Observations

Abstract: Large Language Models (LLMs) promise to transform the social sciences through cost-effective predictions of human behavior. However, despite growing evidence that LLMs can misrepresent such behavior, current approaches to studying causal effects with LLMs require researchers to assume that predicted and observed behavior are *interchangeable*. Instead, we argue that human subjects should serve as a gold standard to correct misrepresentations within a *mixed subjects design*. This paradigm offers valid and more precise estimates of causal effects at a lower cost than experiments relying solely on human subjects. We demonstrate—and extend—prediction-powered inference, a statistical method that instantiates the mixed subjects design. Our innovation is a power analysis for optimally choosing between *informative but costly* human subjects and *less informative but cheap* predictions of human behavior. Mixed subjects designs could enhance scientific productivity and reduce inequality in access to costly evidence on research questions by offering valid, precise, and cost-effective inferences on causal effects and other parameters.

Keywords: Mixed Subjects Design, Prediction-Powered Inference, PPI Correlation, Experiments, Power Analysis, Machine Learning, Large Language Models, Moral Machine experiment, Computational Social Science

1 Introduction

Large language models (LLMs)—neural networks with billions of parameters trained on massive amounts of text data—have been shown to mimic how humans respond to surveys and experimental treatments in various settings. Accurately predicting[1] rather than observing human behavior could serve as a cost-effective and near-instantly available alternative to observing human 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; Abdurahman et al., 2024). Even in settings where LLMs happen to accurately 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. As a result, silicon sampling is of minimal practical benefit since human subjects data must be collected alongside LLM predictions at a scale sufficient to validate the interchangeability assumption. Some have therefore suggested that predictions be confined to exploratory stages of research, such as LLM-powered pilot studies for anticipating effect sizes (Grossmann et al., 2023; Ashokkumar et al., 2024).

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 from human subjects should inform inferences drawn from LLMs in a coherent statistical framework. We propose that observations from human subjects can be leveraged as a gold standard to correct for misrepresentations of human behavior through LLMs. We demonstrate how to implement this approach with prediction-powered inference (PPI) (Angelopoulos et al., 2023, 2024),

a recent statistical method that instantiates the mixed subjects approach. PPI allows researchers to combine observations of human behavior with predictions of that behavior generated by LLMs or other algorithms. So long as there is a correspondence between predicted and observed behavior, PPI produces valid point estimates with narrower confidence intervals than those derived solely from human subjects.

LLMs introduce a trade-off between predicted and observed behavior when estimating parameters such as causal effects. While obtaining predictions from LLMs is more cost-effective than recruiting human subjects, these predictions are less informative for estimating parameters than directly observed behavior. We derive a power analysis that formalizes and resolves this trade-off by balancing the costs of collecting these two types of data with the extent they inform inferences on parameters. Our power analysis allows researchers to allocate a fixed research budget to an optimal mix of human subjects and predictions that maximizes statistical power. Alternatively, researchers can minimize costs with an optimal combination of human subjects and predictions to achieve a given level of power. These functionalities will be integrated into the PPI Python library, available at https://github.com/aangelopoulos/ppi_py. With the extension of a power analysis, PPI becomes a fully usable methodology for conducting mixed subjects studies.

This article demonstrates how researchers can use LLMs 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 benefits of silicon sampling while avoiding its 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 and highlight the potential pitfalls.

2.1 Promises of the Silicon Subjects Design

The silicon subjects design asserts 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 the LLM's accuracy in predicting behavior (Gui and Toubia, 2023). Based on successful replications of canonical experiments with the silicon subjects design, some scholars concluded that LLM predictions were interchangeable with human behavior under certain conditions:

> *These findings could indicate thatâĂŤat least in some instancesâĂŤGPT-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: 9)

> *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: 597)

If silicon subjects could substitute human participants, LLMs may help overcome the limitations of experiments that exclusively draw on responses from human subjects. The first set of issues relates to the cost of conducting experiments with human subjects. Depending on wages paid to survey participants and fees for using online survey panels, a single survey response can cost several dollars. Typical survey experiments in the social sciences require significant numbers of survey participants to identify an effect. For example, researchers need a sample size of $n = 6,570$ to have a 90% chance of detecting an effect of size $d = 0.08$ with a two-sided t-test at $\alpha = 0.05$ (Figure S4 in the Supporting Information). While $d = 0.08$ represents the median effect size in a high-quality sample of online survey experiments in the social sciences (Rauf et al., 2024), the required number of participants is even higher for smaller effects. Larger sample sizes are also required for experiments that systematically assess a broad range of hypotheses (DellaVigna and Pope, 2018; Milkman et al., 2021; Voelkel et al., 2023; Tappin et al., 2023) and those aimed at estimating interaction effects (Gelman, 2018). Across these cases, the costs of 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. The cost of predicting a survey response with an LLM with currently available APIs can be as low as a fraction of a cent (Table S1).

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 on a typical panel for online research (Chandler et al., 2019). For instance, typical online panels for survey research consist of younger, more liberal, and more educated respondents who more are more likely to be White and who earn less on average than the American population (Berinsky et al., 2012; Levay et al., 2016; Zack et al., 2019). Collecting samples that are representative across

multiple dimensions—such as age, gender, income, and education—can be challenging since combinations of these characteristics may be rare among participants available on an online panel. 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 (Grossmann et al., 2023; Bail, 2024).

Finally, experimental research, like other quantitative scholarship, is only as good as the quality of measurements taken. Limited attention spans, insufficient effort, participant attrition, and non-compliance with research protocols are just a few examples of undesirable behaviors by study participants (Stantcheva, 2023). 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 their responses and strictly abide by the researchers' directions (Grossmann et al., 2023). Proponents of the silicon sampling approach could even argue that unrealistic distributions of LLM predictions help estimate 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, but that this estimate exhibits less statistical uncertainty than an estimate obtained from a sample of human subjects.

2.2 Perils of the Silicon Subjects Design

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 are male and from neighborhoods with high socioeconomic 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 and it remains unclear which ones can resolved. LLMs may inaccurately predict outcomes of certain groups of individuals because these have been misrepresented or underrepresented in the models' training data (Wang et al., 2024; Bail, 2024). While improving the representativeness and overall quality of the training data may enhance prediction accuracy, improving

the inputs to LLMs may not be sufficient to rule out errors stemming from the prediction algorithm itself. LLMs have been shown to respond differently depending on the order of a question or give the same answer consistently (Park et al., 2024). Errors may also arise from the complexity of the research design. While LLMs achieved remarkable accuracy in predicted responses to social surveys (Kim and Lee, 2024), predicting responses to experimental stimuli and more complex prediction tasks may result in higher error rates. At a more fundamental level, it remains unclear whether LLM predictions can be trusted without validation against human respondents. A treatment effect estimated based on LLM predictions could be a statistical fluke or an effect that would replicate in studies with human subjects (Harding et al., 2023).

Parameter estimates based on LLM predictions may not only be incorrect but also misleadingly precise. For example, using LLMs to generate many predicted values for an outcome Y and regressing these on an independent variable X results in narrow confidence intervals for a parameter simply because standard errors shrink with sample size. 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. Therefore, silicon samples may create a false sense of precision, leading to overly narrow confidence intervals with incorrect centers. This issue parallels the analysis of Big Data from non-representative samples, where researchers risk being "precisely inaccurate" (McFarland and McFarland, 2015). If biases are not adequately addressed, the availability of large and inexpensive data sources may do more harm than good (Meng, 2018; Bradley et al., 2021). For example, silicon samples could further amplify doubts about the replicability of findings from experimental social science (c.f. 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 a gold standard 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. In the following, we present prediction-powered inference (Angelopoulos

et al., 2023), a recent statistical framework that instantiates the mixed subjects approach.

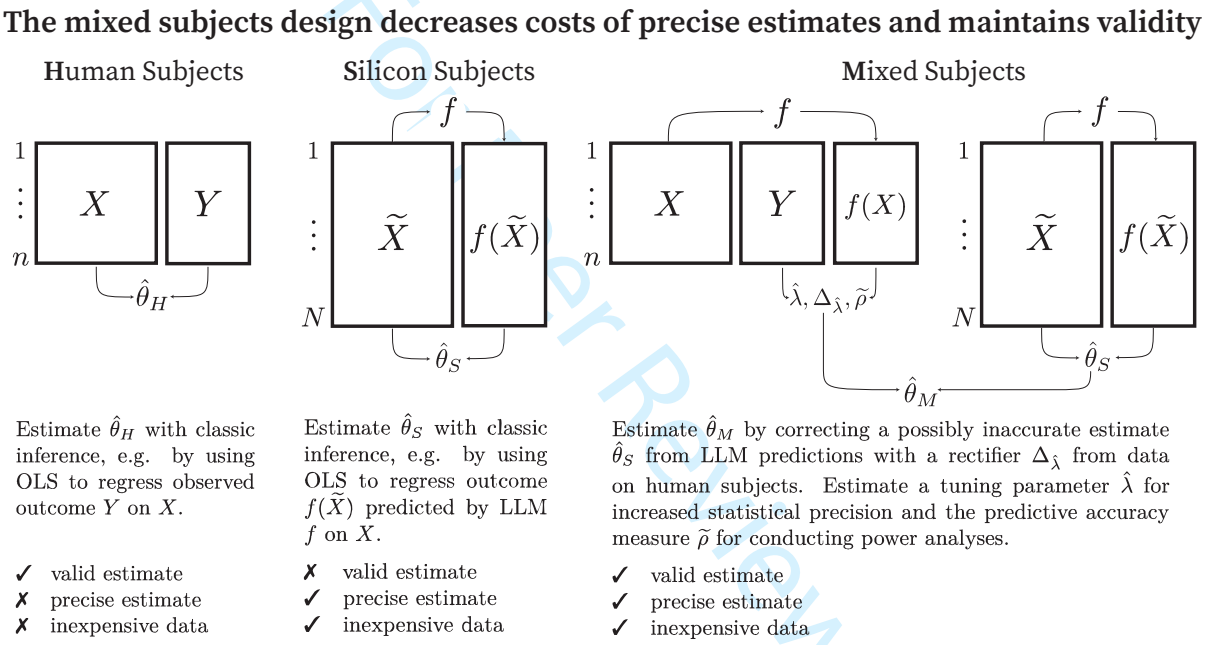

Figure 1: Comparison of experiments with human, silicon, and mixed subjects designs

3.1 The Mixed Subjects Design 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 target the same population parameters as a classical experiment (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 labeled) 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).

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

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 estimate is mostly based on the gold-standard observations.

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 subjects 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 using the software provided by Angelopoulos et al. (2023, 2024).

To apply PPI to a mixed subjects experiment, we need to verify several 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 unlabeled 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 The PPI Correlation

In section 2.2, we argued that using classical inference methods—such as regression—to estimate parameters with a large number 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, a study design that treats LLM predictions not as interchangeable with human subjects must account for how closely LLMs can predict behavior. Therefore, we derived the *PPI correlation* $\tilde{\rho} \in [-1, 1]$ as an empirical measure of the interchangeability assumption. The PPI correlation $\widetilde{\rho}$ measures the correlation between the classical estimator $\hat{\theta}_H$ based on human subjects and the estimator $\hat{\theta}_S$ based on the LLM predictions.[2] To demonstrate that $\tilde{\rho}$ measures interchangeability, we derived an *effective sample size*. The effective sample size in a mixed subjects experiment is the sample size required to achieve the same standard error for a parameter as in a human subjects experiment, allowing for a direct comparison of the two types of data. In section A.2 of the Supporting Information,

we show that the effective sample size when using PPI is given by

$$n_0 = n \cdot \frac{n+N}{n+N-N\tilde{\rho}^2}, \tag{2}$$

where n and N are the number of human and silicon subjects in the mixed subjects experiment and n_0 is the sample size in an equivalent human subjects experiment. When $\tilde{\rho} = 1$, the effective sample size is $n + N$ and human and silicon subjects are treated as equally informative. When $\tilde{\rho} = 0$, the effective sample size is n and only human subjects are used. In empirical applications, $\tilde{\rho}$ will likely be between 0 and 1, indicating that the prediction algorithm is informative but not as informative as human subjects for estimating a parameter. For example, if the PPI correlation is $\tilde{\rho} = 0.75$ and $N/n = 5$ so that for every human subject there are 5 silicon subjects, conducting a mixed subjects experiment PPI is equivalent to a human subjects experiment with 88% more participants. As such, $\tilde{\rho}$ quantifies how informative predictions are in estimating θ and measures the extent to which predictions are interchangeable with gold standard data.

In the following, we show that higher values of the PPI correlation are crucial for obtaining smaller standard errors for parameters, implying narrower confidence intervals, higher statistical power, and lower costs of conducting mixed subjects experiments relative to human subjects experiments. As shown in section A.1 of the Supporting Information, the PPI standard error can be written as

$$\text{SE}(\hat{\theta}^{\text{PP}}) = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{N}{N+n}\tilde{\rho}^2}, \tag{3}$$

where $\sigma/\sqrt{n}$ is the standard error of $\hat{\theta}_H$ in an experiment with n human subjects. 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 $\sigma/\sqrt{n}$. A smaller standard error for $\hat{\theta}^{\text{PP}}$ implies that a mixed subjects design with PPI produces narrower confidence intervals and higher statistical power than a human subjects experiment. Figure 2 illustrates how achieving this higher statistical precision depends on $\tilde{\rho}$ and on the ratio N/n. The PPI standard error becomes narrower than the classical standard error for larger N relative to n, and this benefit is greatest for higher values of the PPI correlation. For example, if $N/n = 5$ and $\tilde{\rho} = 0.5$, then the PPI standard error will be approximately 11% smaller than the classical standard error since $\sqrt{1 - (N/(N+n))\tilde{\rho}^2} \approx 0.89$. 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 same consideration applies to confidence intervals since the ratio of standard errors is equivalent to the ratio of the width of confidence intervals. Finally, higher values of $\tilde{\rho}$, implying a smaller standard error, result in a non-linear increase in statistical power.

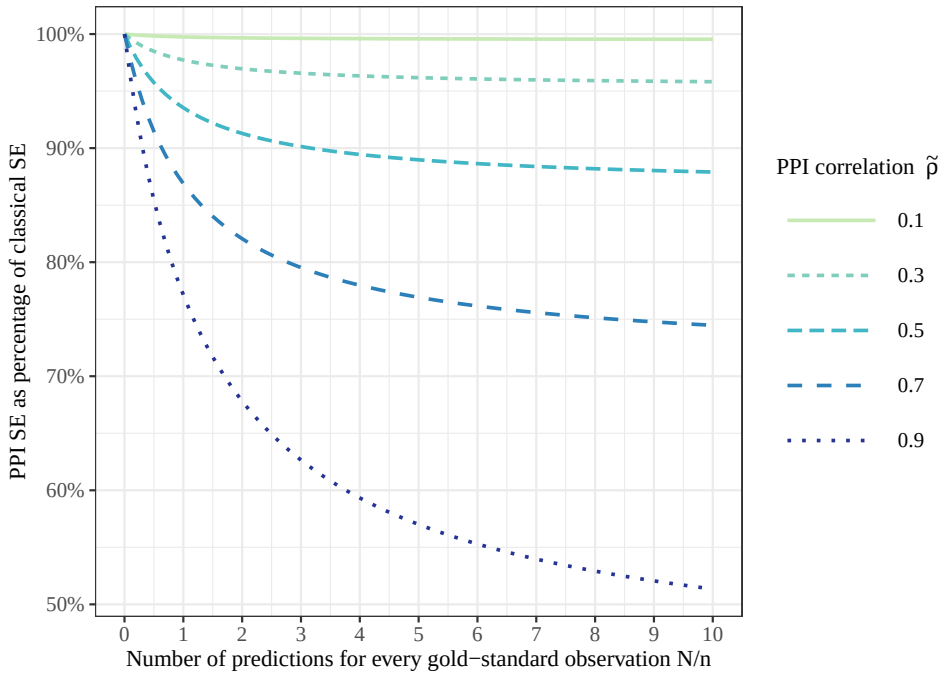

Figure 2: The x-axis shows the ratio N/n of samples sizes with N predictions $f(X_i)$ and n gold standard observations Y_i. The y-axis shows the ratio of the PPI standard error to the classical standard error, defined by $\sqrt{1-(N/(N+n))\tilde{\rho}^2}$.

The PPI standard error directly depends on $\tilde{\rho}$, the ratio of silicon subjects to the total sample size $N/(N+n)$, and the standard error of the classical estimator $\sigma/\sqrt{n}$. As shown in Figure 2, including more silicon subjects is most effective in reducing standard errors when the PPI correlation $\tilde{\rho}$ is closer to 1. Hence, researchers can maximize returns on predictions by using algorithms that accurately portray human behavior. Researchers could increase $\tilde{\rho}$ by choosing more accurate prediction algorithms. These may include LLMs with more parameters, those trained on higher-quality data, models fine-tuned for specific prediction tasks, or models with retrieval-augmented generation (Lewis et al., 2021). Prompt engineering—such as providing more context, examples, or specific instructions—may also enhance prediction accuracy.

3.3 PPI Power Analysis

Power analyses allow researchers to determine the necessary sample size for a desired level of power—i.e., the probability of correctly rejecting the null hypothesis 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.

Our power analysis for the mixed subjects design is based on the trade-off between human and silicon subjects. Obtaining a silicon sample is much more affordable than recruiting human subjects. However, silicon subjects are generally less informative than humans when estimating a parameter, corresponding to a PPI correlation of $\tilde{\rho} < 1$ when human and silicon subjects are not interchangeable. Researchers can make an optimal choice for combining a sample size of human subjects n with N silicon subjects, and this combination depends on the PPI correlation $\tilde{\rho}$, a hypothesized effect size, a desired level of power, the cost of recruiting human and silicon subjects, and the available research budget. Given these parameters, our power analysis allows researchers to optimally decide between recruiting *costly but informative* human subjects or *less informative but cheap* 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 hypothesized effect size, fixed research budget, the cost of recruiting silicon subjects relative to human subjects, and how informative

LLMs are in predicting human behavior $\tilde{\rho}$? Researchers may be particularly interested in finding the *most powerful pair* (n, N) 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 combination of sample sizes (n, N) is the cheapest to sample for a desired level of power, an effect size, the costs of silicon relative to human subjects, and the PPI correlation $\tilde{\rho}$? 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).

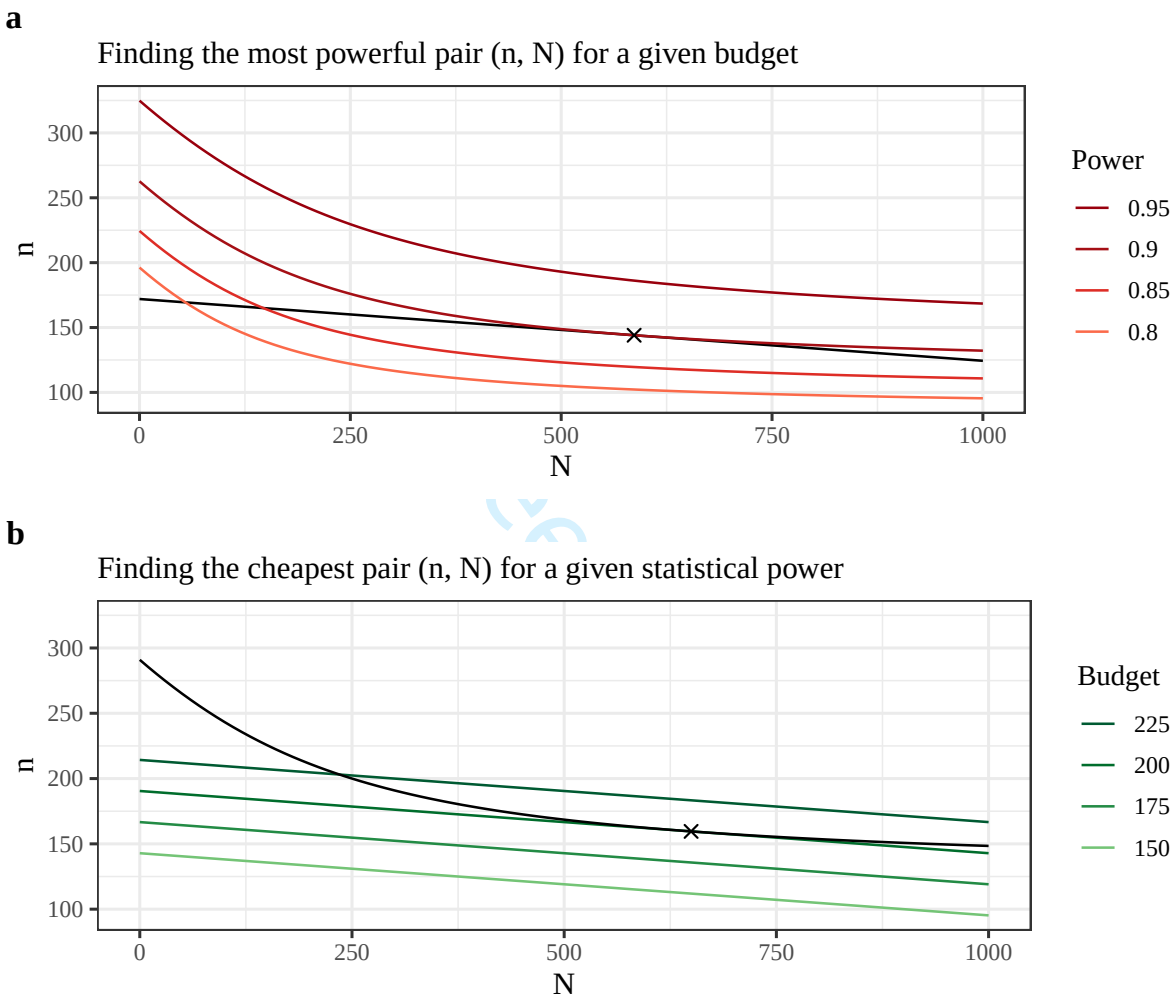

Figure 3: Illustration of the constraint optimization with $\tilde{\rho} = 0.75$, effect size $\delta = 0.2$, classical standard error $\hat{\sigma}/\sqrt{n} = 1$, and a ratio of the costs of sampling silicon subjects to human subjects $\gamma = 0.05$ (**a**) Given a fixed budget, the PPI power analysis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

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 $\tilde{\rho}$ needs to be estimated from a small dataset $\{(X_i, Y_i, f(X_i))\}_{i=1}^{n}$. Researchers could also test hypothetical values for $\tilde{\rho}$, reflecting more or less accurate prediction algorithms. However, applying the PPI software is required to obtain precise estimates. Details on the derivation of the PPI power analysis are given in sections A.4 and A.5 of the Supporting Information.

3.4 Lowering Costs of Data Collection

PPI produces narrower confidence intervals and higher statistical power than classical inference. Whether PPI is also more cost-effective depends on the PPI correlation $\tilde{\rho}$ and γ — the ratio of the costs of surveying silicon and human subjects. That is,

$$\gamma = \frac{c_f}{c_Y}, \tag{4}$$

where c_f is the cost of prompting an LLM to give a prediction, and c_Y is the cost of surveying a human subject. In section A of the Supporting Information, we show that PPI is more cost-effective than classic inference with human subjects if and only if

$$\tilde{\rho}^2 > \frac{4\gamma}{(1+\gamma)^2}. \tag{5}$$

For example, we could assume a cost of \$3.20 for a participant to fill out a 12-minute survey, paying participants the 2024 California minimum wage of \$16.00/hour. Based on the costs of \$0.003 for prompting an LLM for this study, we calculated that any $\tilde{\rho} > 0.06$ would be sufficient for PPI to save costs when compared to classical inference with human subjects only. More generally, if condition (5) is satisfied and researchers use the optimal sample size from the PPI power analysis, conducting mixed subjects can be substantially more cost-effective than experiments with human subjects only. Figure 4 shows that PPI experiments become less expensive than classical experiments as predictions become more affordable, with substantial savings at higher values of the PPI correlation. This theoretical result carries important implications as the field of generative AI continues to advance. The costs of conducting mixed subjects experiments will further decrease relative to classical experiments as costs for prompting LLMs decrease (e.g. API fees) and LLMs become more capable of predicting human behavior.

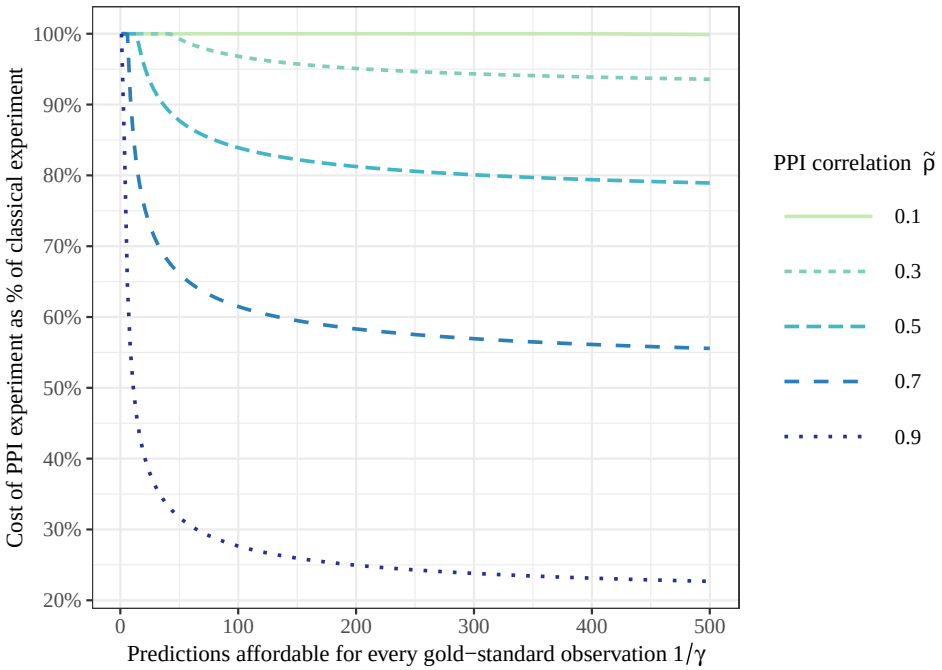

Figure 4: The y-axis shows the costs of conducting a mixed subjects experiments with PPI as a percentage of the costs of conducting a classical experiment with human subjects only, given by $1 - \tilde{\rho}^2(1-\gamma) + 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}$, as a function of the PPI correlation $\tilde{\rho}$ and the number of silicon subjects that researchers can afford for the costs of recruiting a human subject $1/\gamma = c_Y/c_f$.

4 Application to the Moral Machine Experiment

The Moral Machine experiment (Awad et al., 2018) sought to better understand the factors influencing people's decisions in moral dilemmas that self-driving cars might face on the road. In this conjoint experiment, participants were presented with hypothetical scenarios where a sudden brake failure would result in harm to either passengers or pedestrians. Participants could only spare one of the two groups. If participants choose to save the passengers, the autonomous vehicle would drive through a crosswalk where pedestrians are crossing the street. If participants chose to spare the pedestrians instead, the car would crash into a concrete barrier. The experiment measured how attributes such as age, gender, social status, and the number of individuals influenced the probability of participants choosing to save one group over the other. Using a weighted simple linear regression, Awad et al. (2018) estimate the Average Marginal Component Effect (AMCE). The AMCE represents the causal effect of an attribute of a moral dilemma on a respondent's decision to spare passengers or pedestrians.

Our main interests in this application of PPI with LLMs are (**a**) to assess the extent to which including LLM predictions increases statistical precision and (**b**) to compare the validity of point estimates from PPI to those derived from LLM predictions. Analogous to Figure 2, we define increases in precision as the percent reduction in the width of the PPI confidence intervals relative to the width of the confidence interval obtained from human subjects only. We define the validity of a point estimate by the percent of confidence intervals that cover the true causal effect in a specific population. While such population parameters remain of course unknown, we use a quota sample of Americans who responded to the Moral Machine experiment to obtain best possible estimates of the true AMCEs (Figure S3). These AMCEs serve as a benchmark for comparing the validity of silicon sampling and the mixed subjects design with PPI.

4.1 Methods

Awad et al. (2018) obtained a convenience sample with millions of decisions on moral dilemmas from participants worldwide. We used the subset of 492,921 participants who completed an optional demographic survey to obtain a sample of the American population. Using quotas on age, education, gender, and income from the 2016 American Community Survey (Ruggles et al., 2024), we randomly sampled 2,097 Americans who evaluated a total of 22,315 moral dilemmas. Our sample closely resembles the demographics of the United States, except for older individuals, who could not be sufficiently sampled due to their minimal presence in the Moral Machine experiment (Figure S2).

Next, we created the prompts for the LLMs based on the replication data from the Moral Machine experiment. The replication data records the attributes of the moral dilemmas evaluated by survey

participants, such as the number of passengers in the car. We converted this numerical representation into a text description of the dilemmas with computer code adapted from a related study (Takemoto, 2024). We also added a demographic profile to the prompt, including the age, education level, gender, and income of the survey respondent who evaluated the dilemma. Please refer to section B in the Supporting Information for an example. We then used the OpenAI API to prompt four LLMs—GPT4 Turbo (gpt-4-turbo), GPT4o (gpt-4o), and GPT3.5 Turbo (gpt-3.5-turbo-0125)—to predict the decisions of these survey respondents for the moral dilemmas.

We then compared the validity and statistical precision of point estimates derived from PPI against a naive approach that pools human and silicon subjects by incrementally increasing the number of silicon subjects. From the quota sample of 2,097 Americans, we randomly selected $n = 500$ human subjects and added $n \times k = N$ silicon subjects for $k = (.25, .5, .75, 1, 1.5, \ldots, 4.5, 5)$. For each combination of sample sizes n and $N = (125, 250, ..., 2500)$, we repeated the sampling 300 times and calculated the mean width and coverage of the confidence intervals. We estimated the AMCE for each scenario attribute with a weighted simple linear regression (Hainmueller et al., 2014). To assess the validity and precision of the naive approach, we applied the weighted simple linear regression to the pooled sample of size $n + N$. We used the Python library created by Angelopoulos et al. (2023, 2024) to obtain the corresponding PPI estimates of the AMCE.

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. To explore potential ways of improving accuracy, we conducted two supplemental analyses (Table S2). First, we prompted each LLM twice to give 5,000 additional predictions using the same prompts. We then created a composite by taking the mode of the 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, these supplemental analyses yielded very similar correlation coefficients within each LLM. For all subsequent analyses, we focus on the 22,315 predicted survey responses generated by GPT-4 Turbo.

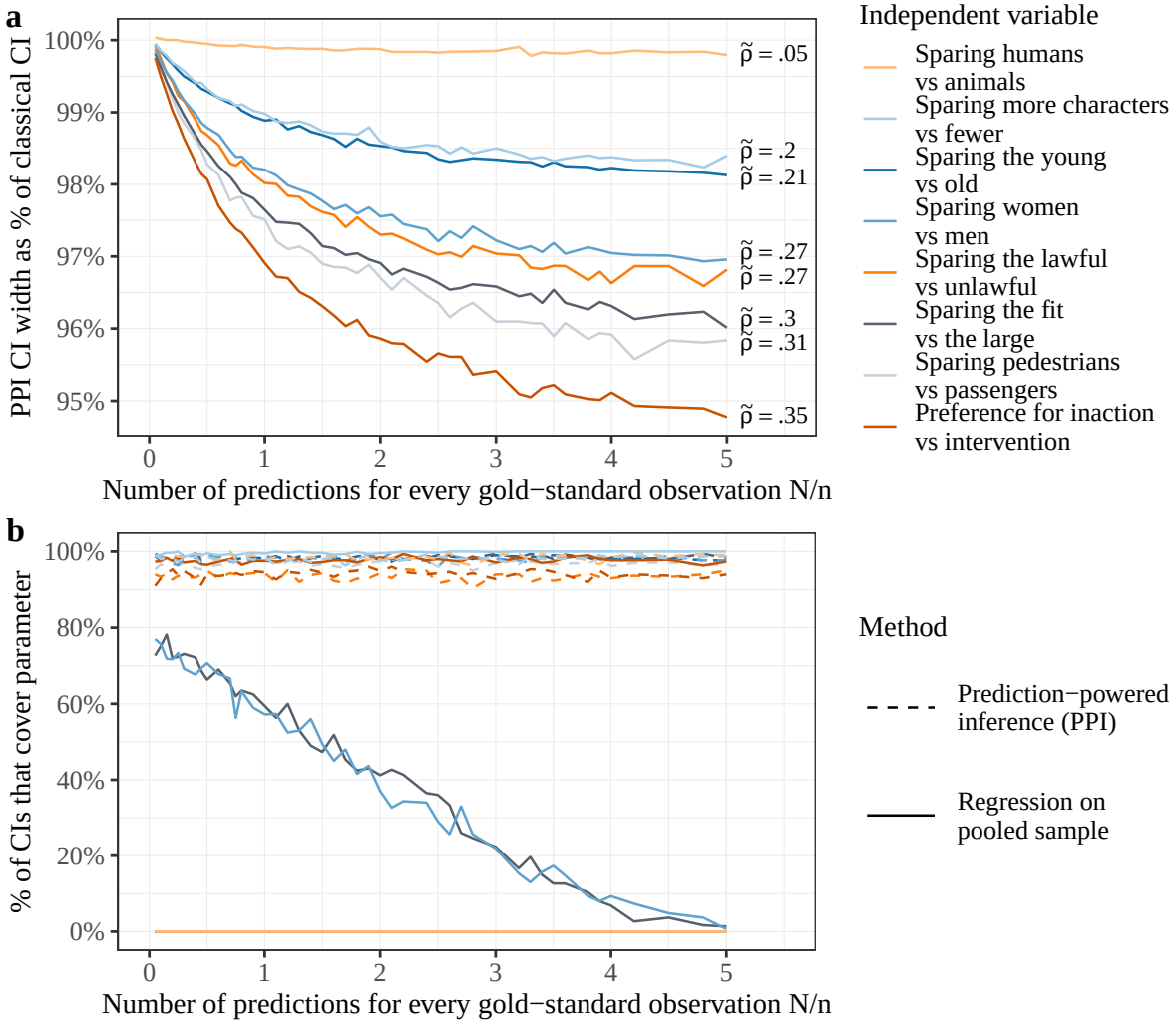

Figure 5: The x-axes show the ratio of sample sizes N/n with N silicon subjects and n gold standard observations. (**a**) The y-axis shows the width of the PPI confidence interval (CI) as a share of the CI from a regression on the pooled sample of size $n + N$. Values smaller than 100% indicate the percent reduction in PPI CI width relative to the regression CI (cf. Figure 2). (**b**) The y-axis shows the percent of CIs calculated with PPI and regression that cover the AMCE estimates from the quota sample of 2,097 Americans; the estimates from this sample are used as the best available approximation of the AMCE parameters.

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 approach that naively pools human and silicon subjects. Figure 5a shows that adding an increasing number of LLM predictions reduces the width of confidence intervals more strongly for larger values of the PPI correlation $\tilde{\rho}$ (cf. Figure 2). Figure 5b shows that the percent of PPI confidence intervals that cover the parameter remains stable at high levels. In contrast, the coverage of the classical intervals computed for the naive approach remains stable only for some independent variables. In sum, this analysis illustrates that PPI produces valid point estimates, with increases in statistical precision depending on the PPI correlation. The silicon sampling approach may also produce valid point estimates but this is impossible to ascertain without validation on human subjects. PPI automatically handles this validation by introducing a statistical correction to the point estimates derived from LLM predictions.

5 Conclusion

Large language models, and generative AI more generally, offer new data sources that may transform the social sciences. Predictions of human behavior—often called "silicon subjects"—provide a cost-effective and near-instantly available alternative to observing behavior in human subjects studies. However, like novel data sources that have come before (Lazer et al., 2009, 2020), computational social scientists must critically assess the limitations of LLMs and develop robust methods to ensure sound conclusions from this emerging data source. We argue that researchers risk drawing incorrect conclusions when treating LLM predictions as interchangeable with observed human behavior. Estimating parameters based on large numbers of predictions can give a false sense of precision because the confidence intervals will be overly narrow, while the point estimates may systematically diverge from those estimated on a sample of human subjects. Even if LLMs become more accurate in predicting human behavior, these predictions remain of minimal benefit because researchers still need to validate the assumption of interchangeability with an appropriately large sample of human subjects.

We propose that LLM predictions be integrated with, rather than replace, human subjects in what we call a mixed subjects design. We demonstrate and extend prediction-powered inference (PPI), a statistical method that adjusts possibly invalid point estimates derived from LLM predictions to produce valid estimates. Mixed subjects studies with PPI also allow researchers to obtain narrower confidence intervals and higher statistical power than studies with human subjects only. Therefore, the mixed-subjects design with PPI allows researchers to combine the strengths of the human and silicon subjects approach.

Our statistical contributions to PPI are two-fold. First, we derive the PPI correlation $\tilde{\rho}$ as an empirical measure of the extent to which human subjects and LLM predictions are interchangeable. We show

that high values of the PPI correlation produce small standard errors for parameters, implying narrower confidence intervals, higher statistical power, and lower costs of conducting mixed subjects experiments relative to human subjects experiments. If LLMs and other algorithms become more capable of predicting human behavior in the future, this improvement will be reflected in higher values for the PPI correlation. More capable algorithms will result in higher statistical precision of PPI estimates and the cost of conducting mixed subjects experiments will further decrease relative to human subjects experiments.

Our second statistical contribution is a power analysis for PPI that addresses the trade-off between silicon and human subjects if they are not fully interchangeable (i.e., $\tilde{\rho} < 1$). The PPI power analysis allows researchers to optimally choose between recruiting *informative but costly* human subjects and *less informative but cheap* silicon subjects. Researchers can allocate a given budget to maximize power or minimize budget expenditure to achieve a desired level of power. Statistical software published alongside this article completes the toolkit necessary to conduct mixed subjects studies with PPI.

Our work points to immediate next steps for mixed-subject experimental design. While we leverage PPI to implement these designs, we note that other methods are also well-suited for this purpose. For instance, researchers could combine samples of silicon and human subjects in a Bayesian regression framework (e.g. Jones et al., 2011). Here, priors on the parameter values correspond to the degree to which researchers want to treat silicon subjects as interchangeable with human subjects. We also want to emphasize that the literature on doubly robust machine learning offers other promising routes for implementing mixed-subjects designs (Egami et al., 2024; Kallus and Mao, 2024). More generally, the development of a robust toolkit of mixed-subject methodologies will allow researchers to leverage LLMs and other forms of generative AI to pursue their research questions.

In a second future direction, the possibility of obtaining valid and precise estimates at low costs from a mixed-subjects design could be leveraged to conduct studies that would otherwise be prohibitively expensive. For instance, identifying small treatment effects or interactions with sufficient statistical power requires thousands of observations, implying costs that may be too high when estimating these effects with human subjects alone. Moreover, studies aimed at systematically exploring a larger number of hypotheses and possible experimental designs have important practical and theoretical implications (Almaatouq et al., 2024), but often require an inordinate number of human subjects. The mixed subjects design could be integrated with existing research infrastructure to facilitate such large-scale experiments (Almaatouq et al., 2021). 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 data for research questions and hypotheses. The resulting savings could also be allocated to other research projects or used to pay higher wages to survey participants.

Notes

1. We recognize that those deploying the methodology of "silicon sampling" prefer to describe LLMs used in this way as "mimicing" or "modeling" human behavior (Argyle et al., 2023; Horton, 2023) instead of "predicting" it. We also acknowledge that there are nuanced differences between these uses for data (see, e.g., Breiman 2001). However, exploring these complexities is beyond the scope of this article, and so we will stick with "predict" as our preferred term.

2. Hence, the PPI correlation $\tilde{\rho}$ does not directly refer to the correlation between predicted and observed values of the dependent variable. The PPI correlation $\widetilde{\rho}$ is defined mathematically in equation (9) in the Supporting Information.

References

Abdurahman, Suhaib, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J. Xue, Jackson Trager, Peter S. Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. "Perils and opportunities in using large language models in psychological research." *PNAS Nexus* 3.

Almaatouq, Abdullah, Joshua Becker, James P. Houghton, Nicolas Paton, Duncan J. Watts, and Mark E. Whiting. 2021. "Empirica: a virtual lab for high-throughput macro-level experiments." *Behavior Research Methods* 53:2158–2171.

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.

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." .

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.

Ashokkumar, Ashwini, Luke Hewitt, Isaias Ghezae, and Robb Willer. 2024. "Predicting Results of Social Science Experiments Using Large Language Models."

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

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

Bail, Christopher A. 2024. "Can Generative AI improve social science?" *Proceedings of the National Academy of Sciences* 121.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20:351–368.

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

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. 1âĂŞ16.

Bradley, Valerie C., Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. 2021. "Unrepresentative big surveys significantly overestimated US vaccine uptake." *Nature* 600:695–700.

Breiman, Leo. 2001. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." *Statistical science* 16:199–231.

Chandler, Jesse, Cheskie Rosenzweig, Aaron J. Moss, Jonathan Robinson, and Leib Litman. 2019. "Online panels in social science research: Expanding sampling methods beyond Mechanical Turk." *Behavior Research Methods* 51:2022–2038.

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

DellaVigna, Stefano and Devin Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *The Review of Economic Studies* 85:1029–1069.

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

Egami, Naoki, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2024. "Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models." arXiv:2306.04746 [cs, stat].

Faul, Franz, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses." *Behavior Research Methods* 41:1149–1160.

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.

Gelman, Andrew. 2018. "You need 16 times the sample size to estimate an interaction than to estimate a main effect | Statistical Modeling, Causal Inference, and Social Science."

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." .

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 DâĂŹAlessandro, NG Laskowski, and Robert Long. 2023. "AI language models cannot replace human research participants." *Ai & Society* pp. 1–3.

Horton, John J. 2023. "Large language models as simulated economic agents: What can we learn from homo silicus?" Technical report, National Bureau of Economic Research.

Jones, Hayley E, David I Ohlssen, Beat Neuenschwander, Amy Racine, and Michael Branson. 2011. "Bayesian models for subgroup analysis in clinical trials." *Clinical Trials* 8:129–143.

Kallus, Nathan and Xiaojie Mao. 2024. "On the role of surrogates in the efficient estimation of treatment effects with limited outcome data." arXiv:2003.12408 [cs, stat].

Kim, Junsol and Byungkyu Lee. 2024. "AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction."

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.

Lazer, David M. J., Alex Pentland, Duncan J. Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, Alondra Nelson, Matthew J. Salganik, Markus Strohmaier, Alessandro Vespignani, and Claudia Wagner. 2020. "Computational social science: Obstacles and opportunities." *Science* 369:1060–1062.

Levay, Kevin E., Jeremy Freese, and James N. Druckman. 2016. "The Demographic and Political Composition of Mechanical Turk Samples." *Sage Open* 6:1–17.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,

Heinrich KÃijttler, Mike Lewis, Wen-tau Yih, Tim RocktÃďschel, Sebastian Riedel, and Douwe Kiela. 2021. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." arXiv:2005.11401 [cs].

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.

Meng, Xiao-Li. 2018. "Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election." *The Annals of Applied Statistics* 12:685–726.

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

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

Rauf, Tamkinat, Jan G. Voelkel, James Druckman, and Jeremy Freese. 2024. "An Audit of Social Science Survey Experiments." Publisher: Open Science Framework.

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.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15:205–234.

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

Tappin, Ben M., Chloe Wittenberg, Luke B. Hewitt, Adam J. Berinsky, and David G. Rand. 2023.

"Quantifying the potential persuasive returns to political microtargeting." *Proceedings of the National Academy of Sciences* 120. Publisher: Proceedings of the National Academy of Sciences.

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

Voelkel, Jan G., Michael Stagnaro, James Chu, Sophia L. Pink, Joseph S. Mernyk, Chrystal Redekopp, and Isaias Ghezae, et al. 2023. "Megastudy identifying effective interventions to strengthen American-sâĂŹ democratic attitudes."

Wang, Angelina, Jamie Morgenstern, and John P. Dickerson. 2024. "Large language models cannot replace human participants because they cannot portray identity groups."

Zack, Elizabeth S., John Kennedy, and J. Scott Long. 2019. "Can Nonprobability Samples be Used for Social Science Research? A cautionary tale." *Survey Research Methods* pp. 215–227.

Supporting Information

A Prediction Powered Inference (PPI)

A.1 PPI Standard Error and Correlation

In this section, we will define the PPI correlation $\tilde{\rho}$ and show that the standard error of the PPI estimator $\hat{\theta}^{\text{PP}}$ is equal to

$$\text{SE}(\hat{\theta}^{\text{PP}}) = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{N}{N+n}\tilde{\rho}^2}. \tag{6}$$

This expression for the standard error is given in equation (3) and used in our power analysis. To define $\tilde{\rho}$ and prove equation (6), we will use some results and concepts from Angelopoulos et al. (2024). In particular, equation (6) only holds when $\hat{\theta}^{\text{PP}}$ is the *power-tuned* PPI estimator – a concept introduced in (Angelopoulos et al., 2024, Section 6) and reviewed here.

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. 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

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

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

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

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

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

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

$$\sqrt{n}\left(\hat{\theta}^{\text{PP}}_\lambda - \theta^\star\right) \xrightarrow{d} \mathcal{N}(0, \Sigma^\lambda), \tag{7}$$

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

$$\begin{aligned}\Sigma^\lambda =& H^{-1}_{\theta^\star}\,\text{Cov}(\nabla\ell_{\theta^\star})H^{-1}_{\theta^\star} + \lambda^2(1+r)H^{-1}_{\theta^\star}\,\text{Cov}(\nabla\ell^f_{\theta^\star})H^{-1}_{\theta^\star} \\ &- \lambda H^{-1}_{\theta^\star}\left(\text{Cov}(\nabla\ell^f_{\theta^\star}, \nabla\ell_{\theta^\star}) + \text{Cov}(\nabla\ell_{\theta^\star}, \nabla\ell^f_{\theta^\star})\right)H^{-1}_{\theta^\star},\end{aligned} \tag{8}$$

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

By equation (7) the standard error of the jth coordinate $\hat{\theta}^{\text{PP}}_{\lambda,j}$ is $\sqrt{\Sigma^{\lambda}_{j,j}/n}$. We will show that if λ is chosen by *power tuning* (Angelopoulos et al., 2024, Section 6), then the PPI standard error will simplify to the expression in (6).

Power tuning (Angelopoulos et al., 2024, Section 6) chooses λ to minimize $\Sigma^{\lambda}_{j,j}$. This is equivalent to choose λ to minimize the standard error of $\hat{\theta}^{\text{PP}}_{\lambda,j}$. From (8), 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

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

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

$$\tilde{\rho}^2_j = \frac{\left[H^{-1}_{\theta^\star}\operatorname{Cov}(\nabla\ell^f_{\theta^\star},\nabla\ell_{\theta^\star})H^{-1}_{\theta^\star}\right]^2_{j,j}}{\left[H^{-1}_{\theta^\star}\operatorname{Cov}(\nabla\ell^f_{\theta^\star})H^{-1}_{\theta^\star}\right]_{j,j}\left[H^{-1}_{\theta^\star}\operatorname{Cov}(\nabla\ell_{\theta^\star})H^{-1}_{\theta^\star}\right]_{j,j}}. \tag{9}$$

The quantity σ^2_j is the asymptotic variance of the classical estimator $\hat{\theta}^{\text{classical}}$ and $\tilde{\rho}_j$ is the PPI correlation.

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

$$\text{SE}(\hat{\theta}^{\text{PP}}_{\lambda^\star_j,j}) = \sqrt{\Sigma^{\lambda^\star_j}_{j,j}/n} = \frac{\sigma_j}{\sqrt{n}}\sqrt{1 - \frac{N}{N+n}\tilde{\rho}^2_j}.$$

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 $\hat{\theta}^{\text{PP}}_{\lambda^\star_j,j}$.

To simplify notation, we will write $\hat{\theta}^{\text{PP}}$ for $\hat{\theta}^{\text{PP}}_{\hat{\lambda}_j,j}$ and σ instead of σ_j and $\tilde{\rho}$ instead of $\tilde{\rho}_j$. With this notation we have

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

as claimed in equations (3) and (6).

A.2 The Effective Sample Size in PPI

The *effective sample size* is the number n_0 of labeled data points that would give the same standard error of using PPI with n labeled points and N unlabeled points. The standard error with n_0 labeled data points is simple $\sigma/\sqrt{n_0}$. Equating this with the PPI standard error in (6) gives

$$n_0 = n \cdot \frac{n+N}{n+N-N\tilde{\rho}^2}.$$

Let $k = \frac{N}{n}$. Then, the effective sample size in PPI increases the number of human subjects by a factor of

$$\frac{1+k}{1+k-k\tilde{\rho}^2} \geq 1.$$

When $\tilde{\rho}^2 = 1$, this factor equals $1+k$ and the effective sample size is $n(1+N/n) = n+N$. That is, the effective sample size is the size of the full pooled sample. When $\tilde{\rho} = 0$, this factor equals 1 and the effective sample size is n meaning that only the labeled samples are used.

A.3 The Cost of PPI and Classical Inference

In this section we derive the percentage of cost saved by PPI as reported in Figure 4. This is done by first finding all pairs of sample sizes (n, N) that achieve the same standard error as a baseline classical experiment with n_0 labeled data points. Then, given the costs of collecting a labeled and unlabeled sample, we find the optimal pair $(n^\star, N^\star)$ that minimizes the cost of PPI while still having the same power as the baseline classical experiment. The results in this section are used in the next two sections to derive the cheapest pair and most powerful pair for our power analysis.

Let $S(n, N)$ be the standard error of the PPI estimate, so that

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

In a baseline experiment n_0 labeled samples, the standard error is $S(n_0, 0) = \sigma/\sqrt{n_0}$. By rearranging the equation $S(n,N) = \sigma/\sqrt{n_0}$, it can be shown that $S(n,N) = \sigma/\sqrt{n_0}$ if and only if

$$n_0(1-\tilde{\rho}^2) < n \leq n_0 \quad \text{and} \quad N = \frac{n(n_0-n)}{n - n_0(1-\tilde{\rho}^2)}. \tag{10}$$

Let c_X and c_Y be the cost of collecting X and Y and let c_f be the cost of computing $f(X)$. The cost of performing PPI with n labeled subjects and N unlabeled subjects is

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

This is because PPI requires X_i, Y_i, and $f(X_i)$ for the n labeled samples and requires $\widetilde{X}_i$ and $f(\widetilde{X}_i)$ for the N unlabeled samples. If we let $\gamma = (c_X + c_f)/c_Y$, then

$$C(n,N) = c_Y((1+\gamma)n + \gamma N)$$

Our goal is to find the pair $(n^\star, N^\star)$ that satisfies the constraints in equation (10) and minimizes $C(n,N)$.

That is, $(n^\star, N^\star)$ is the solution to the optimization problem

$$\begin{array}{ll} \text{minimize} & C(n, N) \\ \text{subject to} & \text{equation (10)} \end{array} \tag{11}$$

This optimization problem can be solved by first substituting $N = \frac{n(n_0-n)}{n-n_0(1-\tilde{\rho}^2)}$ into $C(n, N)$. This gives the cost as a function of n alone. Setting the derivative of this function equal to zero gives

$$n^\star = n_0\left(1-\tilde{\rho}^2+\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}\right) \quad \text{and} \quad N^\star = \frac{n^\star(n_0-n^\star)}{n^\star-(1-\tilde{\rho}^2)n_0}. \tag{12}$$

At the optimal pair $(n^\star, N^\star)$, the cost is

$$C(n^\star, N^\star) = c_Y n_0\left(1-\tilde{\rho}^2+\gamma\tilde{\rho}^2+2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}\right). \tag{13}$$

In contrast, the cost of performing classical inference with n_0 human subjects is

$$C_0(n_0) = (c_Y + c_X)n_0.$$

This equation for the costs in classical inference is simpler since we do not need to compute $f(X_i)$ on the labeled data. It follows that PPI is more cost-effective than classical inference if and only if

$$c_Y\left(1-\tilde{\rho}^2+\gamma\tilde{\rho}^2+2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}\right) \le c_Y + c_X.$$

Furthermore, when this condition is satisfied, the optimal *absolute* cost savings from PPI are

$$C_0(n_0) - C(n^\star, N^\star) = n_0(c_X + c_Y(\tilde{\rho}^2 - \gamma\tilde{\rho}^2 - 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)})).$$

The *relative* cost savings from PPI are

$$\frac{C_0(n_0) - C(n^\star, N^\star)}{C_0(n_0)} = \frac{c_X}{c_Y} + \tilde{\rho}^2 - \gamma\tilde{\rho}^2 - 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}.$$

In the context of mixed subjects experiments, it is natural to take $c_X = 0$. This is because c_X is simply the cost of recording demographic information or treatment assignments which is small compared to the cost of the full survey. When $c_X = 0$, we have $\gamma = c_f/c_Y$ and PPI is strictly more cost efficient than classical inference if and only if

$$\tilde{\rho}^2 - \gamma\tilde{\rho}^2 - 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)} > 0.$$

This is equivalent to

$$\tilde{\rho}^2 > \frac{4\gamma}{(1+\gamma)^2}.$$

That is, the PPI correlation must be sufficiently high compared to the relative cost of collecting a labeled or unlabeled sample. When this condition is met, the expressions for the absolute and relative cost reductions simplify and become

$$\begin{aligned} C_0(n_0) - C(n^\star, N^\star) &= n_0 c_Y(\tilde{\rho}^2 - \gamma\tilde{\rho}^2 - 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}, \\ \frac{C_0(n_0) - C(n^\star, N^\star)}{C_0(n_0)} &= \tilde{\rho}^2 - \gamma\tilde{\rho}^2 - 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}. \end{aligned}$$

Furthermore, the ratio of $C(n^\star, N^\star)$ over $C_0(n_0)$ simplifies to

$$\frac{C(n^\star, N^\star)}{C_0(n_0)} = 1 - \tilde{\rho} + \gamma\tilde{\rho} + 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}. \tag{14}$$

The curves in Figure 4 plot equation (14) as a function of $1/\gamma = c_Y/c_f$ for different values of $\tilde{\rho}$.

A.4 Power Analysis for the Most Powerful Pair

The power analysis for the most powerful pair identifies the pair of sample sizes $(n^\star, N^\star)$ the achieves the smallest standard error subject to a budget constraint. Once the most powerful pair has been computed, the standard error of the PPI estimate can be approximated. Likewise, we can estimate the power of a PPI hypothesis test that uses $(n^\star, N^\star)$.

The inputs required to compute the most powerful pair are the PPI correlation $\tilde{\rho}$, the costs c_Y, c_f, c_X defined above, and a budget B. The PPI correlation would be estimated from data and the costs and budget must be specified by the user. Once these inputs have been provided, $\gamma = (c_X + c_f)/c_Y$ can be computed.

The previous section determined when classical inference was more cost-effective than PPI. When classical inference is more cost-effective, the most powerful pair results from spending all the budget on labeled samples and using no unlabeled samples. Therefore, if

$$c_Y\left(1 - \tilde{\rho}^2 + \gamma\tilde{\rho}^2 + 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}\right) > c_Y + c_X,$$

then

$$n^\star = \frac{B}{c_X + c_Y} \quad \text{and} \quad N^\star = 0.$$

If PPI is more cost-effective than classical inference, then the budget should be allocated so that $C(n^\star, N^\star) = B$ where $n^\star, N^\star$ and $C(n^\star, N^\star)$ are as in equations (12) and (13). This means that if

$$c_Y\left(1 - \tilde{\rho}^2 + \gamma\tilde{\rho}^2 + 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}\right) \leq c_Y + c_X,$$

then

$$n^\star = n_0\left(1 - \tilde{\rho}^2 + \sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}\right) \quad \text{and} \quad N^\star = \frac{n^\star(n_0 - n^\star)}{n^\star - (1-\tilde{\rho}^2)n_0},$$

where

$$n_0 = \frac{B}{c_Y\left(1 - \tilde{\rho}^2 + \gamma\tilde{\rho}^2 + 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}\right)}.$$

Once the most powerful pair $(n^\star, N^\star)$ has been computed, the standard error of the PPI estimate from using $(n^\star, N^\star)$ and the power of using $(n^\star, N^\star)$ can also be approximated. The dataset used to estimate $\tilde{\rho}$ can also be used to estimate σ. Given σ, the standard error of $\hat{\theta}^{\text{PP}}$ when using $(n^\star, N^\star)$ is $S(n^\star, N^\star) = \frac{\sigma}{\sqrt{n^\star}}\sqrt{1 - \frac{N^\star}{N^\star + n^\star}\tilde{\rho}^2}$.

To estimate the power of testing the null hypothesis $\theta = \theta_0$, we can use the supplied dataset to estimate an effect size $\delta = \hat{\theta}^{\text{PP}} - \theta_0$. The power of using $(n^\star, N^\star)$ in a level α test of the hypothesis $\theta = \theta_0$ is

$$1 - \beta = 1 - \Phi\left(z_{1-\alpha/2} - \frac{\delta}{S(n^\star, N^\star)}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\delta}{S(n^\star, N^\star)}\right), \tag{15}$$

where Φ is the cumulative distribution function of the standard normal distribution and z_p is the pth

quantile of the standard normal distribution.

A.5 Power Analysis for the Cheapest Pair

The power analysis for the cheapest pair identifies a pair of sample sizes $(n^\star, N^\star)$ such that achieves a desired standard error of size at most ε. Once the cheapest pair has been computed, the cost of the experiment can be calculated.

The inputs required to compute the cheapest pair are the PPI correlation $\tilde{\rho}$, the standard deviation of the classical estimator σ, the costs c_Y, c_f, c_X and the desired standard error $\varepsilon > 0$. As before, the parameters $\tilde{\rho}$ and σ are estimated from a dataset and c_Y, c_f, c_X and ε are provided by the user. Again, set $\gamma = (c_X + c_f)/c_Y$.

When classical inference is more cost-effective than PPI, the cheapest pair will use only labeled samples. This means that if

$$c_Y\left(1 - \tilde{\rho}^2 + \gamma\tilde{\rho}^2 + 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}\right) > c_Y + c_X,$$

then

$$n^\star = \frac{\sigma^2}{\varepsilon^2} \quad \text{and} \quad N^\star = 0.$$

When PPI is more cost-effective than classical inference, then $(N^\star, n^\star)$ should be chosen so that $S(n^\star, N^\star) = \varepsilon$ where $n^\star$ and $N^\star$ are as in equation (12) and hence $S(n^\star, N^\star) = \frac{\sigma}{\sqrt{n_0}}$. Thus, if

$$c_Y\left(1 - \tilde{\rho}^2 + \gamma\tilde{\rho}^2 + 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}\right) \leq c_Y + c_X,$$

then

$$n^\star = n_0\left(1 - \tilde{\rho}^2 + \sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}^2)}\right) \quad \text{and} \quad N^\star = \frac{n^\star(n_0 - n^\star)}{n^\star - (1-\tilde{\rho}^2)n_0},$$

where

$$n_0 = \frac{\sigma^2}{\varepsilon^2}.$$

Once $(n^\star, N^\star)$ have been computed, the cost of $(n^\star, N^\star)$ can be computed. The cost varies depending on whether classical inference or PPI was used. When classical inference is used, the cost is

$$C_0(n^\star) = (c_Y + c_X)n^\star.$$

When PPI is used, the cost is

$$C(n^\star, N^\star) = (c_Y + c_X + c_f)n^\star + (c_X + c_f)N^\star.$$

The user may wish to provide a level of power, $1 - \beta$, instead of a desired standard error ε. To compute the cheapest pair in this setting, the user must provide or estimate the difference in the alternative and null parameters $\delta = \theta_{H_a} - \theta_{H_0}$. The user must also provide a level of the test α. Once these have been provided, the desired standard error $\varepsilon > 0$ is the solution to the equation

$$1 - \beta = 1 - \Phi\left(z_{1-\alpha/2} - \frac{\delta}{\varepsilon}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\delta}{\varepsilon}\right),$$

where Φ and z_p are as in equation (15). Once ε is known, the power analysis for the cheapest pair proceeds as above.

B Details on the Moral Machine Experiment

B.1 Example of a Moral Dilemma

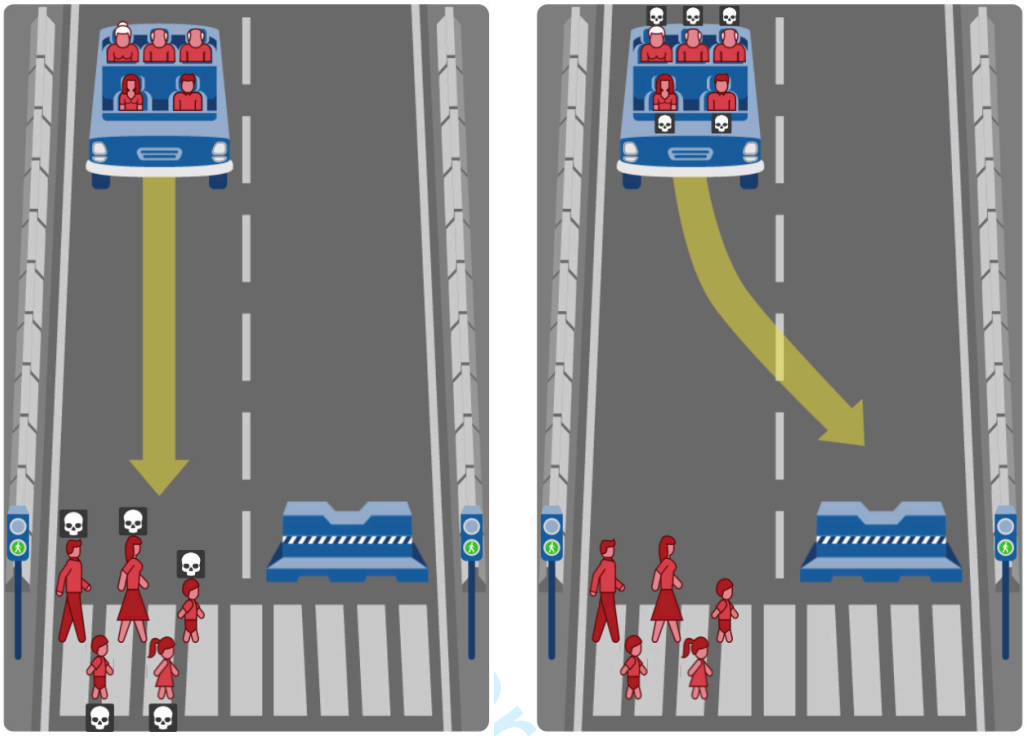

Figure S1: Example dilemma presented to a survey respondent in the Moral Machine Experiment. The image was created on the *moralmachine.net* website with information from the replication data for Awad et al. (2018).

B.2 Example Prompt

The LLM was given the following prompt to predict the survey respondent's decision to the moral dilemma. 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.

B.3 Summary Statistics on Sampling

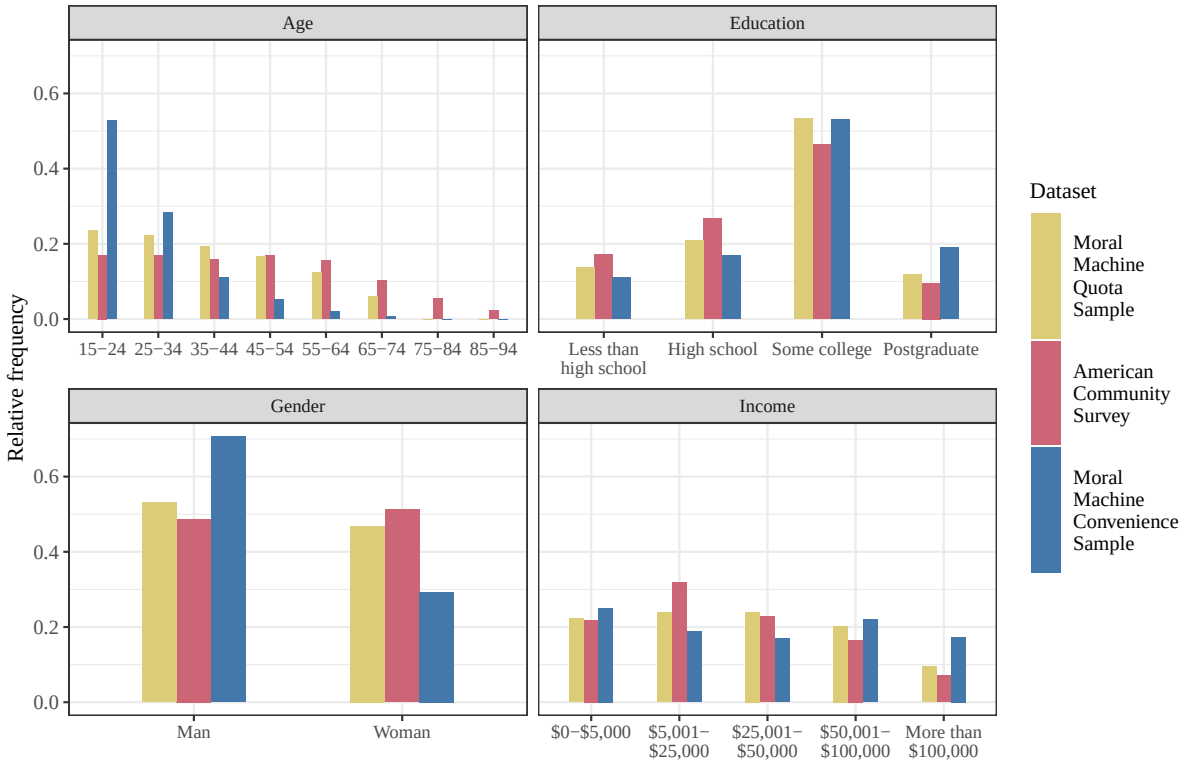

Figure S2: Comparison of demographic distributions of the convenience sample reported in the Moral Machine Experiment, the 2016 American Community Survey, and the random sample from the Moral Machine Experiment with quotas from the American Community Survey.

B.4 Replication of AMCE Estimates

Figure S3 compares the estimated causal effects of scenario attributes on the decision to save characters across samples. Compared to the quota 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).

For completeness, we also show the estimated AMCEs reported in Awad et al. (2018). Although their estimates are based on a cross-country convenience sample and are not expected to align perfectly with those from the quota sample of Americans, the differences are generally small.

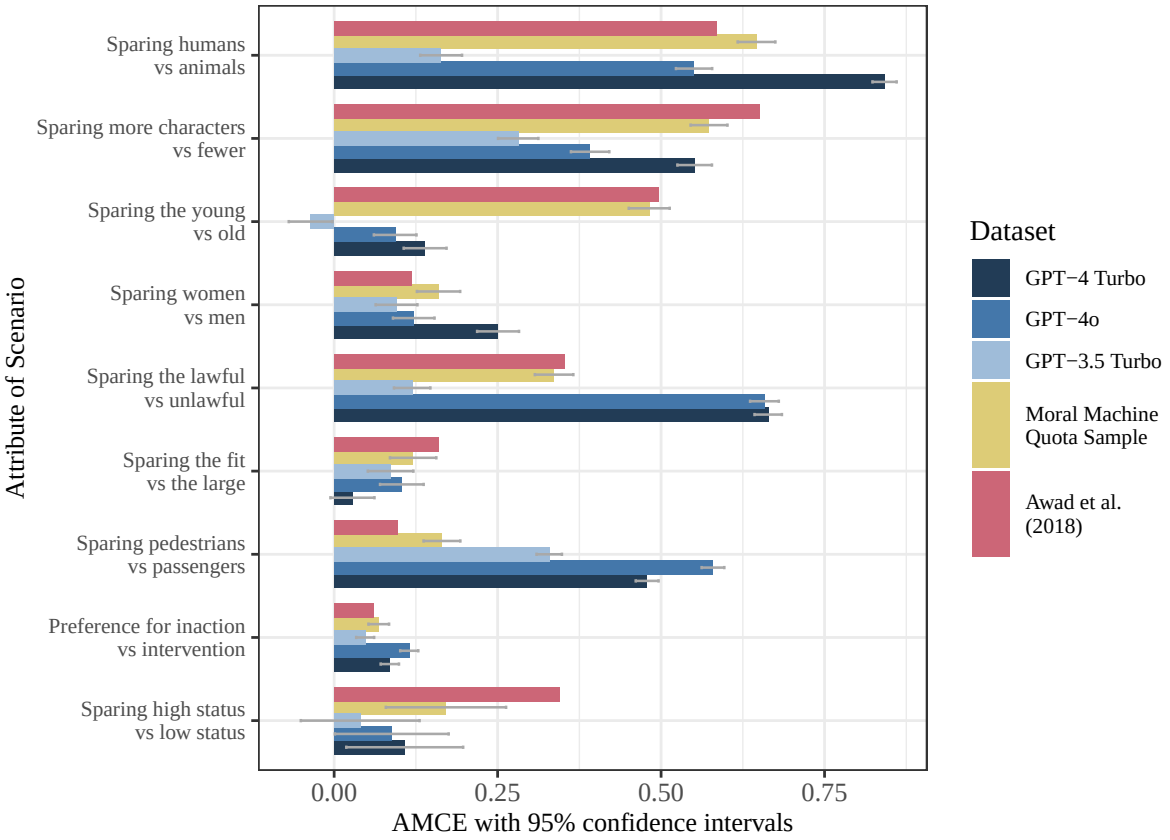

Figure S3: Comparison of AMCE estimates from human subjects against silicon subjects. Note that Awad et al. (2018) do not report confidence intervals due to their negligible width, a consequence of the large sample size.

B.5 Prompting LLMs to Predict Decisions to Moral Dilemmas

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 S1: Details on LLMs used to predict survey responses

Model	Type	Correlation	N
GPT4 Turbo	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
GPT3.5 Turbo	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 S2: 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.

C Example Power Analysis

We used G*Power, a software by Faul et al. (2009), to conduct the power analysis from section 2.1.

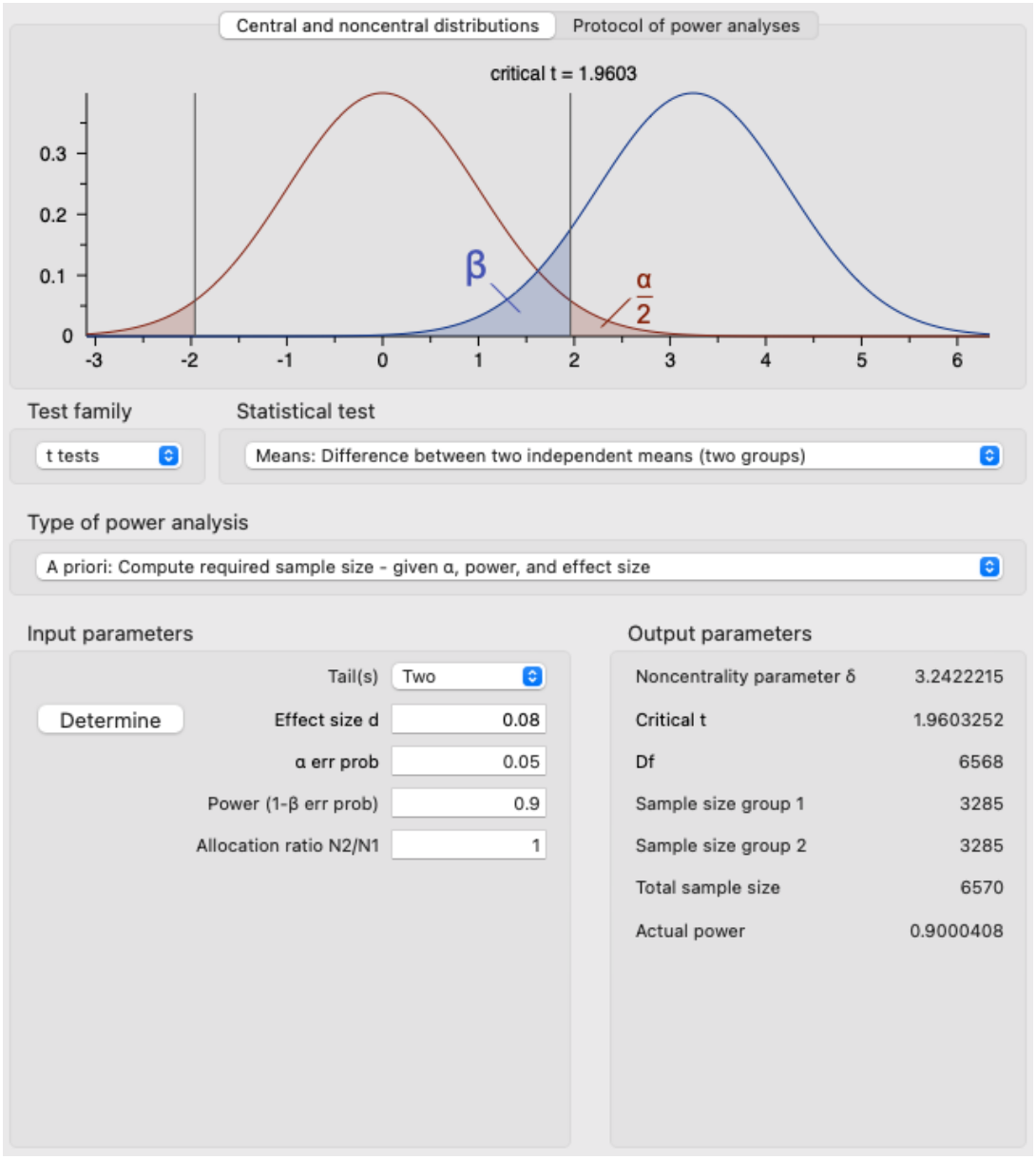

Figure S4: Power analysis with G*Power