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

David Broska, Michael Howes, and Austin van Loon

Abstract: Large Language Models (LLMs) provide cost-effective but possibly inaccurate predictions of human behavior. Despite growing evidence that predicted and observed behavior are often not *interchangeable*, there is limited guidance on using LLMs to obtain valid estimates of causal effects and other parameters. We argue that LLM predictions should be treated as potentially informative observations, while human subjects serve as a gold standard in a *mixed subjects design*. This paradigm preserves validity and offers more precise estimates at a lower cost than experiments relying exclusively on human subjects. We demonstrate—and extend—prediction-powered inference (PPI), a method that combines predictions and observations. We define the *PPI correlation* as a measure of interchangeability and derive the *effective sample size* for PPI. We also introduce a power analysis to optimally choose 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.

Keywords: Mixed Subjects Design, Prediction-Powered Inference, PPI Correlation, Effective Sample Size, PPI Power Analysis, Machine Learning, Large Language Models, 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. This approach to learning about social phenomena, known as "silicon sampling" (Argyle et al., 2023), could 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 rely, either implicitly or explicitly, on what we term the *interchangeability assumption*—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 yields 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 statistical analyses. Currently, the interchangeability of predicted and observed behavior is assessed empirically on a case-by-case basis. As a result, silicon sampling offers 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; Sarstedt et al., 2024).

To address this gap, we propose a *mixed subjects design* for research with LLMs. We suggest leveraging observations from human subjects as a gold standard while treating LLM predictions as potentially informative for estimating parameters. Instead of rejecting the assumption that human behavior and LLM predictions are interchangeable a priori, we argue that this assumption should be assessed empirically. By sampling both human subjects and LLM predictions, this assessment determines the extent to which LLMs can inform parameter estimates. We demonstrate how to implement this approach with prediction-powered inference (PPI) (Angelopoulos et al., 2023, 2024). PPI allows researchers to combine observations of human behavior with predictions generated by LLMs or other algorithms. Unlike silicon sampling and simple data imputation methods, PPI does not assume the predictions to be accurate

(Hoffman et al., 2024). Instead, predictions are considered potentially informative but imperfect proxies for gold standard data. PPI produces valid point estimates with narrower confidence intervals compared to those derived solely from human subjects. Therefore, the mixed subjects design with PPI may offer many benefits of silicon sampling while avoiding the drawbacks that limit its use in confirmatory research.

LLMs introduce a trade-off between predicted and observed behavior when estimating parameters such as causal effects. While obtaining predictions from LLMs is less expensive than recruiting human subjects, these predictions are less informative for estimating parameters than directly observed behavior. We define the PPI correlation $\tilde{\rho}$ as a measure of interchangeability between predicted and observed behavior. Based on the PPI correlation we derive the effective sample size of PPI, quantifying the extent to which LLMs can increase statistical precision by augmenting the sample of human subjects. We also use the PPI correlation in our power analysis, which is our main statistical innovation and extension of PPI. The PPI power analysis optimally balances the costs of collecting predicted and observed behavior with the extent to which these two types of data inform inferences on parameters. One possibility is 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 have been integrated into the PPI Python library, available at `https://github.com/aangelopoulos/ppi_py`.

2 The Silicon Subjects Design

Experiments in surveys, labs, and the field enhance our understanding of causal processes in the social sciences. However, experiments also face limitations, such as high research costs, difficulties in recruiting participants from hard-to-reach populations, and challenges related to measurement and generalizability. Below, we outline how the silicon subjects approach promises to address these issues and highlight 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. The prompt often includes a persona with demographics, attitudes, and other information about the participant. This persona has two purposes in silicon sampling. First, providing relevant information about the participant may enhance the LLM's capacity to mimic human behavior. Second, personas allow researchers to target specific populations, such as eligible voters in the U.S. (Argyle et al., 2023; Bisbee et al., 2024). The composition of the silicon sample is intended to match that of the target population, at least on key demographics. In the

context of experiments, the prompt also includes a description of the condition assigned to the silicon subject. Based on the persona and information about the experimental condition, the LLM predicts the participant's response as if they had participated in the experiment. Silicon sampling aims to estimate the same quantities as studies that rely on human subjects—including point estimates, such as regression coefficients, and measures of statistical precision, such as standard errors (Bisbee et al., 2024). Given successful replications of human subjects studies using the silicon subjects design, some scholars have concluded that LLM predictions can be interchangeable with human behavior:

> *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, 2023:9)

> *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)

> *This paper explores the potential of large language models (LLMs) to substitute for human participants in market research. . . . The paper demonstrates that, for some categories, this new method of fully or partially automated market research will increase the efficiency of market research by meaningfully speeding up the process and potentially reducing the cost.* (Li et al., 2024:254)

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. Survey experiments in the social sciences require large numbers of survey participants to identify typically small effects. 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 S5 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 (e.g., Milkman et al., 2021) 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 (see Table S1 in the Supporting Information).

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 their target population, certain participants remain difficult to recruit through panels for online research (Chandler et al., 2019). For example, typical online panels for survey research consist of younger, more liberal, and more educated respondents who 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 data on these otherwise hard-to-reach populations, providing nearly instant access to diverse participant profiles. Silicon subjects may even serve as alternative study populations when 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 its measurements. 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 strictly follows the researchers' requirements. Such a silicon subject can respond to hundreds of questions rapidly and reliably without experiencing fatigue (Dillion et al., 2023). Similarly, advantages may also be seen in unrealistic statistical properties of LLM predictions. For example, predictions of human responses exhibit less variation than human responses (Park et al., 2024; Bisbee et al., 2024; Mei et al., 2024). Proponents of the silicon sampling approach could argue that this reduced variation allows for a more precise measurement of central tendencies such as the mean (cf. Section 2.2). This property does not imply that researchers obtain an unbiased parameter estimate, but that this estimate has less statistical uncertainty than an estimate from a sample of human subjects.

2.2 Perils of the Silicon Subjects Design

LLMs have been shown to inaccurately predict human behavior, casting doubt on whether the silicon sampling design can yield valid conclusions about parameters. Estimates of regression coefficients computed on silicon samples have been shown to be biased—i.e., differ from those computed for human subjects. Specifically, relying on LLM predictions may lead researchers to overestimate effects (Ashokkumar et al., 2024) or even reach qualitatively different conclusions, as the coefficients can have opposite signs when compared to those from human samples (Bisbee et al., 2024).

Even if estimates based on LLM predictions are accurate for some groups of study participants, research

designs, and types of behaviors, LLMs may be less accurate in other contexts (Messeri and Crockett, 2024; Bisbee et al., 2024). 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. Similarly, Alvero et al. (2024) observe that LLMs, when compared to actual college applicants, write college admissions essays most similarly to those who are male and from affluent neighborhoods. At a more fundamental level, it remains unclear in which contexts LLM predictions can be trusted without validation against human respondents—a treatment effect estimated based solely on LLM predictions may not replicate in studies with human subjects (Harding et al., 2023).

Parameter estimates based on LLM predictions may not only be different from those estimated using human samples—they may also be misleadingly precise. One source of concern is the limited variability in LLM predictions compared to human responses (Park et al., 2024; Bisbee et al., 2024; Mei et al., 2024). As a result, standard errors for coefficients may be too small compared to those computed from human samples. Moreover, the availability of inexpensive and near-instantly available data may give rise to questionable research practices, exacerbating the problem of misleading precision. Analyzing large numbers of predicted values from LLMs results in precise parameter estimates simply because standard errors decrease with sample size. This issue parallels the analysis of Big Data from non-representative samples (Meng, 2018; Bradley et al., 2021), where researchers risk being "precisely inaccurate" (McFarland and McFarland, 2015). Coupled with even small biases in the LLM predictions, smaller standard errors imply narrower confidence intervals with incorrect centers and significant p-values for biased point estimates.[2] Therefore, the silicon sampling design could further amplify doubts about the replicability of findings from experimental social science (e.g., 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.

Silicon sampling may have limited utility even in pilot studies. Researchers often use pilot studies to anticipate the direction and size of effects in experiments. When the variability in the outcome is lower, smaller sample sizes are sufficient to detect effects with a desired level of statistical power. As such, the reduced variability in LLM predictions may lead researchers to underestimate the number of human subjects required to identify an effect (Bisbee et al., 2024). Using silicon samples for pilot studies can therefore lead researchers to pursue experiments with underpowered hypothesis tests.

3 The Mixed Subjects Design

We propose the mixed subjects design, an umbrella term for statistical methods that provide valid inferences about human behavior while including 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 subjecting it to potential disconfirmation via empirical evidence. In this approach, human respondents count as a gold standard and are used to empirically assess the interchangeability of LLM predictions with human subjects (Figure 1). Our approach helps build confidence in LLMs as a research tool by combining human and silicon subjects through statistical methods that produce valid parameter estimates while increasing statistical precision by leveraging low-cost LLM predictions.

In this section, we first describe prediction-powered inference (PPI) (Angelopoulos et al., 2023, 2024), a recent statistical framework that instantiates a mixed subjects approach. We then build on PPI in two ways. First, we define the PPI correlation, an intuitive metric for the degree to which human and silicon subjects are interchangeable in a particular setting. Second, we introduce the PPI power analysis, a tool researchers can use to find the optimal combination of human and silicon subjects given either a set research budget or a desired level of statistical power. We end this section by discussing how to implement a mixed subjects design in practice.

3.1 Introduction to Prediction-Powered Inference

PPI is a general statistical framework for combining a dataset of gold standard observations with predictions to estimate a broad class of parameters (Angelopoulos et al., 2024). For example, researchers can estimate the extent of deforestation in the Amazon rainforest by combining a machine learning model trained to predict forest cover from satellite images. The underlying intuition is that the prediction algorithm captures information about deforestation, allowing for a more precise estimation of forest loss than using gold standard measurements of forest cover alone (Angelopoulos et al., 2023). Information from prediction algorithms can be extracted through PPI and used to draw more precise inferences about population means, regression coefficients, and other parameters.

Before delving into a more technical primer, we highlight two key properties of PPI. First, the validity of the PPI estimator does not depend on the accuracy or unbiasedness of the prediction algorithm. Even when predictions provide minimal information about the outcomes, PPI can leverage these predictions to produce unbiased estimates and valid confidence intervals for the parameter of interest. The second property is that PPI estimates are always at least as precise as classical estimates based solely on gold standard measurements. When the predictions are informative, PPI provides more precise estimates than methods that do not incorporate predictions. Furthermore, the precision of PPI estimates is higher for more informative prediction algorithms.

How does PPI achieve these two properties? PPI adjusts an estimator that only uses gold standard

observations. The form of the adjustment prevents the algorithm from introducing bias. To improve the precision of the estimate, PPI optimizes the magnitude of this adjustment. The optimization sets a lower bound on precision, equal to the amount of precision attained without the use of predictions. We provide a more technical explanation of PPI in the following section.

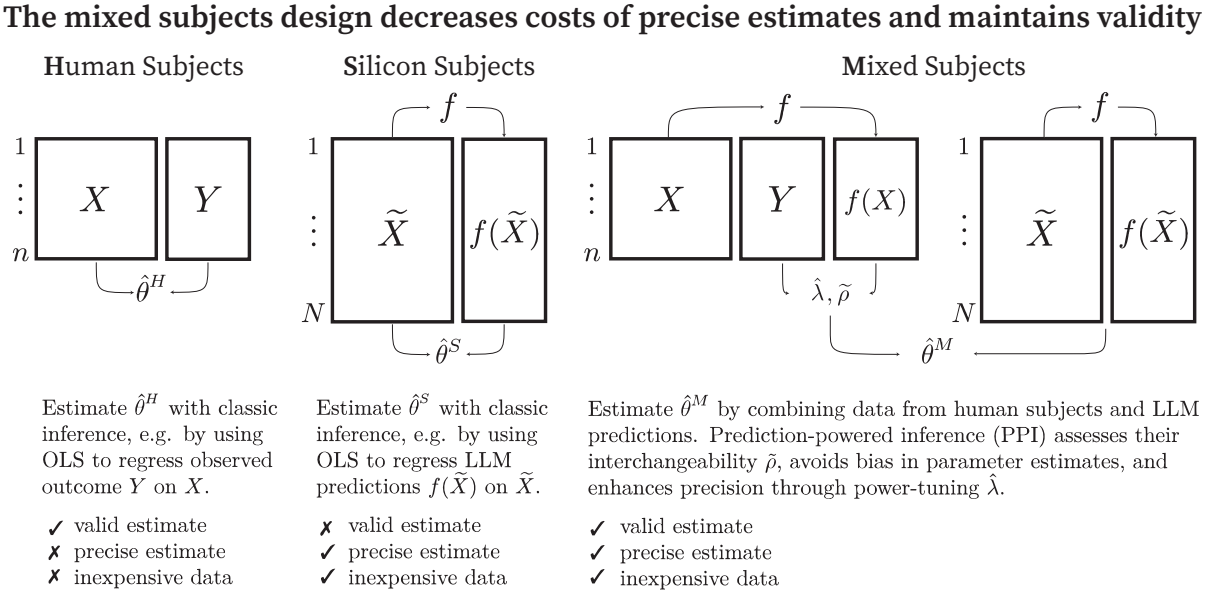

Figure 1: Comparison of experiments with human, silicon, and mixed subjects designs

3.2 Estimating Parameters with Prediction-Powered Inference

To estimate a parameter with PPI, researchers need two independent datasets: a labeled dataset $\{(X_i, Y_i)\}_{i=1}^n$ and an unlabeled dataset $\{\widetilde{X}_i\}_{i=1}^N$. Both are assumed to be independent and identically distributed (i.i.d.) samples. Furthermore, it is assumed that X_i and $\widetilde{X}_i$ are from the same population with the same data-generating process. The variable X_i is the input to an algorithm f for predicting Y_i. PPI requires that the prediction algorithm f is independent of both the labeled and unlabeled datasets. In particular, the datasets should not be used to train f.[3]

In the following, we illustrate how PPI estimates the population mean of Y_i because the formulas for mean estimation are more manageable. That said, the intuition extends to more complex parameters, such as regression coefficients.[4] As a point of reference, we will also discuss the estimators from the human subjects and silicon subjects designs.

Let $\theta^\star = \mathbb{E}[Y_i]$ be the population mean of Y_i, such as the average student test score, loan amount, or time spent on social media. For a human subjects study, the human subjects estimator $\hat{\theta}^H$ of $\theta^\star$ is the sample mean of $\{Y_i\}_{i=1}^n$:

$$\hat{\theta}^H = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

The estimator $\hat{\theta}^H$ gives an unbiased estimate of the population mean of Y_i and has variance $\frac{1}{n}\text{Var}(Y_i)$. For large sample sizes n, $\hat{\theta}^H$ is approximately normally distributed. This approximation can be used to construct confidence intervals for $\theta^\star$.

In contrast, the silicon subjects estimator $\hat{\theta}^S$ uses the unlabeled dataset $\{\widetilde{X}_i\}_{i=1}^N$, and assumes the LLM prediction $f(\widetilde{X}_i)$ are interchangeable with a human response Y_i. The silicon subjects estimator of $\theta^\star$ is then the sample mean of the LLM predictions:

$$\hat{\theta}^S = \frac{1}{N}\sum_{i=1}^{N} f(\widetilde{X}_i)$$

The silicon subjects estimator has a variance of $\frac{1}{N}\,\text{Var}(f(\widetilde{X}_i))$. When the size of the unlabeled dataset N is much larger than the size of the labeled data n, $\hat{\theta}^S$ has much smaller variance than the human subjects estimator $\hat{\theta}^H$. Consequently, confidence intervals constructed from $\hat{\theta}^S$ will be narrower than those based on $\hat{\theta}^H$. However, $\hat{\theta}^S$ may be biased, which means that its confidence intervals may not be centered around the true parameter $\theta^\star$ and thus fail to achieve the nominal coverage probability of $1-\alpha$.

Unlike silicon sampling, the PPI estimator prevents the prediction algorithm f from introducing bias in the parameter estimate. The PPI estimator for the population mean is

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

where λ is a tuning parameter to be discussed later. The estimator $\hat{\theta}^{\text{PP}}_\lambda$ satisfies a central limit theorem provided both sample sizes n and N are large. The central limit theorem can be used to create confidence intervals for $\theta^\star$ and to perform hypothesis tests.

The form of the PPI estimator ensures unbiased estimates of $\theta^\star$ for any prediction algorithm f and any value of λ. The first term in $\hat{\theta}^{\text{PP}}_\lambda$ is the unbiased human subjects estimator $\hat{\theta}^H$ and the second term is the difference between two silicon subject estimators. Crucially, the silicon subjects estimators do not introduce bias to $\hat{\theta}^H$ because X_i and $\widetilde{X}_i$ are sampled from the same distribution. If the two silicon subjects estimators in (1) are biased, then their bias is the same and cancels out on average. Consequently, the distribution of $\hat{\theta}^{\text{PP}}_\lambda$ is always centered at the same parameter targeted by the human subjects estimator:

$$\mathbb{E}[\hat{\theta}^{\text{PP}}_\lambda] = \mathbb{E}[\hat{\theta}^H] = \theta^\star \tag{2}$$

The PPI estimator not only remains unbiased but is also designed to be more precise than the human subjects estimator. This increase in precision arises because PPI leverages the unlabeled dataset $\{\widetilde{X}_i\}_{i=1}^N$ and the information that f has about the relationship between X_i and Y_i. PPI incorporates this information through a statistical technique that Angelopoulos et al. (2024) refer to as *power tuning*.

Specifically, the parameter λ is tuned to minimize the variance of $\hat{\theta}^{\text{PP}}_{\lambda}$. When λ is chosen with power tuning, the PPI estimator achieves at least the same level of precision as the human subjects estimator (Angelopoulos et al., 2024, Section 6). To explain power tuning, we use an analogy with linear regression and the analysis of variance. We show in Section S1.1 of the Supporting Information that $\hat{\theta}^{\text{PP}}_{\lambda}$ in (1) can be rewritten such that the estimator takes the form of an average of residuals:

$$\hat{\theta}^{\text{PP}}_{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - \lambda(f(X_i) - \hat{\theta}^{S})\Big) \tag{3}$$

where Y_i is the dependent variable, $f(X_i) - \hat{\theta}^{S}$ is the independent variable and $\lambda(f(X_i) - \hat{\theta}^{S})$ are the fitted values. If $f(X_i)$ is at all correlated with Y_i, then there is a choice of λ such that the residuals $Y_i - \lambda(f(X_i) - \hat{\theta}^{S})$ have lower variance than Y_i. For the population mean, the power-tuned estimate of λ is

$$\hat{\lambda} = \frac{N}{n+N}\frac{\widehat{\text{Cov}}(Y_i, f(X_i))}{\widehat{\text{Var}}(f(X_i))}. \tag{4}$$

The ratio $\frac{\widehat{\text{Cov}}(Y_i, f(X_i))}{\widehat{\text{Var}}(f(X_i))}$ is the least squares regression coefficient obtained from regressing Y_i on $f(X_i)$. When $f(X_i)$ does not explain much variation in Y_i, then $\widehat{\text{Cov}}(Y_i, f(X_i))$ will be close to zero and $\hat{\theta}^{\text{PP}}_{\hat{\lambda}}$ will be close to the human subjects estimator. Conversely, when $f(X_i)$ is very predictive of Y_i, then most of the variance in Y_i is explained by $f(X_i) - \hat{\theta}^{S}$ and $\hat{\theta}^{\text{PP}}_{\hat{\lambda}}$ will have a much lower variance than $\hat{\theta}^{H}$.

The fraction $\frac{N}{n+N}$ in (4) is a shrinkage factor for the regression coefficient. This factor arises because the silicon subjects estimator $\hat{\theta}^{S}$ adds uncertainty and variation to $\hat{\theta}^{\text{PP}}_{\lambda}$. This uncertainty decreases with N, which is reflected in the size of the shrinkage factor. When the unlabeled dataset is much larger than the labeled dataset ($N \gg n$), then $\frac{N}{n+N}$ is close to 1, giving greater weight to the unlabeled data. Conversely, if $N \leq n$, then $\frac{N}{n+N}$ is at most $\frac{1}{2}$, assigning more weight to the labeled data and making $\hat{\theta}^{\text{PP}}_{\hat{\lambda}}$ more closely resemble the human subjects estimator $\hat{\theta}^{H}$. In summary, PPI uses power tuning to increase statistical precision by balancing the information from both datasets, considering the relative sizes of the datasets and the extent to which $f(X_i)$ correlates with Y_i.

The intuition and statistical properties from mean estimation with PPI extend to estimating other parameters. Angelopoulos et al. (2024) define estimators $\hat{\theta}^{\text{PP}}_{\lambda}$ for generalized linear models, e.g. coefficients in linear or logistic regression. Similar to (1), these estimators evaluate f on both the labeled and unlabeled datasets to ensure that the $\hat{\theta}^{\text{PP}}_{\lambda}$ is targeting the same parameter as the human subjects estimator. Angelopoulos et al. (2024) show that $\hat{\theta}^{\text{PP}}_{\lambda}$ converges to the true regression coefficients and satisfies a multivariate central limit theorem for large sample sizes n and N. As a result, confidence intervals based on $\hat{\theta}^{\text{PP}}_{\lambda}$ achieve the nominal coverage of $1 - \alpha$.

The PPI estimator for a generalized linear model is not only valid but also designed to be more precise than the corresponding human subjects estimator. As in mean estimation, this increase in precision

results from power tuning λ. This tuning parameter is chosen to minimize the variance of a particular coefficient in the vector $\hat{\theta}^{\text{PP}}_{\lambda}$. The estimation of λ is again analogous to mean estimation, resulting in a tuned value of $\hat{\lambda}$ similar to (4). One difference is that for regression coefficients, the tuning parameter λ is restricted to be between 0 and 1. However, even with the restriction that $\hat{\lambda} \in [0, 1]$, the final estimator $\hat{\theta}^{\text{PP}}_{\hat{\lambda}}$ is at least as precise as the human subjects estimator $\hat{\theta}^{H}$ (Angelopoulos et al., 2024, Section 6).

The Python library `ppi_py` implements PPI with power tuning. To compute PPI point estimates or confidence intervals, the user needs to supply the labeled dataset $\{(X_i, Y_i, f(X_i))\}_{i=1}^{n}$ and the unlabeled dataset $\{(\widetilde{X}_i, f(\widetilde{X}_i))\}_{i=1}^{N}$. We use this implementation in all of our examples. The package `ipd` includes an implementation of PPI with power tuning in R (Salerno et al., 2024).

3.3 The PPI Correlation and the Effective Sample Size

In this section, we extend the ideas of Angelopoulos et al. (2024) to define the *PPI correlation*, a measure of the interchangeability of human and silicon subjects. We show that higher values of the PPI correlation correspond to increased precision of parameter estimates. Additionally, we propose that the limitations of silicon subjects should be reflected in the sample size used for inference about parameters. While silicon sampling and simple data imputation methods assume a pooled sample size of $n + N$, the mixed subjects design with PPI uses an *effective sample size*. The effective sample size adjusts the pooled sample size when the silicon subjects and human subjects are not fully interchangeable.

The PPI correlation, $\tilde{\rho}$, is the asymptotic correlation between the human subjects estimator $\hat{\theta}^{H}$ and the silicon subjects estimator $\hat{\theta}^{S}$ as n approaches infinity. In the case of mean estimation, $\tilde{\rho}$ is exactly the correlation between Y_i and $f(X_i)$, while for other parameters its definition of the PPI correlation is more complicated (see equation (S9) in the Supporting Information).

The PPI correlation affects the precision of the power-tuned PPI estimator. Specifically, let $\hat{\theta}^{\text{PP}} = \hat{\theta}^{\text{PP}}_{\hat{\lambda}}$ be the power-tuned estimator for a coefficient in a vector of regression coefficients. Its standard error, $\text{se}(\hat{\theta}^{\text{PP}})$, quantifies the precision of $\hat{\theta}^{\text{PP}}$. The standard error determines the width of the PPI confidence intervals and influences the statistical power of a hypothesis test with PPI. Note that in PPI, each of X_i, Y_i, and $\widetilde{X}_i$ are treated as random. Therefore, the standard error of the PPI estimator is its standard deviation under hypothetical resampling of n labeled data points and N unlabeled data points. In the Supporting Information, we show that the standard error of $\text{se}(\hat{\theta}^{\text{PP}})$ can be written as

$$\text{se}(\hat{\theta}^{\text{PP}}) = \text{se}(\hat{\theta}^{H})\sqrt{1 - \frac{N}{n+N}\tilde{\rho}_{+}^{2}} \tag{5}$$

where $\text{se}(\hat{\theta}^{H})$ is the standard error of the human subjects estimator applied to the labeled dataset, n and N are the sizes of the labeled and unlabeled datasets respectively, and $\tilde{\rho}_{+} = \max\{0, \tilde{\rho}\}$ is the non-

negative part of the PPI correlation. Equation (5) shows that the PPI estimator is always at least as precise as the human subjects estimator $\hat{\theta}^H$ since $\text{se}(\hat{\theta}^H)$ is multiplied by a factor that is no larger than 1. Furthermore, if $N > 0$ and $\tilde{\rho} > 0$, then the PPI estimator will have a strictly smaller standard error than the human subjects estimator.

The PPI correlation determines the effective sample size of PPI. The effective sample size is the number, n_0, of human subjects necessary to achieve the same standard error as PPI. That is, the estimator from a mixed subjects experiment with n human subjects is as precise as the estimator from a human subjects experiment with n_0 subjects. The effective sample size for PPI can be written in terms of n, N and $\tilde{\rho}$

$$n_0 = n \times \frac{n+N}{n+N(1-\tilde{\rho}_+^2)}. \tag{6}$$

This equation shows that $\tilde{\rho}$ is a quantitative measure of the degree to which the interchangeability assumption holds in a mixed subjects experiment. When $\tilde{\rho} = 1$, the effective sample size is $n_0 = n + N$ and the human and silicon subjects are equally informative. When $\tilde{\rho} \leq 0$, the silicon subjects provide no additional information and the effective sample size reduces to the number of human subjects $n_0 = n$. If $\tilde{\rho}$ is strictly between 0 and 1, the silicon subjects are informative but not completely interchangeable. In this case, the effective sample size is strictly between n and $n + N$. Both the PPI correlation and the effective sample size depend on the parameter being estimated. Therefore, a single experiment may have multiple effective sample sizes if there are multiple conditions.

Figure 2 shows that using a larger number of silicon subjects increases the effective sample size (**a**) and reduces the PPI standard error (**b**) with diminishing returns for larger N. The benefits of adding more silicon samples are stronger for higher values of the PPI correlation. To illustrate, suppose that the PPI correlation is $\tilde{\rho} = 0.7$ and the labeled dataset consists of $n = 1{,}000$ human subjects. In a mixed subjects experiment with an additional $N = 10{,}000$ silicon subjects, the effective sample size is $n_0 = 1{,}803$ and the PPI standard error is reduced to 74.5% of the standard error in the human subjects experiment. If instead $\tilde{\rho}$ was 0.9, the effective sample size expands to $n_0 = 3{,}793$ and the PPI standard error is reduced to 51.3% of the human subjects standard error. These considerations about the increases in statistical precision also apply to confidence intervals since the ratio of standard errors is equivalent to the ratio of the width of confidence intervals. We also note that smaller standard errors result in non-linear increases in statistical power.

Researchers will often not have prior information about the PPI correlation. In Section S1.4 of the Supporting Information, we provide a consistent estimator for $\tilde{\rho}$. Our recommendation is that researchers collect a small sample of labeled data $\{(X_i, Y_i, f(X_i))\}_{i=1}^n$ and estimate $\tilde{\rho}$. Researchers can use this estimate to empirically assess the interchangeability assumption and to evaluate the potential for increasing

precision by using PPI on a larger dataset. To help researchers assess these benefits, we have developed a power analysis for PPI, which we describe in the next section.

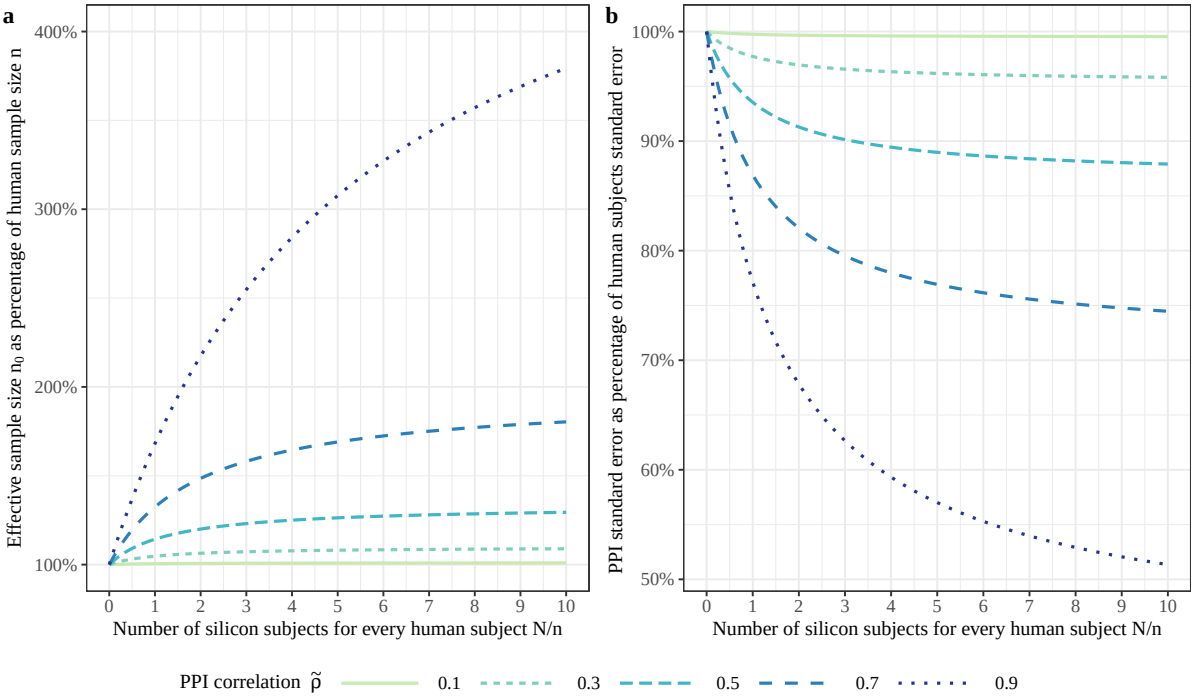

Figure 2: The x-axis shows the ratio N/n of samples sizes with N silicon and n human subjects. Using a larger number of silicon subjects increases the effective sample size (**a**) and reduces the PPI standard error (**b**) with diminishing returns for large N. The ratio of the PPI standard error to the classical standard error in (**b**) is defined by $\sqrt{1-(N/(N+n))\tilde{\rho}^2}$. Adding more silicon samples increases the precision of the PPI estimator more strongly for higher values of the PPI correlation.

3.4 PPI Power Analysis

Power analyses allow researchers to determine the necessary sample size to achieve a desired level of power—i.e., the probability of correctly rejecting the null hypothesis when there is an effect (Cohen, 1988). At present, no such method has been developed for PPI. We introduce our PPI power analysis to complement the toolkit necessary for conducting mixed subjects experiments. The PPI power analysis is based on a trade-off between recruiting *costly but informative* human subjects and *less informative but cheap* silicon subjects. Based on the cost of collecting human and silicon subjects and their interchangeability, our power analysis allows researchers to determine the optimal mix of n human and N silicon subjects.

We highlight that our power analysis is *data-driven*. That is, to run our power analysis, researchers need access to an initial dataset, e.g. from a pilot study, containing covariates X_i, responses Y_i, and predictions $f(X_i)$. This initial dataset is used to estimate the PPI correlation $\tilde{\rho}$. Once $\tilde{\rho}$ is known, our power analysis can be used to find an optimal, larger mixed subjects design with n human samples and N silicon samples. The researcher can then carry out this confirmatory study and be confident that the inference is both statistically valid and sufficiently powered. We do not quantify the uncertainty in estimating $\tilde{\rho}$. If researchers want to account for uncertainty in $\tilde{\rho}$, they may bootstrap our power analysis to get a range of plausible estimates of $\tilde{\rho}$ (Efron and Tibshirani, 1994).

The PPI correlation is one of two inputs to our power analysis. The second is a desired effective sample size n_0. This desired effective sample size should be computed using a power analysis for the human subjects estimator $\hat{\theta}^H$. The desired n_0 depends on the desired level of power $1-\beta$, the significance level α, the hypothesized effect, and the precision of the human subjects estimator. Section S1.6 in the Supporting Information provides a formula specifying how these quantities determine n_0 and Python code to compute n_0. Section S1.6 also explains how to use `ppi_py` to estimate the hypothesized effect size and the precision of the human subjects estimator.

Our power analysis describes all mixed subjects designs with precision equal to that of a human subject design with n_0 human subjects. Such designs are those that satisfy the following two conditions:

$$n_0(1-\tilde{\rho}_+^2) < n \leq n_0 \quad \text{and} \quad N = n \times \frac{n_0 - n}{n - n_0(1-\tilde{\rho}_+^2)}. \tag{7}$$

These conditions imply that a researcher can achieve the desired n_0 with many possible pairs (n, N). Specifically, for every value n that is greater than $n_0(1-\tilde{\rho}_+^2)$ and less than n_0, there is a value of N at which the associated mixed subjects design is as precise as a human subjects design with n_0 human subjects. For example, imagine a setting in which $\tilde{\rho} = 0.7$ and a researcher considers running a human

subjects study with $n_0 = 1{,}000$ participants. Our results show that they could achieve an equally precise estimate in a mixed subjects design by collecting $n = 520$ human subjects and $N = 24{,}960$ silicon subjects. Alternatively, the researcher could collect $n = 900$ human subjects and $N = 231$ silicon subjects. More generally, the researcher could pick any pair of sample sizes satisfying (7) and know that their proposed experiment will have the correct level and power for their proposed effect size.

We provide two methods for researchers to *optimally* pick a design based on the costs of sampling human and silicon subjects. The first design is called the *cheapest pair*. The cheapest pair is the mix of human and silicon subject sample sizes which incurs the lowest cost to achieve a prespecified effective sample size (Figure 3a). To compute the cheapest pair, the researcher needs to provide the costs per human and silicon subject, the PPI correlation $\tilde{\rho}$, and the desired effective sample size. Researchers can use the cheapest pair design if budget constraints are less salient, but resource allocation should still be as efficient as possible.

The second method identifies the design we call the *most powerful pair*. This design maximizes statistical power by identifying the largest possible effective sample size n_0 by optimally combining n and N with a constraint on the total budget of the planned experiment. As illustrated by Figure 3b, multiple pairs of sample sizes (n, N) fall within the budget of an experiment. However, only one such pair can be combined to an effective sample size that maximizes statistical power. Researchers may be particularly interested in finding the most powerful pair (n, N) if a limited research budget is the main constraint and the budget should be allocated to maximize power. To compute the most powerful pair, researchers need to provide the costs per human and silicon subject, the PPI correlation, and the total research budget. A Python implementation of our power analysis is incorporated in the `ppi_py` package.

We conclude this section by emphasizing the importance of power analyses. Power analyses allow researchers to assess the feasibility of an experiment based on a pilot study, considering the required sample size and cost of data collection. Our power analysis determines the most cost-effective mixed subjects design and thus helps researchers assess the feasibility of an experiment. Power analyses also contribute to the credibility of the reported results. 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, thwarts the accumulation of knowledge, and may explain the existence of inconsistent findings (Thye, 2000; Stadtfeld et al., 2020). Our power analysis can help researchers conduct well-powered studies at lower costs.

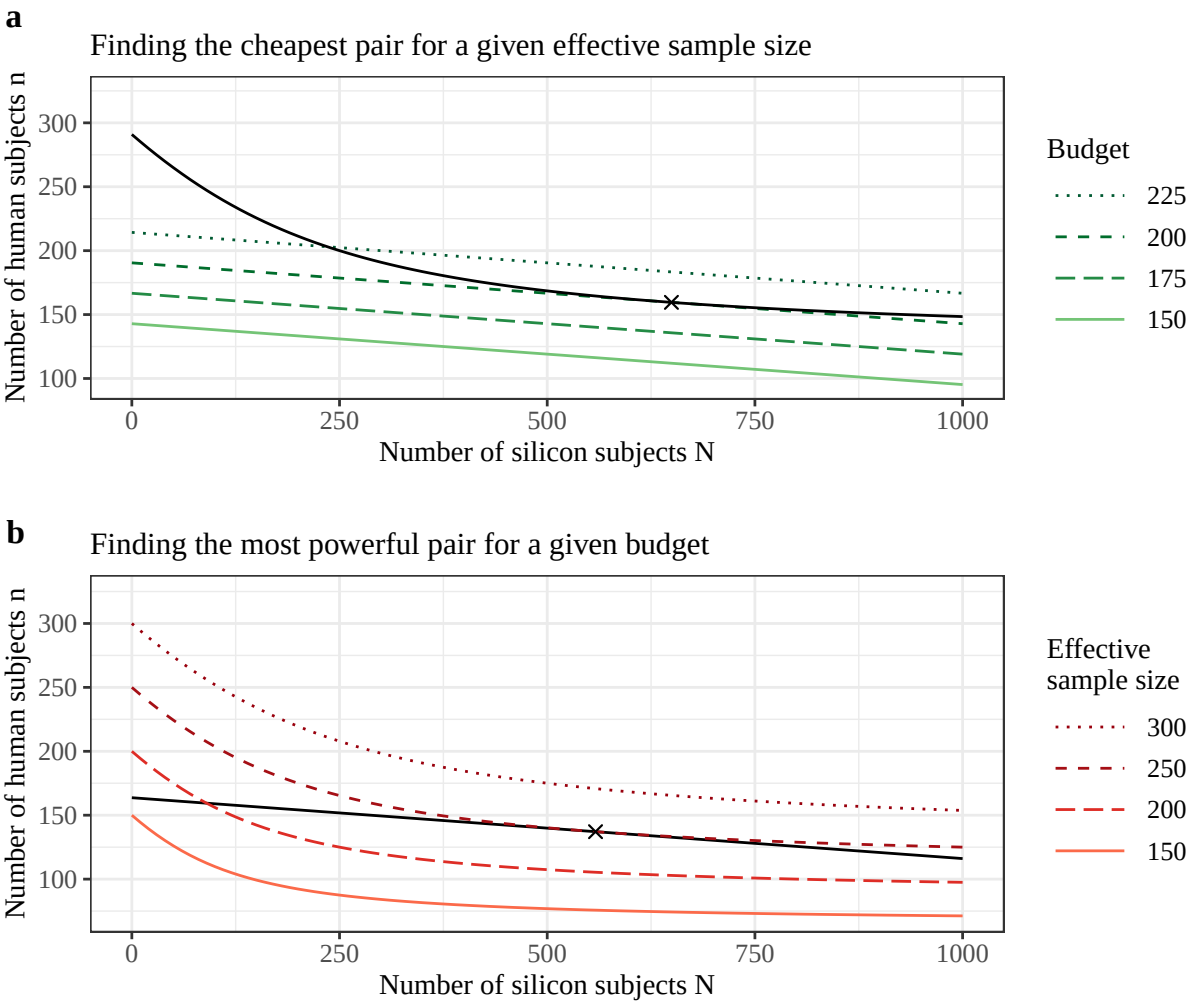

Figure 3: Illustration of constraint optimization with the PPI power analysis to find the optimal mix of n human subjects and N silicon subjects (black cross). (**a**) Researchers have multiple options for combining sample sizes (n, N) to achieve a given level of statistical power (black curve). However, only one such pair (n, N) is combined to an effective sample size that minimizes budget expenditure. (**b**) Researchers could allocate a given budget to multiple combinations of sample sizes (n, N) that do not exceed the budget of the experiment (black line). However, only one such pair (n, N) is combined to an effective sample size that maximizes statistical power.

3.5 Lowering Costs of Data Collection

In this section, we describe the cost reductions researchers can expect when conducting mixed subjects experiments with PPI. We explain that these cost reductions are determined by the PPI correlation and the cost of silicon subjects relative to human subjects. Throughout this section, we assume that researchers have determined the optimal mix of human and silicon subjects with our PPI power analysis. Therefore, the savings described here would apply to a subsequent confirmatory study conducted after the PPI correlation has been estimated in a pilot study.

The savings from using PPI depend on the cost of silicon and human subjects. We use γ to denote the ratio of the costs of surveying silicon and human subjects. That is, $\gamma = \frac{c_f}{c_Y}$ 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 S1.7 in the Supporting Information, we show that PPI with our power analysis is more cost-effective than classic inference with human subjects if and only if

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

For illustration, suppose recruiting a human subject costs \$1, while using an LLM to predict the response of a human subject costs \$0.1. Researchers can therefore recruit $1/\gamma = 10$ silicon subjects for the cost of recruiting one human subject. In this example, any PPI correlation $\tilde{\rho} > 0.575$ would be sufficient for PPI to save costs compared to a study with only human subjects. If we use an informative prediction algorithm ($\tilde{\rho} = 0.7$), the cost of a mixed subjects experiment is 87.5% of the cost of a human subjects experiment. The cost reduces to 51.9% if we use a more informative algorithm ($\tilde{\rho} = 0.9$). The same cost reduction occurs if the cost of silicon subjects decreases enough to recruit $1/\gamma = 12{,}144$ silicon subjects for every human subject, while keeping the PPI correlation at $\tilde{\rho} = 0.7$.

More generally, Figure 4 shows that mixed subjects experiments become less expensive as predictions become more affordable, with substantial savings at higher values of the PPI correlation. As the costs for prompting LLMs decrease (e.g. due to lower API fees) and LLMs become more capable of predicting human behavior (e.g., due to improved prompting strategies and the release of more sophisticated models), the cost of conducting mixed subjects experiments will further decrease relative to human subjects experiments. However, Figure 4 also shows that the savings from a mixed subjects design are bounded even if the costs of prompting LLMs go to zero. If prompting LLMs is free ($\gamma = 0$), then the relative savings from PPI will be $1 - \tilde{\rho}_+^2$. Therefore, if human and silicon subjects are not fully interchangeable ($\tilde{\rho} < 1$), then the cost is always a non-zero percentage of a human subjects experiment. This bound on decreasing costs is relevant to mixed subjects experiments since the cost of prompting LLMs is very small compared to the cost of recruiting human subjects. Therefore, mixed subjects experiments will benefit

more from an improvement in the PPI correlation than a reduction in the cost of prompting LLMs.

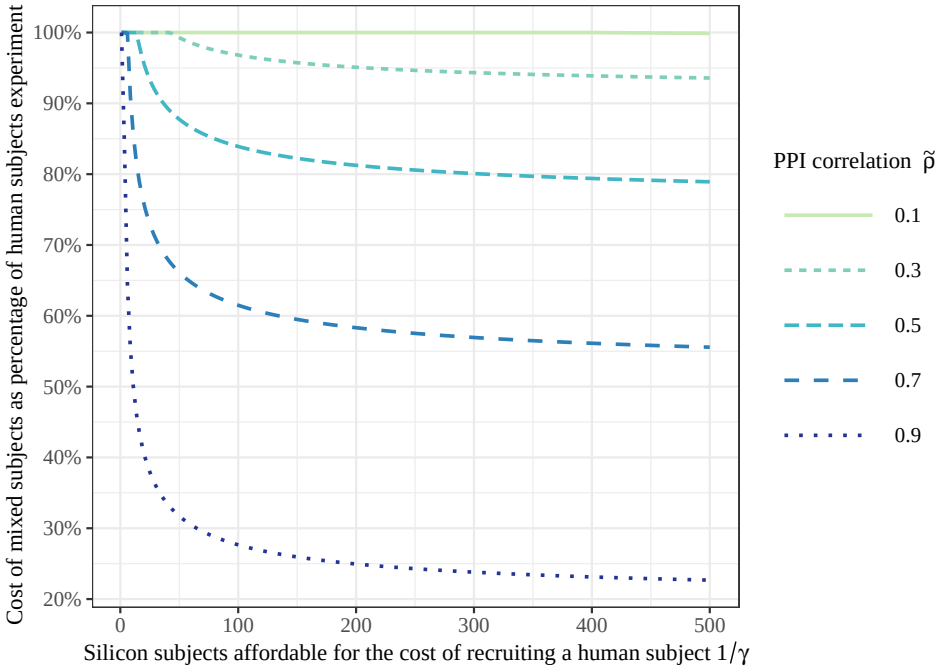

Figure 4: The y-axis shows the cost of conducting a mixed subjects experiment as a percentage of the cost of conducting an experiment with human subjects only, given by $1-\tilde{\rho}_+^2(1-\gamma)+2\sqrt{\gamma\tilde{\rho}_+^2(1-\tilde{\rho}_+^2)}$. We plot this percentage as a function of the PPI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

3.6 Conducting Mixed Subjects Experiments

In this section, we discuss the implementation of a mixed subjects design with PPI. While the exact steps may depend on the specific research project, we generally envision mixed subjects designs to involve two main steps. First, researchers conduct a pilot study with a smaller dataset to find the optimal combination of sample sizes using the PPI power analysis described in the Section 3.4. Then, in a confirmatory study, researchers conduct a second experiment with the optimal sample sizes to estimate parameters with PPI, as explained in Sections 3.1 and 3.2. Table 1 summarizes these steps and the main design considerations for researchers conducting mixed subjects experiments, including prompt design, the assumptions of PPI, and ways to enhance research transparency. We detail these considerations in the following.

Predicting survey responses with LLMs requires converting information about the subjects and the study into prompts (see Section S3 in the Supporting Information for an example). Previously, we considered the prompt as part of the prediction algorithm f. More explicitly, one can denote f as a composition of two functions $f(X_i) = m(p(X_i))$ where m denotes the LLM and $p(X_i)$ is the prompt given to m. The function p can be thought of as a "prompting template" (Hussain et al., 2024). This prompting

Conducting Mixed Subjects Experiments	Design Considerations
• **Step 1**: Conduct a pilot study to find the optimal mix of N silicon and n human subjects using the PPI power analysis • **Step 2**: Conduct a confirmatory study with the sample sizes determined by the power analysis. Estimate parameters and construct confidence intervals with PPI.	• Create a prompting template for the LLM and personas for the subjects • Sample human and silicon subjects from the same population • Apply the LLM consistently to labeled and unlabeled datasets • Ensure separation between labeled dataset and training data for the LLM • Document the PPI correlation, effective sample size, and prediction algorithm for confirmatory study as part of a preregistration

Table 1: Steps and design considerations for mixed subjects studies with PPI

template inserts X_i into a predefined prompt structure. We refer to X_i as the "persona" (e.g., Bisbee et al., 2024), which consists of demographics, attitudes, the assigned treatment, and other information about the subject. The prompting template p includes instructions for the LLM, a description of the treatment, and survey questions. For instance, if the original experiment presents participants with a vignette followed by a series of questions, the prompting template should instruct the LLM to predict a survey response based on the vignette text and the persona. Researchers should provide clear instructions to ensure that the LLM provides predictions in a consistent and usable format (Ziems et al., 2024).

The main assumption underlying PPI is that $f(X_i)$ and $f(\widetilde{X}_i)$ originate from the same distribution. To satisfy this assumption in a mixed subject experiment, two key requirements must be met. First, researchers must ensure that X_i and $\widetilde{X}_i$ are drawn from the same distribution. Second, the prediction algorithm f must be applied consistently across both the labeled and unlabeled datasets.

To ensure that X_i and $\widetilde{X}_i$ are from the same distribution, researchers should think of each silicon subject as a potential human subject. That is, the silicon subjects in the unlabeled dataset correspond to people who would have been surveyed had the sample size n been larger. In particular, the demographics and attitudes of the silicon subjects and human subjects must both be samples from the same population of interest. Likewise, the procedure for assigning human subjects to a treatment must match the procedure used for the silicon subjects. One way to ensure that both demographics and treatments come from the same distribution is to obtain a sample of size $n+N$ from the target population. For every subject in this pooled sample, researchers should assign experimental conditions and collect demographic information. Then, a random subset of size n is selected from the pooled sample to be part of the experiment. These selected subjects provide the labeled dataset $\{(X_i, Y_i)\}_{i=1}^{n}$. The treatment assignments and background information of the remaining subjects provide the unlabeled dataset $\{\widetilde{X}_i\}_{i=1}^{N}$.

While it may not always be feasible to implement every step of this procedure in praxis, researchers could consider reasonable approximations. For example, if eligible voters in the U.S. is the target population,

surveys such as the ANES could serve as the unlabeled datasets (Argyle et al., 2023; Bisbee et al., 2024). To create the labeled dataset, researchers could field an experiment that also targets eligible voters. Treatments can then be assigned at random to both datasets. In this case, researchers must assume they have access to the same population as the survey used to create the unlabeled dataset.

Even if X_i and $\widetilde{X}_i$ are from the same distribution, PPI also requires that the prediction algorithm f is used consistently on X_i and $\widetilde{X}_i$. If we again think of $f(X_i)$ as the composition $m(p(X_i))$, then both the prompting template p and the LLM m must be the same for both datasets. In particular, the same version of the LLM must be used with the same settings and hyperparameters. Additionally, for f to be used consistently on X_i and $\widetilde{X}_i$, the labeled data $\{(X_i, Y_i)\}_{i=1}^n$ must be kept separate from the data used to train the LLM. If the training data contains $\{(X_i, Y_i)\}_{i=1}^n$, then the prediction function f may be more accurate on the labeled data than on the unlabeled data. In this case, the prediction algorithm may have less bias on the labeled data than on the unlabeled data. The prediction algorithm would then introduce bias into the PPI estimator. The simplest way to avoid this problem is by conducting the experiment after the model has been trained.

These assumptions also apply to power analyses with PPI. In our PPI power analysis, researchers estimate the PPI correlation $\tilde{\rho}$ with a pilot study and then select a pair of sample sizes for a subsequent experiment. For this procedure to be valid, $\tilde{\rho}$ must be estimated using a sample from the targeted population. The prediction algorithm f must also be the same across the two experiments. If $\tilde{\rho}$ differs across the two experiments, then the estimated effective sample size may be incorrect and the selected pair of sample sizes will no longer be optimal.

Notably, the use of LLMs in the social sciences raises important questions about transparency and reproducibility (Spirling, 2023; Bail, 2024; Davidson, 2024). We recommend that researchers enhance transparency at every stage of their study. This includes sharing the prompts and code used for interacting with LLMs, as well as detailed information about the models employed—such as version numbers, settings, and relevant hyperparameters. Open-source models provide more transparency than their proprietary counterparts (Hussain et al., 2024). Therefore, open-source models may be better suited to ensure that the same algorithm is used for both the labeled and unlabeled datasets. Additionally, researchers should document the results of their power analysis in a preregistration document (Veer and Giner-Sorolla, 2014; Christensen et al., 2019), noting the estimated PPI correlation, effective sample size, and choice of prediction algorithm. The preregistration should also detail the sampling process for both labeled and unlabeled data, ensuring that others can understand and reproduce the methodology. Whenever possible, code, data, and other materials should be made publicly available in repositories such as the Open Science Framework (`https://osf.io`).

It should be noted that mixed subjects designs with PPI share some of the limitations of the human

subjects approach. One set of concerns relates to internal validity. For example, any bias in the human subjects estimator would also introduce bias into the PPI estimator. As discussed in Section 3.2, the PPI estimator is designed only to prevent the prediction algorithm from introducing bias. Additionally, researchers may have concerns about external validity. For example, some groups within the target population may be difficult to reach. While mixed subjects designs may reduce the number of human subjects needed for a desired level of precision, some populations may not be accessible at all. We therefore emphasize that reducing the cost of obtaining precise estimates is the primary benefit of a mixed subjects design.

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 forced a decision between harming pedestrians or passengers. 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. Awad et al. (2018) estimate the Average Marginal Component Effect (AMCE) to measure the causal effect of an attribute of a moral dilemma on a respondent's decision.

In the following, we present our reanalysis of the Moral Machine experiment. We compare the mixed subjects and silicon subjects design to a human subjects design, which serves as our benchmark. Our reanalysis demonstrates how a mixed subjects design can address the challenges associated with silicon sampling detailed in Section 2.2. Specifically, we use a simulation to show how relying on an increasing number of LLM predictions affects the bias and precision of the AMCE estimates in both designs.

4.1 Methods

Awad et al. (2018) obtained a convenience sample with millions of decisions on moral dilemmas from participants worldwide. For our analysis, we use the subset of 55,893 Americans who completed an optional demographic survey.[5] On average, these participants evaluated 10.4 dilemmas. Each dilemma presented participants with two options, resulting in a total sample size of 1,163,962 decisions. We treat this dataset as an artificial population for our simulation.[6] In particular, AMCEs computed on

this dataset are treated as the ground truth parameters. To create i.i.d. samples of human and silicon subjects, we sample individual decisions on moral dilemmas.

For the mixed subjects approach, we randomly select $n = 10{,}000$ decisions from the sample of 1,163,962 decisions. For the unlabeled dataset, we randomly select $N = n \times k$ decisions, where $k = (.25, .5, .75, 1, 1.5, ..., 9.5, 10)$ are multiples of the human subjects sample size. For the silicon subjects approach, we take this unlabeled dataset of size N as the silicon sample. For each combination of sample sizes n and N, we repeat the sampling 500 times. In each repetition, we estimate the AMCE for nine scenario attributes for both designs. For the mixed subjects design we use the Python library created by Angelopoulos et al. (2023, 2024) to obtain the corresponding PPI estimates and confidence intervals. For the silicon subjects design, we use the predicted decision as the dependent variable. For both designs, we estimate the AMCEs with simple weighted linear regressions (Hainmueller et al., 2014).

For each of the nine AMCEs, we evaluate the bias, precision, and coverage of confidence intervals in both designs. We calculate bias by subtracting the average AMCE estimate across repetitions from the ground truth AMCE computed on the entire sample. We define precision as the width of a confidence interval as a percentage of the width of the confidence interval computed using n human subjects. As such, the precision of estimates in the mixed and silicon subjects designs is defined relative to a human subjects design. Finally, we calculate coverage as the percentage of confidence intervals across repetitions that include the ground truth AMCEs. Coverage can be considered a summary measure of bias and precision since coverage is influenced by both.

We use the replication data from the Moral Machine experiment to create the prompts for the LLMs. These prompts describe the moral dilemmas that survey participants evaluated, allowing us to predict their decisions using LLMs. We convert the numerical representations of the dilemmas in the replication data into descriptive text. For instance, the replication data records the number of passengers and pedestrians in the moral dilemmas, how many of those were children, adults, or elderly, and whether the participant had to intervene to spare one of the two groups. Using computer code adapted from a related study (Takemoto, 2024), we automatically created these descriptions in English. We also add a demographic profile to each prompt, including the age, education level, gender, and income of the survey respondent who evaluated the dilemma. Please refer to Section S3 in the Supporting Information for an example of a prompt. We then use the OpenAI API to request predictions from three LLMs—GPT-4 Turbo (gpt-4-turbo), GPT-4o (gpt-4o), and GPT-3.5 Turbo (gpt-3.5-turbo-0125)—to predict the survey respondents' decisions in the moral dilemmas. We report the results for GPT-4 Turbo, the LLM with the best performance in predicting the survey responses in the Moral Machine experiment.

4.2 Results

Our simulation with the Moral Machine experiment corroborates our arguments about bias in the mixed subject and silicon subjects design (Figure 5a and b). A mixed subjects design with PPI provides unbiased estimates of causal effects while silicon sampling does not. We found that bias in silicon sampling can be as large as 38.3% of the total scale of the variable, which ranges from zero (harm) to one (save). Large biases are particularly consequential if the underlying effect is small. While the silicon subject estimate suggests that there is a 49 percentage point difference in the preference to spare pedestrians rather than passengers, the ground truth AMCE indicates that this preference amounts to only 11 percentage points. This bias corresponds to about 363% of the original effect.

We now comment on our findings on the precision of parameter estimates in the mixed subjects and silicon subjects design (Figure 5c and d). We confirm that the degree to which LLM predictions reduce the width of confidence intervals depends greatly on the PPI correlation. We find that LLM predictions of decisions to moral dilemmas from GPT-4 Turbo are mostly not interchangeable with observed decisions from survey respondents—the PPI correlation ranged from 0.049 to 0.353 for the nine different scenario characteristics. As a result of this modest interchangeability, complementing the 10,000 human observations with 100,000 LLM predictions results in an effective sample size of at most 11,275 (Table S2 in the Supporting Information). Hence, the tightening of confidence intervals is also modest.

In contrast, Figure 5d shows that the silicon subjects design first gives a wider confidence interval than the human subjects design. The confidence intervals for the silicon subjects become increasingly narrow as the sample size increases. As noted in Section 2.2, overly narrow confidence intervals give a false sense of precision if the estimates are biased.

Figure 5e shows that PPI, unlike the estimates derived from silicon subjects alone, maintains a nominal coverage rate of 95%. PPI therefore correctly quantifies uncertainty about the true population parameter. Figure 5f demonstrates that none of the silicon subjects confidence intervals achieve nominal coverage for larger sample sizes. This lack of coverage results from a combination of bias and overconfidence. Without an appropriate correction, even small biases cause a lack of coverage, with coverage decreasing more rapidly for parameter estimates with greater bias.

Finally, Figures 5b and c demonstrate an interesting property of PPI—bias and interchangeability do not exclude each other. The attribute with the highest bias in silicon sampling (sparing pedestrians vs passengers) also achieved the second-highest PPI correlation. As such, biased information from the LLM can still be useful for increasing precision in PPI.

We end this section with two broader points. First, our reanalysis of the Moral Machine experiment illustrates that PPI produces valid point estimates, with increases in statistical precision depending on

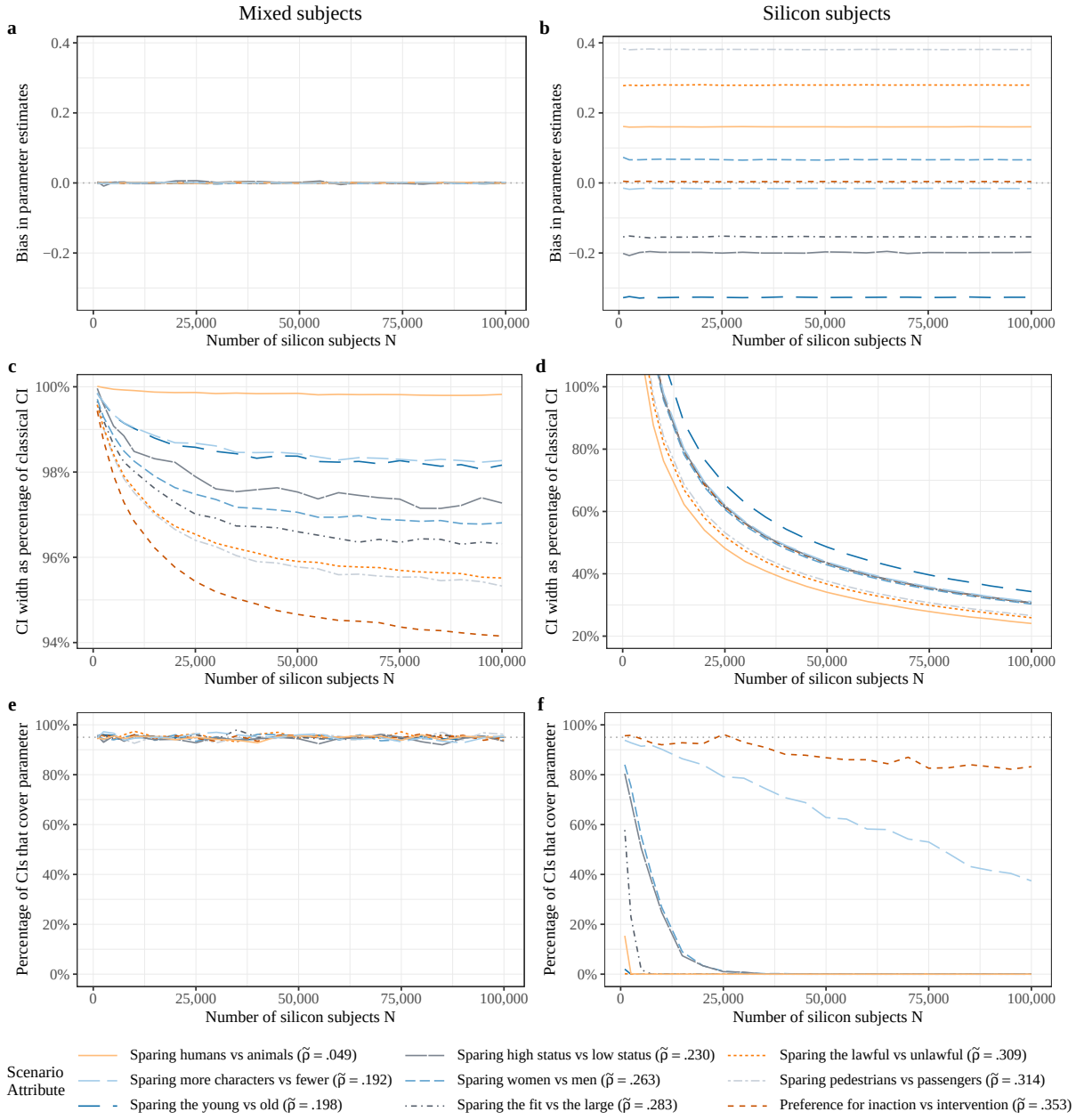

Figure 5: Comparison of bias (**a**, **b**), precision (**c**, **d**), and coverage of confidence intervals (**e**, **f**) in the mixed subjects and silicon subjects designs. The PPI correlations $\tilde{\rho}$ were computed separately from the simulation, using sample sizes n =1,000,000 and N=163,962.

the PPI correlation. The silicon sampling approach may produce correct point estimates in some settings some of the time, but a researcher has no way of knowing without validation on human subjects. In contrast, researchers can have confidence in point estimates from mixed subjects experiments, as PPI automatically prevents the silicon subjects from introducing bias.

Second, our empirical example illustrates that the PPI correlation and the effective sample size are intuitive metrics for gauging the degree to which LLM predictions can increase precision. Using these metrics, researchers can assess whether incorporating LLMs meaningfully reduces the cost of obtaining

precise estimates in a mixed subjects design. Naturally, this decision also depends on factors such as data collection costs and budget constraints (see Section 3.4). Our PPI power analysis supports researchers in making informed choices about how to implement their experiments. For a demonstration of our power analysis, please refer to the GitHub repository at `https://github.com/aangelopoulos/ppi_py/blob/6b8475b53f0303e4abc4d1eb0b94aa8052f9b940/examples/power_analysis.ipynb`

5 Conclusion

Large language models, and generative AI more generally, offer new data sources that may transform the social sciences. Predictions of human behavior offered by generative AI—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 when these predictions systematically diverge from actual human behavior. Even if LLMs become more accurate in predicting human behavior than they are today, these predictions remain of minimal benefit because researchers must 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 prevents LLMs from introducing bias to parameter estimates. Mixed subjects studies with PPI also allow researchers to obtain narrower confidence intervals and higher statistical power at a lower cost 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 approaches.

Our statistical contributions to PPI are two-fold. First, we define 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. As algorithms become more capable, PPI estimates will achieve higher statistical precision, and the cost of conducting mixed subjects experiments will decrease further 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 complements the toolkit necessary to conduct mixed subjects studies with PPI.

6 Future Directions

Our work points to immediate next steps in designing mixed subjects experiments. We first highlight the further development of PPI as a promising avenue for future research (e.g., Zrnic and Candès, 2024; Gligorić et al., 2024; Fisch et al., 2024). Researchers could adopt PPI to handle correlated observations and extend PPI beyond currently supported generalized linear models (e.g., survival analysis). Moreover, extending PPI to incorporate multiple prediction algorithms could further enhance the precision of its parameter estimates. While we leverage PPI to implement mixed subjects designs, other methods are also well-suited for this purpose. There is a broader literature on using predictions as data to perform valid inference on parameters (Hoffman et al., 2024). We expect interest in these statistical methods to grow as prediction algorithms become increasingly powerful. Therefore, we believe that systematic comparisons of PPI with comparable frameworks are warranted (e.g., Blackwell et al., 2017; Egami et al., 2024).

Second, we note that both silicon subjects and mixed subjects designs benefit when predictions become more interchangeable with observed behavior. Understanding which types of data enhance predictions could thus advance both approaches. Beyond demographics and attitudes from quantitative surveys (e.g., Argyle et al., 2023; Bisbee et al., 2024), researchers could include an individual's social network, biography, or other contextual information to help LLMs better predict human behavior. Moreover, the performance of LLMs also depends on the prompt structure. Carefully developed best practices, such as prompting guidelines (e.g., Ziems et al., 2024), could further enhance the benefits of LLMs. Finally, social scientists could collaborate with computer scientists and statisticians to build models that are most useful for research in the social sciences (Bail, 2024). Specialized systems developed through such collaborations may be more capable of predicting human behavior than relying on LLMs "out of the box."

In a third future direction, we encourage researchers to leverage the possibility of obtaining valid and precise estimates at lower costs. A mixed subjects design could facilitate studies that would be prohibitively expensive if conducted solely with human subjects. For instance, identifying small treatment effects or interactions with sufficient statistical power requires recruiting thousands of human subjects,

which incurs costs that exceed many research budgets. Moreover, studies aimed at systematically exploring a large number of hypotheses have important practical and theoretical implications, but require an extraordinary number of human subjects (e.g., DellaVigna and Pope, 2018; Milkman et al., 2021; Tappin et al., 2023; Voelkel et al., 2024; Almaatouq et al., 2024). Resource constraints can even prevent those with more generous budgets from testing all promising hypotheses (e.g., Chu et al., 2024). Put differently, the cost of recruiting human subjects can be a "real constraint that should not be ignored" (Salganik, 2019). By reducing the cost of obtaining precise and valid parameter estimates, 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.

Data and Code Availability Statement

A replication package is available at `https://github.com/davidbroska/MixedSubjects`. Our two statistical contributions—the PPI correlation and the PPI power analysis—are implemented in the `ppi_py` Python library at `https://github.com/aangelopoulos/ppi_py`.

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 (e.g., Breiman 2001). For the purposes of this article, we consider LLMs to be part of a broader class of algorithms that provide potentially informative proxies for observations. We will therefore stick with "predict" as our preferred term. We refer the reader to introductions by Korinek (2023) and Hussain et al. (2024) for more details on LLMs.

2. We observe that silicon sampling can give a false sense of precision in our empirical application in Section 4 and in our simulation study in Section S2 in the Supporting Information.

3. In Section 3.6, we explain how the assumptions for PPI apply to LLMs as prediction algorithms.

4. Section S1.2 in the Supporting Information describes the PPI estimator for ordinary linear regression and Section S1.3 contains a general overview on PPI estimators.

5. For a comparison of the demographic distribution of participants in the Moral Machine experiment and the American population, please refer to Figure S3 in the Supporting Information.

6. We also include a pure simulation in Section S2 in the Supporting Information. In this simulation, we can specify the ground truth parameters, the bias in the prediction algorithm, and the PPI correlation. The findings in the simulation mirror those of the empirical example.

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., and Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, 3(7).

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

Alvero, A. J., Lee, J., Regla-Vargas, A., Kizilcec, R. F., Joachims, T., and Antonio, A. L. (2024). Large language models, social demography, and hegemony: Comparing authorship in human and synthetic text. *Journal of Big Data*, 11(1).

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671):669–674.

Angelopoulos, A. N., Duchi, J. C., and Zrnic, T. (2024). Ppi++: Efficient prediction-powered inference.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351.

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

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., and Henrich, J. (2023). Which humans?

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563(7729):59–64.

Bail, C. A. (2024). Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21).

Berinsky, A. J., Huber, G. A., and Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3):351–368.

Binz, M. and Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6).

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., and Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4):401–416.

Blackwell, M., Honaker, J., and King, G. (2017). A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods & Research*, 46(3):303–341.

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, 600(7890):695–700.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., and Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, 51(5):2022–2038.

Christensen, G., Freese, J., and Miguel, E. (2019). *Transparent and reproducible social science research: How to do open science*. University of California Press.

Chu, J. Y., Voelkel, J. G., Stagnaro, M. N., Kang, S., Druckman, J. N., Rand, D. G., and Willer, R. (2024). Academics are more specific, and practitioners more sensitive, in forecasting interventions to strengthen democratic attitudes. *Proceedings of the National Academy of Sciences*, 121(3).

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, 2nd edition.

Davidson, T. (2024). Start Generating: Harnessing Generative Artificial Intelligence for Sociological Research. *Socius*, 10.

DellaVigna, S. and Pope, D. (2018). What Motivates Effort? Evidence and Expert Forecasts. *The Review of Economic Studies*, 85(2):1029–1069.

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

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC Press.

Egami, N., Hinck, M., Stewart, B. M., and Wei, H. (2024). Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models. arXiv:2306.04746 [cs, stat].

Fisch, A., Maynez, J., Hofer, R. A., Dhingra, B., Globerson, A., and Cohen, W. W. (2024). Stratified prediction-powered inference for hybrid language model evaluation.

Freese, J. and Peterson, D. (2018). The Emergence of Statistical Objectivity: Changing Ideas of Epistemic Vice and Virtue in Science. *Sociological Theory*, 36(3).

Gelman, A. (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.

Gligorić, K., Zrnic, T., Lee, C., Candès, E. J., and Jurafsky, D. (2024). Can unconfident llm annotations be used for confident conclusions?

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., and Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, 380(6650):1108–1109.

Hainmueller, J., Hopkins, D. J., and Yamamoto, T. (2014). Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis*, 22(1):1–30.

Harding, J., D'Alessandro, W., Laskowski, N., and Long, R. (2023). Ai language models cannot replace human research participants. *Ai & Society*.

Hoffman, K., Salerno, S., Afiaz, A., Leek, J. T., and McCormick, T. H. (2024). Do we really even need data?

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

Hussain, Z., Binz, M., Mata, R., and Wulff, D. U. (2024). A tutorial on open-source large language models for behavioral science. *Behavior Research Methods*, 56(8):8214–8237.

Korinek, A. (2023). Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computational Social Science. *Science*, 323(5915):721–723.

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., and Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.

Levay, K. E., Freese, J., and Druckman, J. N. (2016). The Demographic and Political Composition of Mechanical Turk Samples. *Sage Open*, 6(1):1–17.

Li, P., Castelo, N., Katona, Z., and Sarvary, M. (2024). Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis. *Marketing Science*, 43(2):254–266.

McFarland, D. A. and McFarland, H. R. (2015). Big Data and the danger of being precisely inaccurate. *Big Data & Society*, 2(2).

Mei, Q., Xie, Y., Yuan, W., and Jackson, M. O. (2024). A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9).

Meng, X.-L. (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(2):685–726.

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

Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., Park, Y., Rai, A., Bazerman, M., Beshears, J., Bonacorsi, L., Camerer, C., Chang, E., Chapman, G., Cialdini, R., Dai, H., Eskreis-Winkler, L., Fishbach, A., Gross, J. J., Horn, S., Hubbard, A., Jones, S. J., Karlan, D., Kautz, T., Kirgios, E., Klusowski, J., Kristal, A., Ladhania, R., Loewenstein, G., Ludwig, J., Mellers, B., Mullainathan, S., Saccardo, S., Spiess, J., Suri, G., Talloen, J. H., Taxer, J., Trope, Y., Ungar, L., Volpp, K. G., Whillans, A., Zinman, J., and Duckworth, A. L. (2021). Megastudies improve the impact of applied behavioural science. *Nature*, 600(7889):478–483.

Park, P. S., Schoenegger, P., and Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6):5754–5770.

Rauf, T., Voelkel, J. G., Druckman, J., and Freese, J. (2024). An Audit of Social Science Survey Experiments.

Salerno, S., Miao, J., Afiaz, A., Hoffman, K., Neufeld, A., Lu, Q., McCormick, T. H., and Leek, J. T. (2024). ipd: An r package for conducting inference on predicted data.

Salganik, M. (2019). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Sarstedt, M., Adler, S. J., Rau, L., and Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, 41(6):1254–1270.

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature*, 616(7957):413–413.

Stadtfeld, C., Snijders, T. A. B., Steglich, C., and van Duijn, M. (2020). Statistical Power in Longitudinal Network Studies. *Sociological Methods & Research*, 49(4):1103–1132.

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

Takemoto, K. (2024). The moral machine experiment on large language models. *Royal Society Open Science*, 11(2).

Tappin, B. M., Wittenberg, C., Hewitt, L. B., Berinsky, A. J., and Rand, D. G. (2023). Quantifying

the potential persuasive returns to political microtargeting. *Proceedings of the National Academy of Sciences*, 120(25).

Thye, S. R. (2000). Reliability in Experimental Sociology. *Social Forces*, 78(4):1277–1309.

Veer, A. v. t. and Giner-Sorolla, R. (2014). Public template for pre-registration. https://osf.io/k5wns/.

Voelkel, J. G., Stagnaro, M. N., Chu, J. Y., Pink, S. L., Mernyk, J. S., Redekopp, C., Ghezae, I., Cashman, M., Adjodah, D., Allen, L. G., Allis, L. V., Baleria, G., Ballantyne, N., Van Bavel, J. J., Blunden, H., Braley, A., Bryan, C. J., Celniker, J. B., Cikara, M., Clapper, M. V., Clayton, K., Collins, H., DeFilippis, E., Dieffenbach, M., Doell, K. C., Dorison, C., Duong, M., Felsman, P., Fiorella, M., Francis, D., Franz, M., Gallardo, R. A., Gifford, S., Goya-Tocchetto, D., Gray, K., Green, J., Greene, J., Güngör, M., Hall, M., Hecht, C. A., Javeed, A., Jost, J. T., Kay, A. C., Kay, N. R., Keating, B., Kelly, J. M., Kirk, J. R. G., Kopell, M., Kteily, N., Kubin, E., Lees, J., Lenz, G., Levendusky, M., Littman, R., Luo, K., Lyles, A., Lyons, B., Marsh, W., Martherus, J., Maurer, L. A., Mehl, C., Minson, J., Moore, M., Moore-Berg, S. L., Pasek, M. H., Pentland, A., Puryear, C., Rahnama, H., Rathje, S., Rosato, J., Saar-Tsechansky, M., Almeida Santos, L., Seifert, C. M., Shariff, A., Simonsson, O., Spitz Siddiqi, S., Stone, D. F., Strand, P., Tomz, M., Yeager, D. S., Yoeli, E., Zaki, J., Druckman, J. N., Rand, D. G., and Willer, R. (2024). Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. *Science*, 386(6719).

Zack, E. S., Kennedy, J., and Long, J. S. (2019). Can Nonprobability Samples be Used for Social Science Research? A cautionary tale. *Survey Research Methods*.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2024). Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291.

Zrnic, T. and Candès, E. J. (2024). Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15).

The Mixed Subjects Design
Supporting Information

S1 Prediction Powered Inference (PPI)

S1.1 Alternative Expression for the PPI Mean Estimator

In Section 3.2, we proposed that the PPI estimator for the population mean can be rewritten to take the form of an average of residuals. In the following, we detail the steps for rewriting the PPI estimator. Recall that the PPI estimator for the mean of Y_i from equation (1)

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

and that the silicon subjects estimator is $\hat{\theta}^S = \frac{1}{N}\sum_{i=1}^{N} f(\widetilde{X}_i)$. Note also that $\hat{\theta}^S = \frac{1}{n}n\hat{\theta}^S = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}^S$. We therefore have

$$\begin{aligned}
\hat{\theta}^{\text{PP}}_{\lambda} &= \frac{1}{n}\sum_{i=1}^{n} Y_i - \lambda\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{N}\sum_{i=1}^{N} f(\widetilde{X}_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i - \lambda\left(\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) - \hat{\theta}^s\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i - \lambda\left(\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}^s\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n}\lambda(f(X_i) - \hat{\theta}^s) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \lambda(f(X_i) - \hat{\theta}^S))\right),
\end{aligned}$$

as claimed in equation (3).

S1.2 PPI for Linear Regression

In this section, we explain PPI for linear regression. This is an important case since many causal estimators can be written as a coefficient in a linear regression. This explanation builds on the explanation of PPI for estimating the mean given in Sections 3.1 and 3.2 of the main text. As in Sections 3.1 and 3.2, we focus on how PPI avoids introducing bias and why PPI is more precise than classical inference. A more general presentation of how PPI estimates parameters is given in Section S1.3.

Let $\{(X_i, Y_i)\}_{i=1}^{n}$ be a labeled dataset with $X_i \in \mathbb{R}^d$. Let $\{\widetilde{X}_i\}_{i=1}^{N}$ be the unlabeled dataset. Let $\theta^\star$ be

the vector of population-level linear regression coefficients. That is,

$$\theta^\star = \mathbb{E}[X_i X_i^\top]^{-1}\mathbb{E}[X_i^\top Y_i]$$

The classical—or human subjects—estimator of $\theta^\star$ is $\hat{\theta}^H$ given by

$$\hat{\theta}^H = \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{Y}\right)$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the matrix with rows X_i and $\mathbf{Y} \in \mathbb{R}^n$ is the vector with entries Y_i. For $\lambda \in [0,1]$, the PPI estimator $\hat{\theta}^{\text{PP}}_\lambda$ is

$$\hat{\theta}^{\text{PP}}_\lambda = \left(\frac{1-\lambda}{n}\mathbf{X}^\top\mathbf{X} + \frac{\lambda}{N}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right)^{-1}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{Y} - \lambda\left(\frac{1}{n}\mathbf{X}^\top f(\mathbf{X}) - \frac{1}{N}\widetilde{\mathbf{X}}^\top f(\widetilde{\mathbf{X}})\right)\right) \tag{S1}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are as above, $\widetilde{\mathbf{X}} \in \mathbb{R}^{N\times d}$ is the matrix with rows $\widetilde{X}_i$, $f(\mathbf{X}) \in \mathbb{R}^n$ is the vector with entries $f(X_i)$ and $f(\widetilde{\mathbf{X}}) \in \mathbb{R}^N$ is the vector with entries $f(\widetilde{X}_i)$.

We will now compare the OLS estimator in (S1) to the mean estimator in equation (1) of the main text. Define $\widehat{\mathbf{H}}_\lambda = \frac{1-\lambda}{n}\mathbf{X}^\top\mathbf{X} + \frac{\lambda}{N}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$ and $\mathbf{H} = \mathbb{E}[X_i X_i^\top]$. By definition, we have

$$\theta^\star = \mathbf{H}^{-1}\mathbb{E}[X_i^\top Y_i] \quad \text{and} \quad \hat{\theta}^{\text{PP}}_\lambda = \widehat{\mathbf{H}}_\lambda^{-1}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{Y} - \lambda\left(\frac{1}{n}\mathbf{X}^\top f(\mathbf{X}) - \frac{1}{N}\widetilde{\mathbf{X}}^\top f(\widetilde{\mathbf{X}})\right)\right) \tag{S2}$$

By the assumption that X_i and $\widetilde{X}_i$ are drawn from the same distribution, we have that

$$\begin{aligned}
\mathbb{E}[\widehat{\mathbf{H}}_\lambda] &= \mathbb{E}\left[\frac{1-\lambda}{n}\mathbf{X}^\top\mathbf{X} + \frac{\lambda}{N}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right] \\
&= \mathbb{E}\left[\frac{1-\lambda}{n}\sum_{i=1}^n X_i X_i^\top + \frac{\lambda}{N}\sum_{i=1}^N \widetilde{X}_i\widetilde{X}_i^\top\right] \\
&= \frac{1-\lambda}{n}\sum_{i=1}^n \mathbb{E}[X_i X_i^\top] + \frac{\lambda}{N}\sum_{i=1}^N \mathbb{E}[\widetilde{X}_i\widetilde{X}_i^\top] \\
&= (1-\lambda)\mathbb{E}[X_i X_i^\top] + \lambda\mathbb{E}[\widetilde{X}_i\widetilde{X}_i^\top] \\
&= \mathbb{E}[X_i X_i^\top] \\
&= \mathbf{H}
\end{aligned}$$

Furthermore, by the law of large numbers, if n and N go to infinity, then $\widehat{\mathbf{H}}_\lambda$ converges in probability to H. To simplify the exposition, we will replace $\widehat{\mathbf{H}}_\lambda$ in (S2) with $\mathbf{H}$.[1] That is, we will study the estimator

$$\hat{\theta}^{\text{PP}}_\lambda = \mathbf{H}^{-1}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{Y} - \lambda\left(\frac{1}{n}\mathbf{X}^\top f(\mathbf{X}) - \frac{1}{N}\widetilde{\mathbf{X}}^\top f(\widetilde{\mathbf{X}})\right)\right) \tag{S3}$$

Note that

$$\begin{aligned}
\hat{\theta}^{\text{PP}}_\lambda &= \mathbf{H}^{-1}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{Y} - \lambda\left(\frac{1}{n}\mathbf{X}^\top f(\mathbf{X}) - \frac{1}{N}\widetilde{\mathbf{X}}^\top f(\widetilde{\mathbf{X}})\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{H}^{-1}X_i Y_i - \lambda\left(\frac{1}{n}\sum_{i=1}^n \mathbf{H}^{-1}X_i f(X_i) - \frac{1}{N}\sum_{i=1}^N \mathbf{H}^{-1}\widetilde{X}_i f(\widetilde{X}_i)\right)
\end{aligned} \tag{S4}$$

[1] By Slutsky's Theorem, changing $\widehat{\mathbf{H}}_\lambda$ to $\mathbf{H}$ will not change asymptotic distribution of $\hat{\theta}^{\text{PP}}_\lambda$.

The final expression for $\hat{\theta}^{\text{PP}}_{\lambda}$ in equation (S4) is directly analogous to the mean estimator in equation (1). In equation S4, we have a term involving the true labels ($\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}^{-1}X_iY_i$) and the difference of two terms involving the prediction algorithm ($\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}^{-1}X_if(X_i) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{H}^{-1}\widetilde{X}_if(\widetilde{X}_i)$). Thus, as with the mean estimator, $\hat{\theta}^{\text{PP}}_{\lambda}$ is unbiased for the population-level linear regression coefficients.

$$\begin{aligned}
\mathbb{E}[\hat{\theta}^{\text{PP}}_{\lambda}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}^{-1}X_iY_i - \lambda\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}^{-1}X_if(X_i) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{H}^{-1}\widetilde{X}_if(\widetilde{X}_i)\right)\right] \\
&= \mathbf{H}^{-1}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_iY_i - \lambda\left(\frac{1}{n}\sum_{i=1}^{n}X_if(X_i) - \frac{1}{N}\sum_{i=1}^{N}\widetilde{X}_if(\widetilde{X}_i)\right)\right] \\
&= \mathbf{H}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_iY_i] - \lambda\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_if(X_i)] - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\widetilde{X}_if(\widetilde{X}_i)]\right)\right) \\
&= \mathbf{H}^{-1}\left(\mathbb{E}[X_iY_i] - \lambda\left(\mathbb{E}[X_if(X_i)] - \mathbb{E}[\widetilde{X}_if(\widetilde{X}_i)]\right)\right) \\
&= \mathbf{H}^{-1}\mathbb{E}[X_iY_i] \\
&= \theta^{\star}
\end{aligned}$$

Furthermore, as in the case of mean estimation, the PPI estimator is designed to be more precise than the classical estimator. The parameter λ is *power tuned* to minimize the variance of $\hat{\theta}^{\text{PP}}_{\lambda,j}$—a particular coefficient in the regression. Power tuning achieves a reduction in variance by representing $\hat{\theta}^{\text{PP}}_{\lambda}$ as an average of residuals. This was also shown for mean estimation in equation (3). For the OLS estimator, let $\hat{\theta}^{S} = \mathbf{H}^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\widetilde{X}_if(\widetilde{X}_i)\right)$. By an argument similar to the one given in Section S1.1, we have

$$\hat{\theta}^{\text{PP}}_{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{H}^{-1}X_iY_i - \lambda(\mathbf{H}^{-1}X_if(X_i) - \hat{\theta}^{S})\right) \tag{S5}$$

Thus, $\hat{\theta}^{\text{PP}}_{\lambda}$ can be thought of as an average of residuals with dependent variable $\mathbf{H}^{-1}X_iY_i$ and independent variable $\mathbf{H}^{-1}X_iY_i - \hat{\theta}^{S}$. This analogy is not perfect because $\mathbf{H}^{-1}X_iY_i$ and $\mathbf{H}^{-1}X_iY_i - \hat{\theta}^{S}$ are both vectors in $\mathbb{R}^d$. This is why power tuning focuses on a particular coordinate j. When focusing on a particular coordinate, we have

$$\hat{\theta}^{\text{PP}}_{\lambda,j} = \frac{1}{n}\sum_{i=1}^{n}\left([\mathbf{H}^{-1}X_iY_i]_j - \lambda\left[\mathbf{H}^{-1}X_if(X_i) - \hat{\theta}^{S}\right]_j\right)$$

Now the dependent variable is the scalar $[\mathbf{H}^{-1}X_iY_i]_j$ and the independent variable is $\left[\mathbf{H}^{-1}X_if(X_i) - \hat{\theta}^{S}\right]_j$. The optimal value of λ has a form that is similar to a regression coefficient:

$$\begin{aligned}
\lambda^{\star}_j &= \frac{N}{n+N}\frac{\text{Cov}([\mathbf{H}^{-1}X_iY_i]_j, [\mathbf{H}^{-1}X_if(X_i)]_j)}{\text{Var}([\mathbf{H}^{-1}X_if(X_i)]_j)} \\
&= \frac{N}{n+N}\frac{[\mathbf{H}^{-1}\,\text{Cov}(X_iY_i, X_if(X_i))\mathbf{H}^{-1}]_{j,j}}{[\mathbf{H}^{-1}\,\text{Cov}(X_if(X_i))\mathbf{H}^{-1}]_{j,j}}
\end{aligned}$$

where the last line follows from properties of variance-covariance matrices. Note the similarities between the above expression for $\lambda^{\star}_j$ and equation (4) for mean estimation. As shown in Angelopoulos et al. (2024), when $\lambda = \lambda^{\star}_j$, PPI is always at least as precise as the human subject estimator for estimating the parameter $\theta^{\star}_j$.

The package `ppi_py` automatically implements power tuning for mean estimation, linear regression, logisitic regression, and Poisson regression. For examples, see the documentation `https://ppi-py.readthedocs.io/en/latest/index.html`.

S1.3 PPI Standard Error and Correlation

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

$$\text{se}(\hat{\theta}^{\text{PP}}) = \text{se}(\hat{\theta}^{H})\sqrt{1 - \frac{N}{n+N}\tilde{\rho}_{+}^{2}} \tag{S6}$$

This expression for the standard error is given in equation (5) and used in our power analysis. To define $\tilde{\rho}$ and prove equation (S6), we will use some results and concepts from Angelopoulos et al. (2024). In particular, equation (S6) 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 3.2. 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

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

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. More generally, if $\ell_\theta(x, y)$ is the negative log-likelihood in a generalized linear model (glm), then $\theta^\star$ is the vector of glm coefficients.

Angelopoulos et al. (2024) introduce a family of estimators $\hat{\theta}_\lambda^{\text{PP}}$ for $\theta^\star$. These estimators depend on a tuning parameter $\lambda \in [0, 1]$ and are given by

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

Taking $\lambda = 0$ corresponds to the classical—or human-subjects—estimator for $\theta^\star$

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

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

$$\sqrt{n}\left(\hat{\theta}_\lambda^{\text{PP}} - \theta^\star\right) \xrightarrow{d} \mathcal{N}(0, \Sigma^\lambda) \tag{S7}$$

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

$$\begin{aligned}\Sigma^\lambda =& H_{\theta^\star}^{-1}\,\text{Cov}(\nabla\ell_{\theta^\star})H_{\theta^\star}^{-1} + \lambda^2(1+r)H_{\theta^\star}^{-1}\,\text{Cov}(\nabla\ell_{\theta^\star}^{f})H_{\theta^\star}^{-1} \\ &- \lambda H_{\theta^\star}^{-1}\left(\text{Cov}(\nabla\ell_{\theta^\star}^{f}, \nabla\ell_{\theta^\star}) + \text{Cov}(\nabla\ell_{\theta^\star}, \nabla\ell_{\theta^\star}^{f})\right)H_{\theta^\star}^{-1},\end{aligned} \tag{S8}$$

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

By equation (S7) the standard error of the jth coordinate $\hat{\theta}_{\lambda,j}^{\text{PP}}$ is $\sqrt{\Sigma_{j,j}^\lambda/n}$. 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 (S6).

Power tuning (Angelopoulos et al., 2024, Section 6) chooses λ to minimize $\Sigma^{\lambda}_{j,j}$. This is equivalent to choosing λ to minimize the standard error of $\hat{\theta}^{\mathrm{PP}}_{\lambda,j}$. From (S8), 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

$$\tilde{\rho}_j = \frac{\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}, \nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j}}{\sqrt{\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j} \left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j}}}. \tag{S9}$$

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

Note that $\tilde{\rho}_j$ and $\lambda^\star_j$ always have the same sign. This is because they are related through the equation

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

The terms $1+r$, $\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell^f_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j}$ and $\left[H^{-1}_{\theta^\star} \operatorname{Cov}(\nabla \ell_{\theta^\star}) H^{-1}_{\theta^\star}\right]_{j,j}$ are all positive. This is because $r > 0$, and because the second two are diagonal entries of positive definite matrices. The $\tilde{\rho}_j = c\lambda^\star_j$ where $c > 0$. Thus, $\tilde{\rho}_j$ is zero, positive or negative if and only if $\lambda^\star_j$ is zero, positive or negative.

If we replace r with n/N, then the standard error of $\hat{\theta}^{\mathrm{PP}}_{\lambda^\star_j,j}$ from a sample of n labeled data points and N unlabeled data point becomes

$$\operatorname{se}(\hat{\theta}^{\mathrm{PP}}_{\lambda^\star_j,j}) = \sqrt{\Sigma^{\lambda^\star_j}_{j,j}/n} = \frac{\sigma_j}{\sqrt{n}} \sqrt{1 - \frac{N}{n+N} \tilde{\rho}^2_j} = \operatorname{se}(\hat{\theta}^H) \sqrt{1 - \frac{N}{n+N} \tilde{\rho}^2_j}. \tag{S10}$$

In practice, $\lambda_j^\star$ has to be estimated. Angelopoulos et al. (2024) provide a consistent estimator $\hat{\lambda}_j$ for $\lambda_j^\star$ and show that $\hat{\theta}^{\text{PP}}_{\hat{\lambda}_j,j}$ achieves the same asymptotic variance as $\hat{\theta}^{\text{PP}}_{\lambda_j^\star,j}$.

One problem is that the optimal value $\lambda_j^\star$ or the estimator $\hat{\lambda}_j$ may lie outside of the interval $[0,1]$. When this occurs Angelopoulos et al. (2024) clip the $\hat{\lambda}_j$ to lie between $[0,1]$. This may reduce the precision of PPI but Angelopoulos et al. (2024) note this rarely occurs in practice and propose a "one-step" estimator that obtains that same standard error as $\hat{\theta}^{\text{PP}}_{\lambda_j^\star}$ even when $\lambda_j^\star$ lies outside of $[0,1]$. In our power analysis, we assume that $\lambda_j^\star$ is always at most 1 or that the one-step estimator is used. As noted above, when $\lambda_j^\star < 0$ we also have that $\tilde{\rho}_j < 0$ and in this case the PPI estimator equals the classical estimator $\hat{\theta}_j^H$ with standard error $\text{se}(\hat{\theta}_j^H)$. This gives the the following expression for $\text{se}(\hat{\theta}^{\text{PP}}_{\hat{\lambda}_j,j})$

$$\text{se}(\hat{\theta}^{\text{PP}}_{\lambda_j,j}) = \text{se}(\hat{\theta}^H)\sqrt{1 - \frac{N}{n+N}(\tilde{\rho}_j)_+^2} = \begin{cases} \text{se}(\hat{\theta}^H) & \text{if } \tilde{\rho}_j \leq 0, \\ \text{se}(\hat{\theta}^H)\sqrt{1 - \frac{N}{n+N}\tilde{\rho}_j^2} & \text{if } \tilde{\rho}_j > 0. \end{cases} \tag{S11}$$

The difference between equations (S10) and (S11), is that (S10) applies when $\lambda_j^\star \in [0,1]$ or when the one-step estimator is used. In this case PPI is strictly more precise than classical inference if and only if $\tilde{\rho}_j \neq 0$. This is because $\tilde{\rho}_j$ is squared. Thus, strictly negative values of $\tilde{\rho}_j$ can improve inference if $\lambda_j^\star$ is allowed to be negative. When the one-step estimator is not used, equation (S10) is the correct formula for the standard error. In this case, PPI is strictly more precise if and only if $\tilde{\rho}_j > 0$. When $\tilde{\rho}_j$ is negative, the predictions are ignored and PPI reduces to classical inference.

Finally, to simplify notation, let $\hat{\theta}^{\text{PP}} = \hat{\theta}^{\text{PP}}_{\hat{\lambda}_j,j}$ and $\tilde{\rho} = \tilde{\rho}_j$. With this notation

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

as claimed in equations (5) and (S6).

S1.4 Estimating the PPI Correlation

Recall that

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

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 $H_{\theta^\star} = \mathbb{E}[\nabla^2 \ell_{\theta^\star}(X,Y)]$. To estimate $\tilde{\rho}_j$ from mixed subjects data $\{(X_i, Y_i, f(X_i))\}_{i=1}^n$ we first estimate $\theta^\star$ with the human subjects estimator

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

We then estimate $H_{\theta^\star}$ with $\widehat{H}_{\hat{\theta}^H} = \frac{1}{n}\sum_{i=1}^{n} \nabla^2 \ell_{\hat{\theta}^H}(X_i, Y_i)$ and likewise use the following sample covariance estimates

$$\widehat{\mathrm{Cov}}(\nabla \ell^f_{\hat{\theta}^H}, \nabla \ell_{\hat{\theta}^H}) = \frac{1}{n}\sum_{i=1}^{n} \left(\nabla \ell_{\hat{\theta}^H}(X_i, f(X_i)) - \overline{\ell^f_{\hat{\theta}^H}}\right)\left(\nabla \ell_{\hat{\theta}^H}(X_i, Y_i) - \overline{\ell_{\hat{\theta}^H}}\right)^\top$$

$$\widehat{\mathrm{Cov}}(\nabla \ell^f_{\hat{\theta}^H}) = \frac{1}{n}\sum_{i=1}^{n} \left(\nabla \ell_{\hat{\theta}^H}(X_i, f(X_i)) - \overline{\ell^f_{\hat{\theta}^H}}\right)\left(\nabla \ell_{\hat{\theta}^H}(X_i, f(X_i)) - \overline{\ell^f_{\hat{\theta}^H}}\right)^\top$$

$$\widehat{\mathrm{Cov}}(\nabla \ell_{\hat{\theta}^H}) = \frac{1}{n}\sum_{i=1}^{n} \left(\nabla \ell_{\hat{\theta}^H}(X_i, Y_i) - \overline{\ell_{\hat{\theta}^H}}\right)\left(\nabla \ell_{\hat{\theta}^H}(X_i, Y_i) - \overline{\ell_{\hat{\theta}^H}}\right)^\top$$

where $\overline{\ell^f_{\hat{\theta}^H}} = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_{\hat{\theta}^H}(X_i, f(X_i))$ and $\overline{\ell_{\hat{\theta}^H}} = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_{\hat{\theta}^H}(X_i, Y_i)$. The estimator for $\tilde{\rho}_j$ is then

$$\hat{\rho}_j := \frac{\left[\widehat{H}^{-1}_{\hat{\theta}^H} \widehat{\mathrm{Cov}}(\nabla \ell^f_{\hat{\theta}^H}, \nabla \ell_{\hat{\theta}^H}) \widehat{H}^{-1}_{\hat{\theta}^H}\right]_{j,j}}{\sqrt{\left[\widehat{H}^{-1}_{\hat{\theta}^H} \widehat{\mathrm{Cov}}(\nabla \ell^f_{\hat{\theta}^H}) \widehat{H}^{-1}_{\hat{\theta}^H}\right]_{j,j} \left[\widehat{H}^{-1}_{\hat{\theta}^H} \widehat{\mathrm{Cov}}(\nabla \ell_{\hat{\theta}^H}) \widehat{H}^{-1}_{\hat{\theta}^H}\right]_{j,j}}}$$

Since $\hat{\theta}^H$ is a consistent estimator for $\theta^\star$, each of the estimators $\widehat{H}_{\hat{\theta}^H}$, $\widehat{\mathrm{Cov}}(\nabla \ell^f_{\hat{\theta}^H}, \nabla \ell_{\hat{\theta}^H})$, $\widehat{\mathrm{Cov}}(\nabla \ell^f_{\hat{\theta}^H})$ and $\widehat{\mathrm{Cov}}(\nabla \ell_{\hat{\theta}^H})$ are also consistent. This implies that $\hat{\rho}_j$ is a consistent estimator for $\tilde{\rho}_j$. We do not have a derivation of the sampling distribution for $\hat{\rho}_j$, but it can be approximated using the bootstrap. When using the bootstrap, we recommend applying Fisher's z-transformation to $\hat{\rho}_j$ which should stabilize the variance of $\hat{\rho}_j$ (Fisher, 1915).

S1.5 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 $\sigma/\sqrt{n_0}$ and the PPI standard error is

$$\frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{N}{n+N}\tilde{\rho}^2_+}$$

Equating this two standard errors gives

$$n_0 = n \cdot \frac{n+N}{n + N(1-\tilde{\rho}^2_+)}$$

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

$$\frac{1+k}{1+k(1-\tilde{\rho}^2_+)} \geq 1$$

When $\tilde{\rho} = 1$, 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 $\tilde{\rho} \leq 0$, this factor equals 1 and the effective sample size is n meaning that only the labeled samples are used.

Our power analysis offer researchers a range of designs each of which have the same effective sample size. The effective sample size for PPI is

$$\mathrm{ESS}(n, N) = n \cdot \frac{n+N}{n + N(1-\tilde{\rho}^2_+)}$$

Setting $\text{ESS}(n, N) = n_0$ and rearranging for N gives

$$n_0(1 - \tilde{\rho}_+^2) < n \leq n_0 \quad \text{and} \quad N = n \cdot \frac{n_0 - n}{n - n_0(1 - \tilde{\rho}_+^2)}. \tag{S12}$$

Each of the pairs (n, N) satisfying (S12) give PPI estimators with the same standard error and level of precision.

S1.6 Computing a Desired Effective Sample Size

To compute the desired effective sample size researchers should perform a classical power analysis. This power analysis depends on the confidence level α, the level of power $1-\beta$, the effect size and the variance of the human subjects estimator. If the hypothesized effect size is $\delta = \theta_1 - \theta_0$ and the standard error of the human subjects estimator $\hat{\theta}^H$ is $\sigma/\sqrt{n}$, then the desired effective sample size is the solution to the following equation:

$$\mathbb{P}(|\delta/\sigma\sqrt{n_0} + Z| > z_{1-\alpha/2}) = 1 - \beta$$

where $Z \sim \mathcal{N}(0, 1)$ and $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. Python code that computes n_0 as a function of δ, σ, α and β is given below.

```python
from scipy.optimize import brentq
from scipy.stats import norm, foldnorm

def effective_n_from_effect_size(delta, sigma, power, alpha=0.05):
    assert 0 < power < 1, "Power must be between 0 and 1"
    assert 0 < alpha < 1, "Level must be between 0 and 1"
    assert 0 != delta, "Effect size must be non-zero"
    assert power > alpha, "Power must be greater than alpha level"
    q = norm.ppf(1 - alpha / 2)
    beta = 1 - power

    def f(mu):
        return foldnorm.cdf(q, c=mu) - beta

    a = 0
    b = q - norm.ppf(beta)
    mu_star = brentq(f, a, b)
    d = delta/sigma
    effective_n = (mu_star/d)**2
    return round(effective_n)
```

Both δ and σ can be estimated using the `ppy_py` package. This is illustrated below with pseudo code. The text `[estimand]` should be replaced with one of `mean`, `ols`, `logisitic` or `poisson` depending on the parameter of interest, `coord` is an integer that refers to the position of the parameter estimate $\hat{\theta}$ in a vector of estimates. The number `theta_0` is the value of θ under the null hypothesis. Each of `X`, `Y`, `Yhat`, `X_unlabeled` and `Yhat_unlabeled` refer to previously collected data, for example from a pilot study.

```python
from ppi_py import ppi_[estimand]_pointestimate, ppi_[estimand]_ci


coord   = ...
theta_0 = ...
theta_1 = ppi_[estimand]_pointestimate(X,
                                       Y,
                                       Yhat,
```

```
                                        X_unlabeled,
                                        Yhat_unlabeled,
                                        lam = 0)[coord]
delta = theta_1 - theta_0
lower, upper = ppi_[estimand]_ci(X,
                                 Y,
                                 Yhat,
                                 X_unlabeled,
                                 Yhat_unlabeled,
                                 lam = 0,
                                 alpha = 0.05)
se = (upper[coord] - lower[coord])/2/1.96
sigma = se * X.shape[0]**0.5

level = 0.05
power = 0.9
effective_n = effective_n_from_effect_size(delta, sigma, power, level)
```

S1.7 The Cost of PPI and Classical Inference

In this section, we determine when PPI is more cost-effective than classical inference and the percentage of cost saved by PPI as reported in Figure 4 of the main text. This is done by finding the most cost-effective design that achieves a desired effective sample size. These calculations are then used in our power analysis.

Recall that the possible designs (n, N) achieving an effective sample size of n_0 are given by

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

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 using PPI 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 (S12) and minimizes $C(n, N)$. That is, $(n^\star, N^\star)$ is the solution to the optimization problem

$$\begin{array}{ll} \text{minimize} & C(n, N) \\ \text{subject to} & \text{equation (S12)} \end{array}$$

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

$$n^\star = n_0\left(1 - \tilde{\rho}_+^2 + \sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}_+^2)}\right) \quad \text{and} \quad N^\star = \frac{n^\star(n_0 - n^\star)}{n^\star - (1-\tilde{\rho}_+^2)n_0}. \tag{S13}$$

For this pair to be a valid design, we need that $n^\star$ and $N^\star$ are both integers and that $n_0 \geq n^\star > n_0(1-\tilde{\rho}_+^2)$. To make $n^\star$ and $N^\star$ integers, we simply round the values in (S13). The constraint $n^\star > n_0(1-\tilde{\rho}_+^2)$ and

$n^\star \leq n_0$ are equivalent to $\gamma > 0$ and $\tilde{\rho}_+^2 \geq \gamma/(1+\gamma)$.

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

$$C(n^\star, N^\star) = c_Y n_0 \left(1 - \tilde{\rho}_+^2 + \gamma\tilde{\rho}_+^2 + 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}_+^2)}\right). \tag{S14}$$

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 is is because with classical inference 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

$$c_Y \left(1 - \tilde{\rho}_+^2 + \gamma\tilde{\rho}_+^2 + 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}_+^2)}\right) \leq c_Y + c_X \quad \text{and} \quad \tilde{\rho}_+^2 \geq \frac{\gamma}{1+\gamma}.$$

Furthermore, when these conditions are satisfied, the optimal *absolute* cost savings from PPI are

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

The *relative* cost savings from PPI are

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

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

$$\tilde{\rho}_+^2 - \gamma\tilde{\rho}_+^2 - 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}_+^2)} > 0 \quad \text{and} \quad \tilde{\rho}_+^2 > \frac{\gamma}{1+\gamma}.$$

The first inequality is equivalent to

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

which implies the inequality $\tilde{\rho}_+^2 > \frac{\gamma}{1+\gamma}$ provided $\gamma \in [0,1]$ which occurs as long as $c_f \leq c_Y$ which is necessary for PPI to more cost effective than classical inference. Thus, the PPI correlation must be sufficiently high compared to the relative cost of collecting a labeled or unlabeled sample.

When $c_X = 0$ and $\tilde{\rho} > \frac{2\sqrt{\gamma}}{1+\gamma}$, the expressions for the absolute and relative cost reductions simplify and become

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

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

$$\frac{C(n^\star, N^\star)}{C_0(n_0)} = 1 - \tilde{\rho}_+^2(1-\gamma) + 2\sqrt{\gamma\tilde{\rho}_+^2(1-\tilde{\rho}_+^2)}. \tag{S15}$$

The curves in Figure 4 plot equation (S15) as a function of $1/\gamma = c_Y/c_f$ for different values of $\tilde{\rho}$.

S1.8 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 largest effective sample size subject to a budget constraint. Once the most powerful pair has been computed, the effective sample size can be estimated.

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.

Section S1.7 determined when classical inference is 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

$$c_Y\left(1 - \tilde{\rho}_+^2 + \gamma\tilde{\rho}_+^2 + 2\sqrt{\gamma\tilde{\rho}_+^2(1-\tilde{\rho}_+^2)}\right) > c_Y + c_X \quad \text{or} \quad \tilde{\rho}_+^2 < \frac{\gamma}{1+\gamma}$$

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 (S13) and (S14). This means that if

$$c_Y\left(1 - \tilde{\rho}_+^2 + \gamma\tilde{\rho}_+^2 + 2\sqrt{\gamma\tilde{\rho}_+^2(1-\tilde{\rho}_+^2)}\right) \leq c_Y + c_X \quad \text{and} \quad \tilde{\rho}_+^2 \geq \frac{\gamma}{1+\gamma}.$$

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 $(n^\star, N^\star)$ has been computed, the effective sample size of $(n^\star, N^\star)$ can be computed. If classical inference is used, then the effective sample size is

$$\text{ESS}(n^\star, N^\star) = \frac{B}{c_X + c_Y}.$$

If PPI is used, then the effective sample size is

$$\text{ESS}(n^\star, N^\star) = \frac{B}{c_Y\left(1 - \tilde{\rho}_+^2 + \gamma\tilde{\rho}_+^2 + 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}_+^2)}\right)}.$$

S1.9 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 effective sample size at the lowest cost. 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 costs c_Y, c_f, c_X and the desired effective sample size n_0. As before, the parameters $\tilde{\rho}$ is estimated from a dataset and c_Y, c_f, c_X and n_0 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

$$c_Y \left(1 - \tilde{\rho}_+^2 + \gamma\tilde{\rho}_+^2 + 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}_+^2)}\right) > c_Y + c_X \quad \text{or} \quad \tilde{\rho}_+^2 < \frac{\gamma}{1+\gamma},$$

then

$$n^\star = n_0 \quad \text{and} \quad N^\star = 0.$$

When PPI is more cost-effective than classical inference, then $(N^\star, n^\star)$ should be chosen so that $\text{ESS}(n^\star, N^\star) = n_0$ where $n^\star$ and $N^\star$ are as in equation (S13). Thus, if

$$c_Y \left(1 - \tilde{\rho}_+^2 + \gamma\tilde{\rho}_+^2 + 2\sqrt{\gamma\tilde{\rho}^2(1-\tilde{\rho}_+^2)}\right) \leq c_Y + c_X \quad \text{and} \quad \tilde{\rho}_+^2 \geq \frac{\gamma}{1+\gamma},$$

then

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

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

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

S2 Simulation Study

In this section, we conduct a simulation to illustrate the problems associated with silicon sampling (Section 2.2) and demonstrate how a mixed subjects design with PPI can address these issues (Section 3.1). We examine the impact of biased predictions when estimating regression coefficients as we increasingly rely on LLMs to estimate parameters. We define the true coefficient to be $\theta = 1$ and introduce a small bias of $b = 0.1$, constituting 10% of the effect. We demonstrate that even such a small bias can substantially compromise the validity of conclusions drawn from silicon sampling.[2] As such, the purpose of this simulation study parallels the one using empirical data from the Moral Machine experiment (Section 4). However, the pure simulation presented in the following affords full transparency over the data-generating process and thus complements the simulation with empirical data.

The data for this simulated example are generated using the following linear models:

[2]Similarly, even small inaccuracies of LLMs in predicting the categories of documents can lead to biased point estimates if those classifications are used in regression analyses (Egami et al., 2024).

$$
\begin{aligned}
Y_i &= \theta_0 + \theta_1 X_i + Z_{i,1} \\
f(X_i) &= \theta_0 + (\theta_1 + b) X_i + Z_{i,2} \\
f(\widetilde{X}_i) &= \theta_0 + (\theta_1 + b) \widetilde{X}_i + \widetilde{Z}_i,
\end{aligned}
$$

where

$$
\begin{aligned}
X_i &\overset{\text{iid}}{\sim} \mathcal{N}(0,1) \\
\begin{bmatrix} Z_{i,1} \\ Z_{i,2} \end{bmatrix} &\overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \right) \\
\widetilde{X}_i &\overset{\text{iid}}{\sim} \mathcal{N}(0,1) \\
\widetilde{Z}_i &\overset{\text{iid}}{\sim} \mathcal{N}(0,1).
\end{aligned}
$$

The parameters θ_0, θ_1 are the true parameters and b represents the bias in silicon sampling parameters. Note that $X_i \overset{d}{=} \widetilde{X}_i$ and $f(X_i) \overset{d}{=} f(\widetilde{X}_i)$ originate from the same distribution, ensuring that the assumptions of PPI hold. Moreover, the error terms in the two models are correlated. In practice, this correlation arises from relevant information that the LLM has about the relationship between Y_i and X_i, beyond what is captured by the variables used in the regression to estimate parameters. The extent to which the LLM provides relevant information about this relationship is reflected in the PPI correlation. In this simulation, the PPI correlation can be computed analytically and it is equal to

$$
\tilde{\rho} = \frac{r}{\sqrt{2b^2 + 1}}.
$$

For our simulation, we assume that the LLM is informative ($\tilde{\rho} = 0.9$).

We generate $n = 1,000$ observations for the labeled datasets $\{(X_i, Y_i, f(X_i))\}_{i=1}^{n}$. To create the unlabeled data, we generate datasets $\{(\widetilde{X}_i, f(\widetilde{X}_i))\}_{i=1}^{N}$ for $N = n \times k$ where $k = (0.1, 0.25,\ 0.5, 0.75,\ 1, 1.5, \ldots, 9.5, 10)$. For the silicon subjects approach, we take these unlabeled datasets of size N as the silicon samples.

For the silicon subjects approach, we use an OLS regression to estimate the parameter θ with the sample of N silicon subjects. For the mixed-subjects design, we apply PPI using the labeled and unlabeled datasets. We repeat the process of generating samples and estimating coefficients 1,500 times. We average across these repetitions to estimate the bias $\mathbb{E}[\hat{\theta} - \theta]$ of the coefficient estimate, the standard error $\text{se}(\hat{\theta})$, the percentage of confidence intervals that cover θ (i.e., the coverage rate), and the Root Mean Square Error RMSE $= \sqrt{b^2 + \text{se}(\hat{\theta})^2}$.

Across all sample sizes of silicon subjects N, the average bias in silicon sampling is close to the bias of 0.1 defined for our simulation, whereas the average bias for PPI is close to zero (Figure S1a). As equations (5) and S6 suggests, the PPI standard errors stabilize, preventing overconfidence in parameter estimates. As N increases, the standard error for the silicon sampling estimate decreases with the sample size. The RMSE is approximately equal to the bias as the standard error shrinks toward zero (RMSE $\approx |b|$). These dynamics impact the validity of conclusions. The shrinking of the standard of the standard error results in narrower confidence intervals for the silicon sampling estimate. Due to the bias, these narrow confidence intervals have incorrect centers. Figure S1c shows that coverage rates decrease rapidly. When standard errors are much smaller than the bias, the coverage rate approaches zero. For instance, in our simulation when $N = 4{,}000$ and the standard error is less than one-fifth of the effect size, we have about

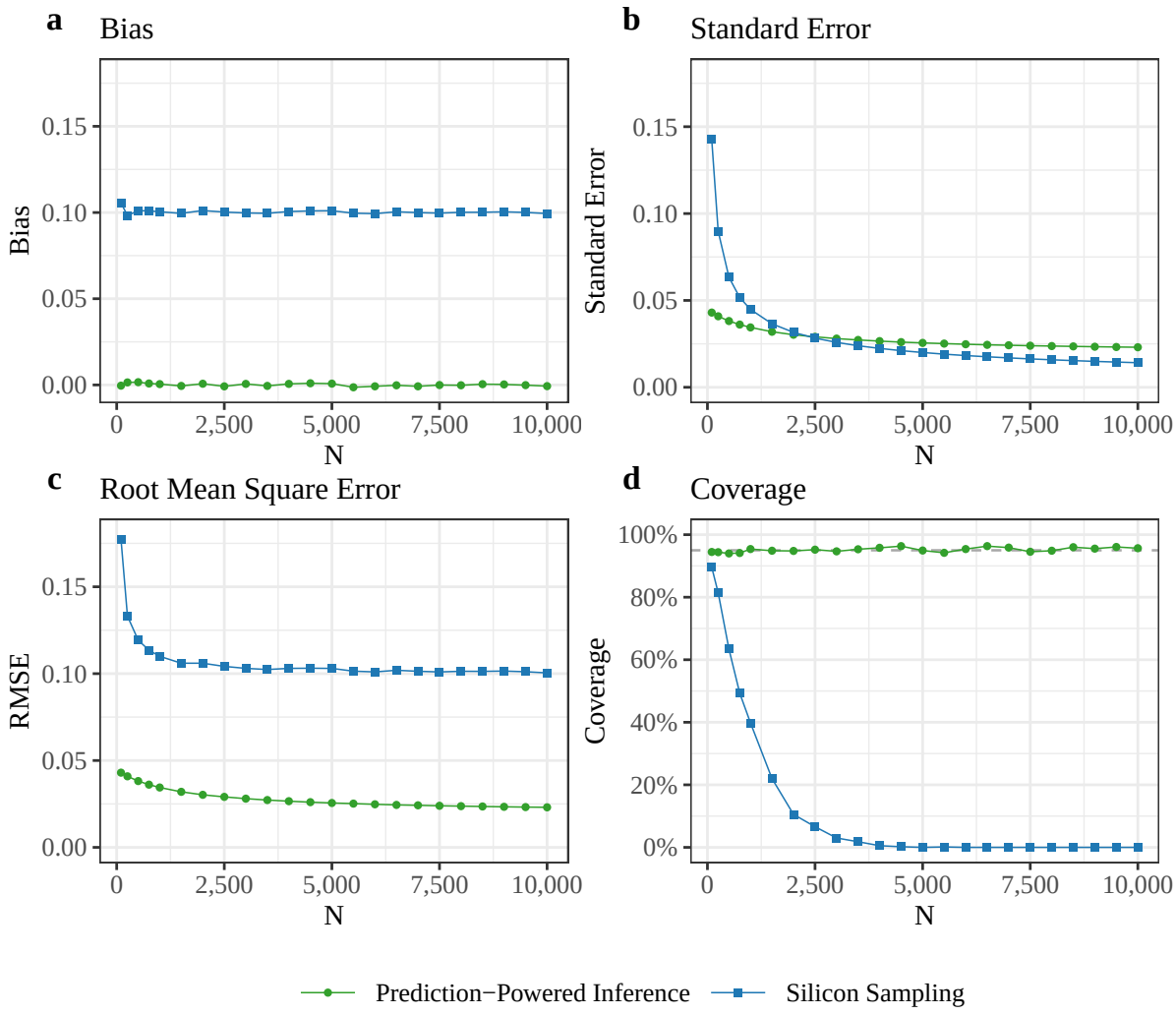

Figure S1: Results from a simulation demonstrating the estimation of a regression coefficient using biased predictions from an LLM. Unlike the coefficient estimates from PPI, the estimates from the silicon sampling approach are biased (**a**, **c**). As standard errors shrink with the number of LLM predictions N (**b**), the confidence intervals from silicon sampling become more narrowly centered around incorrect coefficient estimates. The combination of bias and overconfidence results in a lower percentage of confidence intervals that cover the true parameter. In contrast, PPI estimates maintain a nominal coverage rate of 95% (**d**).

zero coverage. At the same time, PPI maintains a nominal coverage rate of 95%.

In this simulation, we demonstrated even small biases of 10% can compromise the validity of conclusions about parameters estimated with the silicon subjects design. Biases lead to the over- or underestimation of the parameter and decrease the coverage of confidence intervals. PPI estimates are not biased and therefore maintain nominal coverage.

S3 Details on the Moral Machine Experiment

In this section, we provide additional details about the empirical application of the mixed subjects and silicon subjects design to the Moral Machine experiment.

S3.1 Example of a Moral Dilemma

Figure S2 illustrates a dilemma presented to a survey respondent in the Moral Machine experiment. The image was created with the scenario design tool on moralmachine.net.

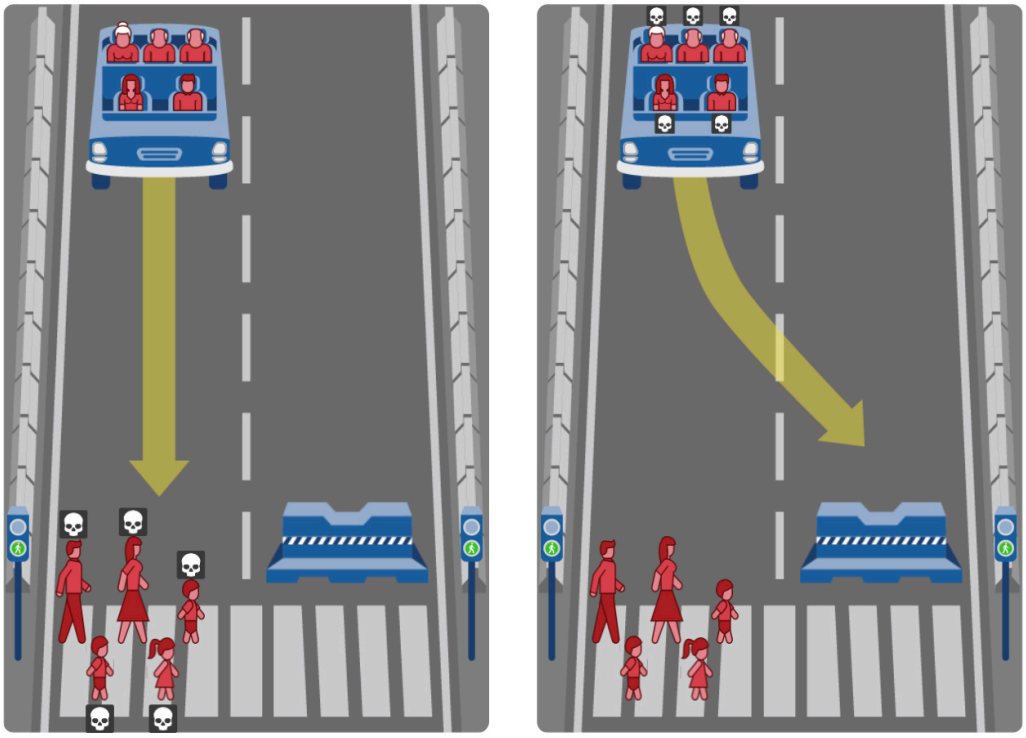

Figure S2: Example dilemma presented to a survey respondent in the Moral Machine Experiment.

S3.2 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.

S3.3 Prompting LLMs to Predict Decisions to Moral Dilemmas

Table S1 provides an overview of the LLMs considered in our analysis. We prompted several models with subsets of the data from the Moral Machine experiment to explore which LLM best predicts decisions to moral dilemmas. While the PPI correlation was the main criterion for selecting GPT-4 Turbo for our study, we also report the Pearson correlation between predicted and observed decisions.

Additionally, we detail the number of API requests for each LLM. Note that each dilemma in the Moral Machine experiment presented respondents with two options. Each dilemma thus resulted in two outcomes, sparing one group and harming the other. Therefore, creating the sample 1,163,962 of predicted decisions from U.S. respondents required half as many API requests.

Language Model	**Context window**	**Training data**	**Input cost 1,000 tokens**	**Output cost 1,000 tokens**	**API requests**	**Pearson Correlation**
gpt-4-turbo	128,000 tokens	Dec 2023	$0.01	$0.03	581,981	0.358
gpt-4o	128,000 tokens	Oct 2023	$0.005	$0.0150	22,315	0.311
gpt-3.5-turbo-0125	16,385 tokens	Sep 2021	$0.0005	$0.0015	22,315	0.113

Table S1: Details on LLMs used to predict decisions to moral dilemmas in the Moral Machine experiment

S3.4 Summary Statistics on Demographics

Figure S3 summarizes the demographic distribution in the sample used in our reanalysis of the Moral Machine experiment in Section 4.

We used the subset of 55,893 Americans who completed an optional demographic survey for our analysis. On average, these participants evaluated 10.4 dilemmas, each with two options. Each of these dilemmas presents participants with two options, resulting in a total sample size of 1,163,962 decisions.

While these decisions represent our primary unit of analysis, we also provide additional details on the demographics of the survey respondents. Compared to demographic distribution in the American Community Survey in 2016, the U.S. sample from the Moral Machine experiment overrepresents young, male, more educated, and individuals with higher incomes.

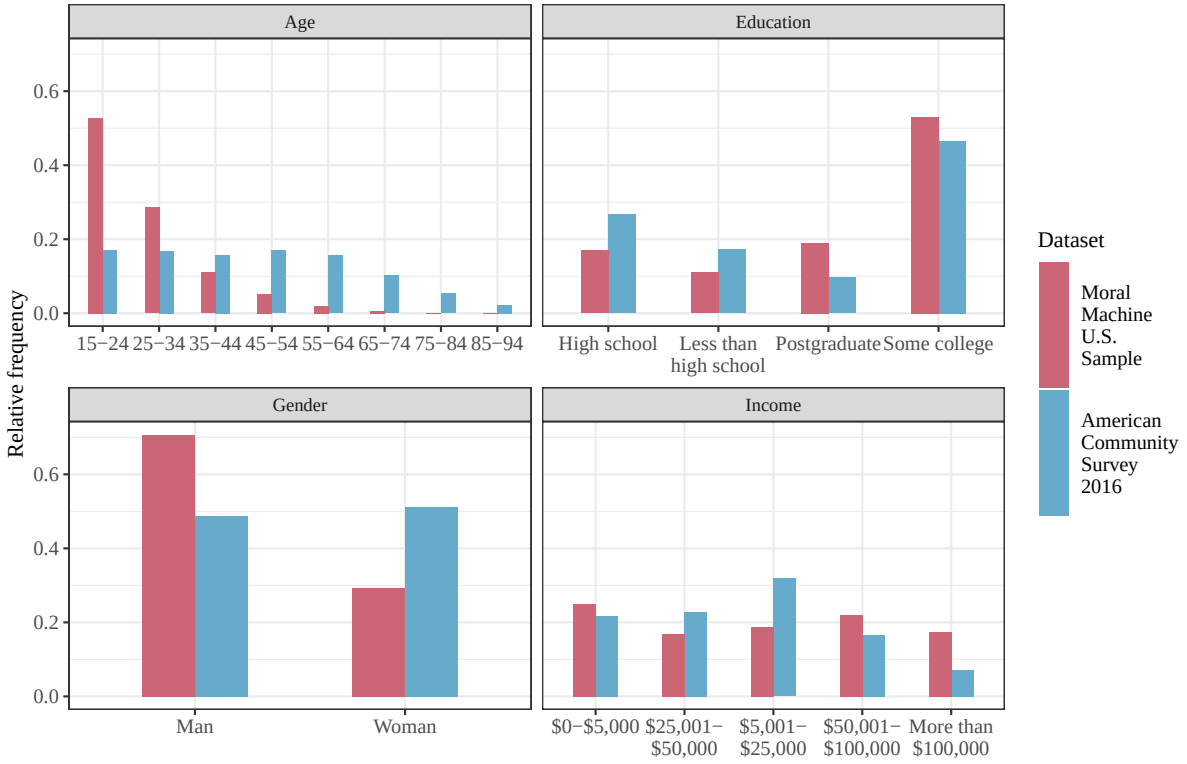

Figure S3: Comparison of demographic distributions of the U.S. sample from the Moral Machine experiment and the 2016 American Community Survey

S3.5 AMCE Estimates

Figure S4 shows the estimated causal effects of scenario attributes on participants' decision to save characters. We compare three sets of AMCE estimates.

Human subjects U.S. sample (red): We estimate AMCEs based on 1,163,962 decisions from a sample of U.S. respondents. These estimates represent the human subjects approach and serve as the ground truth for our simulation in Section 4.

Silicon Subjects Approach (blue): We compute AMCEs based on the 1,163,962 decisions predicted by GPT4-Turbo, representing the silicon subjects approach. While this method yields AMCEs similar to the human subjects approach in some instances, the LLM-derived AMCEs often differ notably.

Human subjects across countries (yellow): For completeness, we also present the AMCEs reported by Awad et al. (2018). These estimates are based on a sample from multiple countries. Therefore, the estimates are not expected to align perfectly with those from the sample of U.S. respondents (red). However, the observed differences are generally small.

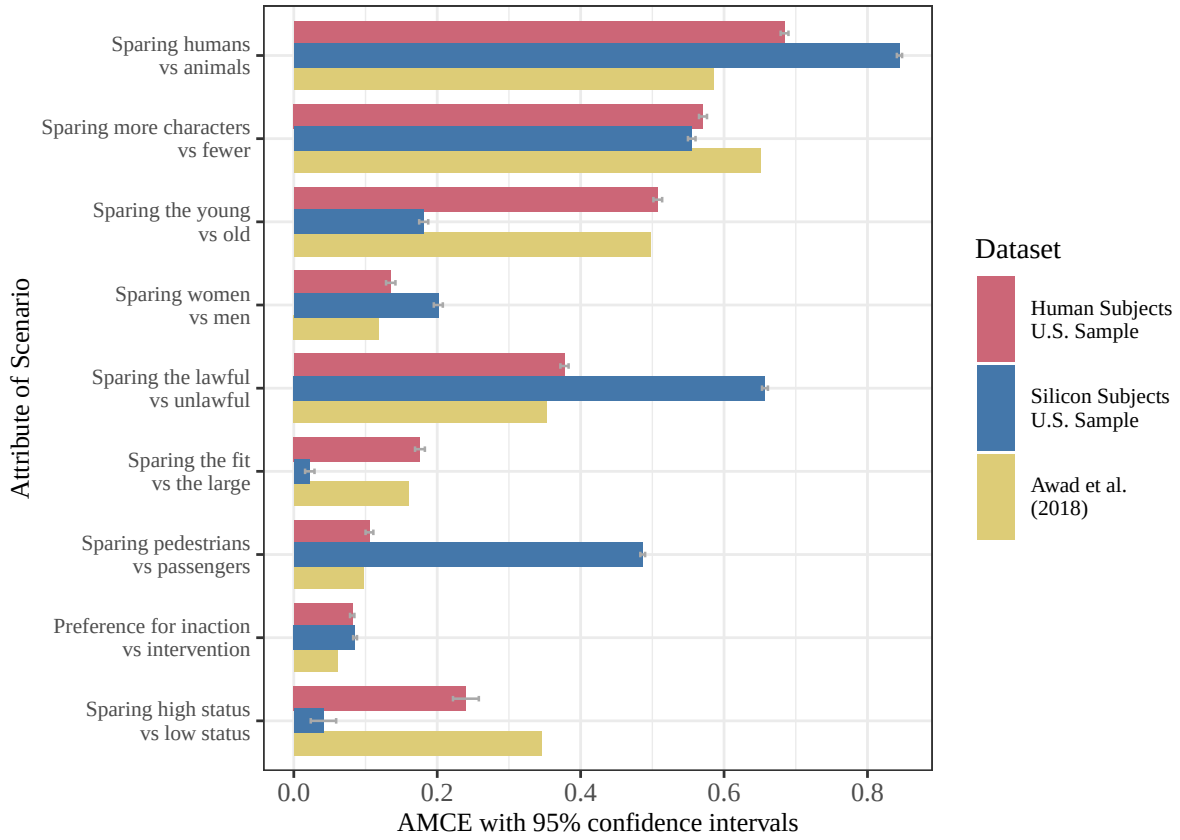

Figure S4: Comparison of AMCE estimates from human subjects against silicon subjects. Note that Awad et al. (2018) do not report confidence intervals due to their negligible width, a consequence of the large sample size.

S3.6 Effective sample size

Table S2 reports the PPI correlations and effective sample sizes. To calculate the effective sample size, we used equation (6) with $n = 10{,}000$ human subjects and $N = 100{,}000$ silicon subjects.

Label	PPI correlation	Effective sample size
Preference for inaction vs intervention	0.353	11,275
Sparing pedestrians vs passengers	0.314	10,986
Sparing the lawful vs unlawful	0.309	10,950
Sparing the fit vs the large	0.283	10,786
Sparing women vs men	0.263	10,670
Sparing high status vs low status	0.230	10,504
Sparing the young vs old	0.198	10,371
Sparing more characters vs fewer	0.192	10,347
Sparing humans vs animals	0.049	10,022

Table S2: Effective sample size for each attribute in the Moral Machine experiment

S4 Example Power Analysis

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

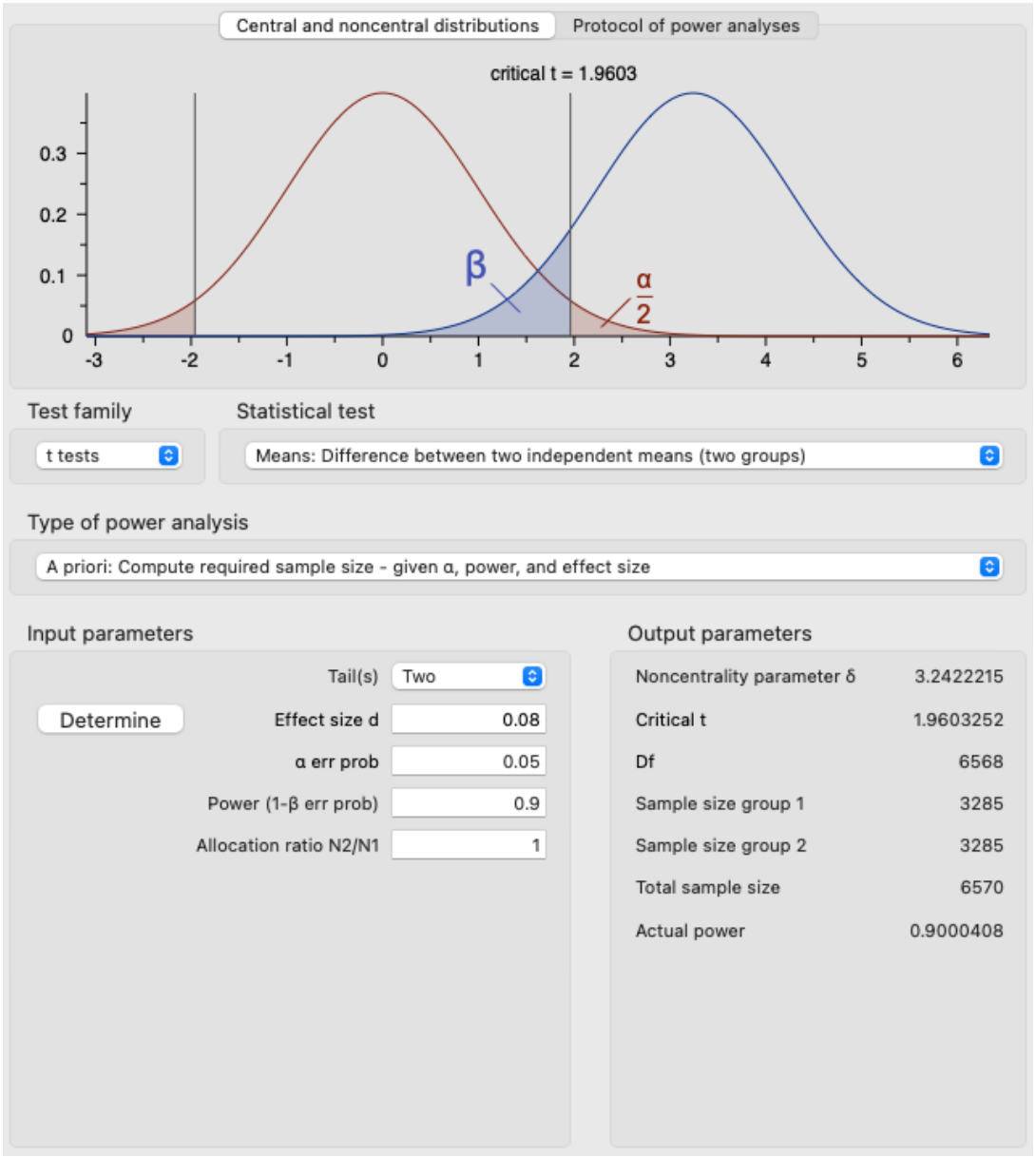

Figure S5: Power analysis with G*Power

References

Angelopoulos, A. N., Duchi, J. C., and Zrnic, T. (2024). Ppi++: Efficient prediction-powered inference.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563(7729):59–64.

Egami, N., Hinck, M., Stewart, B. M., and Wei, H. (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, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4):1149–1160.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.