Article

Sociological Methods & Research
2025, Vol. 54(3) 1156–1196

DOI: 10.1177/00491241251330582
journals.sagepub.com/home/smr

Machine Bias. How Do Generative Language Models Answer Opinion Polls?[1]

Julien Boelaert[1], Samuel Coavoux[2], Étienne Ollion[3], Ivaylo Petev[4], and Patrick Präg[2]

Abstract

Generative artificial intelligence (AI) is increasingly presented as a potential substitute for humans, including as research subjects. However, there is no scientific consensus on how closely these in silico clones can emulate survey respondents. While some defend the use of these "synthetic users," others point toward social biases in the responses provided by large language models (LLMs). In this article, we demonstrate that these critics are right to be wary of using generative AI to emulate respondents, but probably not for the right reasons. Our results show (i) that to date, models cannot replace research subjects for opinion or attitudinal research; (ii) that they display a strong bias and a low variance on each topic; and (iii) that this bias randomly varies from one topic to the next. We label this pattern "machine bias," a concept we define, and whose consequences for LLM-based research we further explore.

[1]CERAPS, Faculté des sciences juridiques politiques et sociales, Université de Lille, France
[2]CREST, ENSAE, Institut Polytechnique de Paris, Paris, France
[3]CREST, Ecole Polytechnique, Institut Polytechnique de Paris, Paris, France
[4]CREST, CNRS, Institut Polytechnique de Paris, Paris, France

Corresponding Author:
Julien Boelaert, CERAPS. 1 place Déliot, 59000 Lille, France.
Email: julien.boelaert@univ-lille.fr

Data Availability Statement included at the end of the article

Keywords

LLMs, bias, generative artificial intelligence, computational social sciences, machine learning, survey research

Introduction

The release of ChatGPT in November 2022 has thrust debates about generative artificial intelligence (AI) onto the public stage. Our collective imagination has been running wild ever since. Optimistic visions of streamlined, accelerated work and life are met with fears of omnipotent artificial intelligence and of mass disappearance of jobs. Social scientists have seized the moment to discuss the implications of these technologies on various occupations (Brynjolfsson, Li, and Raymond 2023), including their own work (Ziems et al. 2024). Generative AI holds a promise to change the way we conduct social science research (Bail 2024). It could perform literature reviews, annotate documents, improve computer programming, generate research hypotheses, and even emulate human behavior in studies.

In the present study we focus on this latter aspect, and more specifically on the capacity of large language models (LLMs) to simulate human populations for opinion research. Surveys are an essential tool in social science research, as well as in commercial marketing, political campaigning, forecasting, etc. But high-quality surveys—robust questionnaire design, repeated measures, representative samples—demand advanced skills, take time, and are costly. They are also in crisis. Survey research response rates have been declining for decades (to as low as 1 percent in the polls for the 2024 U.S. presidential election) (Williams and Brick 2018), which to some extent stems from the decline of landline phones and associated problems of drawing random samples (Dutwin and Buskirk 2021), and issues of motivating participants to fill in questionnaires online (Daikeler, Bošnjak, and Lozar Manfreda 2019). Massive cohort study endeavors such as the U.S. National Children's Study and the U.K. Life Study had to be canceled in the last decade due to lack of volunteer participants (Pearson 2015). These low response rates hinder their representativity, especially for hard-to-reach parts of the population.

Academics have started to explore the potential of generative AI, more specifically LLMs, to generate human-like responses. The argument is the following: because they are trained on massive volumes of human-produced data, LLMs could, if prompted properly and when provided with the right context, replicate the answers actual humans would give. They would also

do so for a fraction of the time and cost associated with standard surveys. Although most have remained cautious in their conclusions and admitted that there is uncertainty as to LLMs' abilities, hopes were high and pointed to a possible replacement of human subjects.[2]

Some commercial actors were not so prudent and immediately followed suit by offering solutions to complement or replace human samples.[3] It is therefore important to verify these claims: should LLMs manage to imitate humans, if only in part, the consequences for research and marketing alike would be immense. We label the idea that LLMs can accurately simulate human populations as the "representative hypothesis."

Despite, or maybe because of this bold promise, many scholars started questioning the direct use of LLMs as research subjects (Dominguez-Olmedo, Hardt, and Mendler-Dünner 2024; Santurkar et al. 2023; Wang, Morgenstern, and Dickerson 2024). Their findings pointed to a mismatch between results generated by a machine and responses provided by humans. Several studies blamed the ingrained social bias of LLMs. Trained on data that are disproportionately produced by geographically and socially situated populations, LLMs tend, they argued, to propagate the point of view of these social groups. This line of reasoning echoes a now classic finding about the algorithmic discrimination of minority groups (O'Neil 2016). Just like widely used predictive tools based on statistical learning (Buolamwini and Gebru 2018), LLMs could harbor massive social biases (Bender et al. 2021). Together, these elements cast a shadow over LLMs' ability to effectively mimic diverse human populations in survey simulations. While they might represent some groups properly, they may be failing others. We label this view the "social bias hypothesis."[4]

In this article, we propose a third alternative. Based on a targeted experiment, our results show that LLMs do not accurately represent subpopulations, and thus cannot generally serve as in silico research subjects for survey research. At least, current models cannot do so in the classical "zero-shot" setting advocated by some.[5] The reason, however, is not primarily social bias. Our experiment shows that current LLMs not only fail at accurately predicting human responses, but that they do so in ways that are unpredictable, as they favor different social groups from one question to another. Moreover, their answers exhibit a substantially lower variance between subpopulations than what is found in real-world human data. We call this theory the "machine bias" hypothesis, according to which LLM errors do not stem primarily from imbalanced training data.

The article proceeds as follows. We first review the existing empirical work that uses LLMs to replicate human subjects, both in the human and social sciences and in computer science. We especially show that some

empirical elements in favor of the machine bias hypothesis have been found in previous studies, but that they were not fully theorized. Second, we describe the experiment we designed to adjudicate between the three hypotheses, based on a "silicon sampling" task, asking several survey questions to a language model. Third, we report our results. We provide evidence against the representative and social bias hypotheses. We then demonstrate that LLMs have a narrow set of possible answer distributions, centered around an arbitrary tendency that varies across models and questions, and that is not consistently correlated with specific social groups. These results are robust to model choice (GPT, Llama, Mixtral), data-generation strategies, prompting strategies, measurement strategies, countries, and periods. We finally discuss the methodological and theoretical implications of our findings.

Literature Review

Can LLMs Imitate (at Least Some) Humans?

Over the past two years, studies into the potential use of GPT and other LLMs for scientific purposes have multiplied. In the social sciences, one heated debate has focused on their ability to mimic human subjects. The promise is alluring. If properly prompted, generative LLMs could help survey hard-to-reach populations, run pilot surveys prior to large-scale release, reduce the cost of experiments, or help impute missing data. Across disciplines, researchers have demonstrated the value of such an approach, often in stimulating ways. In psychology, Dillion et al. (2023) presented ChatGPT with moral scenarios and asked to rate them on a scale from "very immoral" to "very moral." The comparison of its responses to those of human interviewees in four surveys yielded a very high correlation coefficient. In economics, Horton (2023) put GPT 3 to the test in four rationality experiments and came to the conclusion that "homo silicus" bears strong similarities to "homo œconomicus." In a cross-disciplinary comparison, Aher, Arriaga, and Kalai (2023) subjected GPT to a series of experiments drawn from linguistics, behavioral economics, and social psychology to conclude, once more, that generative AI is capable of emulating human behavior (see also Ashokkumar et al. 2024 who replicated 70 survey experiments). This is also the conclusion reached by Kozlowski, Kwon, and Evans (2024), who write that LLMs could create "digital doubles," opening up a path for an "in silico sociology." This list is by no means exhaustive.

One study in particular has captured the imagination of social scientists. Argyle et al. (2023) prompted GPT to predict political opinions and voting

behavior of various American sociodemographic groups. The results show a strong correlation between the predictions of the models and observed data. Based on this experiment, the authors proposed that generative LLMs can mimic the behavior of specific social groups. They called this capacity of LLMs "algorithmic fidelity," and the simulated data "silicon samples": "*Our studies show that the same language model, when properly conditioned, is able to produce outputs biased both toward and against specific groups and perspectives in ways that strongly correspond with human response patterns along fine-grained demographic axes*" (Argyle et al. 2023: 338).

Many objected. In an article directly conversing with Argyle et al. (2023), Bisbee et al. (2024) demonstrated the existence of an important divergence between LLMs' and human responses when asked to express their feelings toward certain groups. They also show that the quality of feeling thermometer prediction drops when certain variables (such as party identification and political ideology) are not provided to the model. In one large-scale experiment, computer scientists found that a mismatch between the answers of various LLMs and U.S. respondents was the rule more than the exception, and tended to favor "lower income, moderate, and Protestant or Roman Catholic groups" (Santurkar et al. 2023). Similarly, Heyde, Haensch, and Wenz (2023) concluded from their study of LLM's answers to political questions in Germany that the model they tested leaned toward the views of certain parties.

These results are in line with a number of articles that have shown, outside of silicon sampling tasks, that LLMs have significant biases. By this, their authors mean that the responses of the models are skewed toward a certain country, one or a few social groups, or a given political ideology. In a large comparative study, Atari et al. (2023) found a near-perfect correlation between a WEIRDness scale for each country and the distance between the LLM prediction and the actual human response. Probing ChatGPT, Motoki, Pinho Neto, and Rodrigues (2024) found evidence of a political bias favoring liberal views (see also Hartmann, Schwenzow, and Witte 2023; Rozado 2024; Rutinowski et al. 2024). Others confirmed this insight (Cao et al. 2023), while Johnson et al. (2022) wittily noted that GPT-3 "has an American accent." The conclusions, and especially the methods used to determine the alignment of the model to a given population, vary, but all articles converge on one aspect: the models favor certain groups.

This second line of argumentation has received a lot of support. This is, in part, a question of numbers. In this area, articles showing negative results are more prevalent than research with conclusive evidence. This is also due to the fact that it echoes an abundant literature on algorithmic discrimination. Over

the past decade, many researchers in computer science have forcefully argued that AI systems favored certain social groups. In an oft-cited article, Emily Bender and her coauthors alerted computer scientists on the role of the data used to train the models. They especially pointed to the risks of LLMs suppressing diversity because of the relative absence of minority views in the texts provided to train the LLMs (Bender et al. 2021: 613). A comprehensive investigation assessing the prevalence of a variety of stereotypes revealed the ubiquitous presence of gender bias in language models (Nangia et al. 2020), a trait once again attributed to the training data. In fact, when commenting on the nonrepresentative outputs of the LLMs on various survey responses or experimental tasks, scholars often blame their faulty response on the lack of diversity, or on the disproportionate presence of the views of hegemonic groups, in the data ingested by the models.

In this rapidly expanding literature, we aim to contribute to the debates on the ability of LLMs to replace human subjects, in two ways. First, given the centrality of the discussion about algorithmic biases, we suggest a systematic approach that helps to determine their presence in the responses of LLMs when prompted to imitate the views of diverse groups, and propose an operationalization of the notion of social bias. Second, our article offers a new take in the debate between representativity versus bias of LLMs. We build on an emerging literature showing that LLMs are either brittle (Barrie, Palaiologou, and Törnberg 2024; Berglund et al. 2024; Plaza-del-Arco, Nozza, and Hovy 2024; X. Wang et al. 2024), not appropriate for closed-form answering (Röttger et al. 2024), or that they do not necessarily exhibit a consistent social bias. Other articles also exhibit results in the models that cannot be linked to any subsample of the population, although they do so mostly incidentally. We expand on this argument, which we test in a rigorous way. We call this third approach the *machine bias* hypothesis, a concept we develop and clarify in the following section along with other important concepts.

Defining and Measuring Bias

Bias can be understood as "prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" (Mehrabi et al. 2021). As such, it is certainly one of the most important objects of contemporary sociological investigation. Algorithmic bias, the bias produced by algorithms that make use of social data, has also been extensively analyzed. To cite only a few studies, Shankar et al. (2017) showed that two main image datasets used to train computer vision models lacked geographical diversity. Most of the pictures they rely on originate from the United States, and a vast

majority from North America and European countries. In a now classic article about facial recognition, Buolamwini and Gebru (2018) demonstrated that two commonly used facial detection algorithms disproportionately feature photographs of light-skinned men. As a result, these algorithms perform best on pictures of white men and systematically misrecognize women and men of color. Similar biases with respect to gender, race or education have also been documented for recommendation systems (Schnabel et al. 2016), in medical care (Chen, Szolovits, and Ghassemi 2019), or when using word embeddings (Bolukbasi et al. 2016).

A recent survey focused on LLMs proposed a taxonomy specific to these models (Gallegos et al. 2024). It lists more than half a dozen possible ways to discriminate against someone using language, ranging from erasure (omission of the idiolect of a group) to stereotyping (false, immutable association of a group to a property) to differential recognition of forms of speech. The consequences can be important, as they affect hiring practices, access to markets (such as credit, housing) or to services, and other aspects of life. In our case, the effects seem less dire. After all, if the model displays biases, the main consequence should (hopefully) be that LLMs won't be used in lieu of human research subjects. This outcome is nevertheless not certain, so it is well worth investigating these possible biases.

But what do we call bias in the context of assessing LLMs' answers to surveys? We build on these studies to adopt a restrictive but tractable definition. The term bias should not only point to a numerical deviation as in mathematical statistics, but also convey the idea of an affinity with a specific sociodemographic group, that is, signal that the LLM's answers are better aligned with those of a certain subpopulation (e.g., American highly educated males).

Indeed, for any single question, even if the simulated data is produced randomly, one subpopulation will almost certainly be better represented than the others. If the favored group varies from one question to another, we would conclude to the existence of biases in the sense of bad predictions, but not to the presence of a social bias. To determine the existence of a *social bias*, we need to find *consistent* deviations in favor of the same subpopulation, one that is repeated across questions.

This specification of bias calls in turn for a disambiguation of the term variance. Alvero et al. (2024) found a low variance in LLM-generated essays, in the sense of a lower diversity of writing styles than what is found in human texts (see also A. Wang et al. 2024). In the case of numerical answers to questionnaires, Bisbee et al. (2024) report low variance in the sense that LLMs have a more limited answer range than humans within

each subpopulation. For categorical answers, a similar definition could be related to low entropy, where the answers of each subpopulation would be mostly concentrated on a single answer level.[6]

In this article, we use yet another measure of variance: in line with our understanding of bias as an affinity towards specific subpopulations, we measure the capability of LLMs to adapt to different sociodemographic prompts. In order to avoid confusion, we refer to this form of variance as *adaptability*. Low adaptability designates a situation where the LLM always outputs more or less the same answer distribution for a given question, irrespective of the sociodemographic group it was asked to emulate. Conversely, a model displays high adaptability if it outputs different answer distributions for different sociodemographic groups.[7]

These definitions can be readily related to the different hypotheses evaluated in this article, which we summarize in Table 1.

- The *representative* hypothesis requires low bias and high adaptability for the LLM to accurately represent the true answers of all social groups.
- The *social bias* hypothesis requires only a consistent bias toward the answers of a certain sociodemographic, and is independent from the adaptability to sociodemographic controls.
- The *machine bias* hypothesis implies both an absence of systematic bias toward a certain population and a low adaptability.

Data and Methods

To test the ability of LLMs to replicate the views of human beings, we use the "silicon sampling" task implemented by several researchers (Argyle et al.

Table 1. Relationship Between Hypotheses and Our Definitions of Bias and Adaptability.

	Bias	Adaptability
Representative	Low	High
Social bias	Socially consistent	(Unrelated)
Machine bias	Socially random	Low

Note: LLM = large language models.
Reading example: Machine bias is defined by socially random biases (the LLM favors different subpopulations across questions) and low adaptability (LLM answers show limited variations across subpopulations).

2023; Santurkar et al. 2023). The task uses sociodemographic information to condition an LLM to "think" as a person with a given profile, then asks the model a survey question. Insofar as LLMs are tools aimed at predicting the next word of a text, the task takes the model's prediction to be the answer of the given social profile to the survey question. We then compare this silicon answer to a ground truth—the answers given to the same question by human survey respondents.

Human Data: The World Values Survey

We use the World Values Survey (WVS; Inglehart et al. 2022) as the ground truth. The WVS is a long-running series of nationally representative, cross-sectional surveys that have long been the backbone of cross-national research in the sociology of values, attitudes, and representations. Its design facilitates international comparisons as well as analysis of social change over time. The WVS provides more diversity of topics of inquiry, geographical and cultural variation, and historical depth than the surveys usually used to test silicon sampling. The latter tend to have a topical focus (often politics), a cultural focus (the United States), and a historical focus (the 2010s).

We restricted the WVS data set to three periods, five countries, and four questions, in order to maximize diversity while avoiding unnecessary experiments and energy consumption. Given that LLMs are partly trained on internet data, we wanted to test whether they would be able to predict contemporary opinions and attitudes more accurately than those of the past decades, from which less content is present on the web. Therefore, we use three WVS waves separated in time by about a decade, that correspond to different moments of internet history: Wave 3 (1995–1997; prebroadband internet), Wave 5 (2005–2006; early Web 2.0) and Wave 7 (2017–2018; the contemporary internet, the latest available data before COVID-19). The countries are Australia, Germany, Mexico, Russia, and the United States. The United States is the *de facto* benchmark in the literature. We chose the other countries to have a diverse range of languages and levels of cultural proximity to the United States.[8]

Finally, we use four survey items: political ideology, social trust, religious attendance, and happiness. Politics is the *de facto* benchmark, as it has been the object of most of the initial attention in the literature. In this study, we measure it with the question "In political matters, people talk of 'the left' and 'the right.' How would you place your views on a scale of 0 to 9, where 0 is left and 9 is right, generally speaking?"[9] The scale ranges from 1 ("left") to 10 ("right"). The other items were chosen because they are available in all waves, cover core areas of sociological inquiry and include

opinions as well as attitudes. Social trust (e.g., Schilke, Reimann, and Cook 2021) is measured with the question "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?," with response options being "Most people can be trusted" (A) and "Need to be very careful" (B). Religious attendance (e.g., Voas and Chaves 2016) is measured with the question "Apart from weddings and funerals, about how often do you attend religious services these days?," with response options ranging from "More than once a week" (A) to "Never, practically never" (G). Happiness (e.g., Glass, Simon, and Andersson 2016) is measured with the question "Taking all things together, how happy are you?" and response options ranging from "Very happy" (A) to "Not at all happy" (D).

To steer LLMs toward social profiles, we select a set of stratifying variables. Two of them specify the general context (survey year and country), and five specify the individual context (sex, age, educational level, marital status, employment status). Again, this choice is driven by availability and use in the sociological literature. Despite its prominence in the U.S. context, we did not include ethnicity as it is not consistently available in the rest of the world.

The stratifying variables are coded as follows: sex (male and female), age in years (subsequently binned into age groups for the data analysis: 15–24, 25–34, 35–44, 45–54, 55–64, and 65+ years), education (distinguishing between low, medium, and high), employment status (working, retired, homemaker, student, unemployed), and marital status (married, cohabiting, divorced or separated, widowed, and single/never married).

Generated Data: Surveying an LLM

Producing LLM-generated answers for comparison against human survey data involves several technical steps: choosing an LLM, constructing a suitable prompt to interrogate it, choosing a generation strategy, using the LLM outputs to compute answer distributions for sociodemographic groups, and choosing a suitable distance measure between generated and ground-truth answers.

To formulate our prompts, we use the "interview-style" prompting strategy proposed in the second study of (Argyle et al. 2023).[10] The control variables are given to the model in a question–answer format that imitates an interview situation: an interviewer asks questions and provides the available answers, and an interviewee chooses one of the answers (except for the year and age questions, where they answer by a number). The sociodemographic controls

are provided in the prompt, and the LLM is asked to fill in the interviewee's answer to the last question. This prompting format serves two purposes. First, it provides the model with the sociodemographic information of each persona it is supposed to simulate. Second, it conditions the LLM to select one of the expected answers suggested by the interviewer. On our data, we find this strategy able to make LLMs generate not only properly formatted answers, but also probability distributions that are both credible at first glance and varying over sociodemographic profiles.

Here is an abridged example prompt[11]:

> Question: What year is it?
>
> Answer: 1995
>
> Question: What country are we in? Is it "A. Australia" or "B. Germany" or "C. Mexico" or "D. Russia" or "E. United States"?
>
> Answer: A. Australia
>
> Question: How old are you?
>
> …
>
> Question: Taking all things together, how happy are you? Are you "A. Very happy" or "B. Quite happy" or "C. Not very happy" or "D. Not at all happy"?
>
> Answer: (Response by the model).

We report results for an answer-generation method called next-token probability (NTP), on three LLMs considered state of the art in their respective classes at the time of our experiments: Llama-3-70B 5bit-quantized (Llama), Mixtral-8 × 7B-0.1 4bit-quantized (Mixtral), and GPT-4-turbo (GPT, undisclosed number of parameters).[12] Results for older, smaller or unquantized models, as well as with a full answer generation technique all lead to the same conclusions (see Supplemental Materials section D).

The NTP method consists in observing the first token, that is, word or part of word, generated by each prompt. It leverages the fact that when prompted, an LLM deterministically outputs a probability distribution over the thousands of tokens in its vocabulary, indicating the most likely

candidates for the next token in its text-prediction task. While the common use of LLMs as text generators involves many successive iterations of this probability estimation, building sentences through repeated token prediction, a single iteration can be enough in the case of survey answers, as each available answer can fit into a single token (typically a single digit or uppercase letter, e.g., choice "A" or "B" or "C"). This is the approach followed in NTP: for each unique interview prompt, we obtain a probability distribution over the valid answers by normalizing the probabilities associated to the valid tokens to 1.[13] As NTP is deterministic, only unique prompts are presented to the model, and are subsequently reweighted to match the WVS sample.[14] Our choice of questions, country-year waves and controls results in a total of approximately 56,000 prompts per model.

Measuring the Quality of Generated Data

The goal of our experiment is to evaluate the ability of LLMs to accurately predict the answers to survey questions, where success means that the responses of the language model match the observed answers of WVS human respondents, based on their sociodemographic properties. The precise quality of this prediction can only be evaluated by reasoning over groups of observations, as a one-to-one observation comparison would be highly affected by the randomness inherent to surveys.[15] In order to preserve the multivariate nature of social profiles, we group the observations into subpopulations based on the combination of their sociodemographic properties, using a greedy algorithm.[16] This yields a total of 687 subpopulations, with an average size of 39 WVS respondents (range of N: 20–115).[17]

For each subpopulation, we compute an LLM answer distribution and its ground-truth counterpart obtained from the survey data. To measure the distance between these two probability distributions, we use the Earth-Mover's Distance (EMD), also known as 1 − Wasserstein Distance. This is the standard statistical approach when it comes to comparing distributions of ordinal data.[18] Consider two samples of equal size, giving distributions P and Q over the same n available answer levels. We set a uniform cost to moving one observation from one level to the previous or next level. EMD is defined as the total cost of the less costly solution that transforms P into Q by moving observations along the levels. It is then divided by the theoretical maximum $(n - 1)$, so that the resulting normalized measure ranges between 0 and 1. To compute EMD and its normalization nEMD, we consider the

probability values P_i and Q_i for n discrete levels i, and use the following formula:

$$d_0 = 0$$
$$d_{i+1} = P_i + d_i - Q_i$$
$$\text{EMD}(P, Q) = \sum_{i=0}^{n} |d_i|$$
$$\text{nEMD}(P, Q) = \frac{1}{n-1}\text{EMD}(P, Q)$$

The normalized Earth-Mover's Distance ranges from 0 to 1, where 0 is perfect agreement, and 1 is complete disagreement. A distance of 1 on political ideology, for instance, would correspond to the situation where everyone in a given subpopulation is far left but the LLM predicts everyone to be far right.[19] A normalized Earth-Mover's Distance of .1 on a political ideology scale with an odd number of possible answers means that 10 percent of observations need to be moved from far left to far right for the two distributions to become identical (or, similarly, 10 percent from center to far left and 10 percent from center to far right). Figure 1 displays an illustration on a four-level Likert scale.

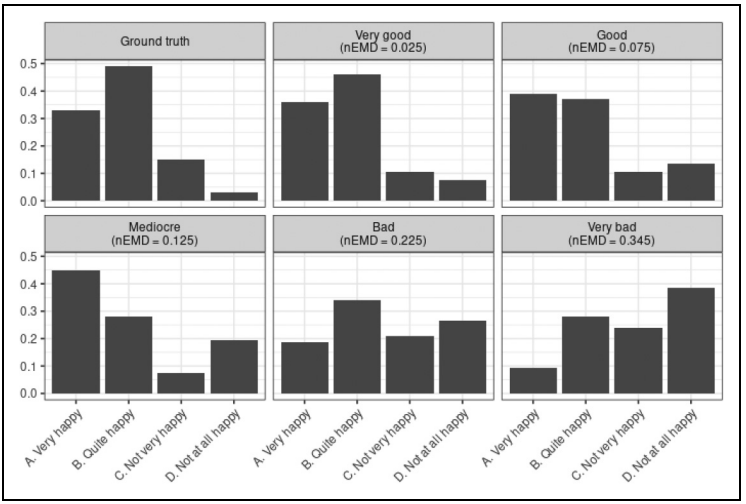

Figure 1. Examples of normalized Earth-Mover's Distances between the ground truth and other answer distributions, illustrating the different quality categories.

The distance metric cannot be interpreted in a linear fashion, as there is greater variation in bad prediction scores than in good ones. We therefore suggest the following arbitrary labels for distance thresholds: very good (< .05), good (< .1), mediocre (< .15), bad (< .3), very bad (> .3). In Figure 1, we clearly see that even relatively low values such as .075 already signal a significant gap between the two distributions. We deliberately chose generous thresholds to ensure a conservative test of our hypotheses. In other words, LLMs could have passed our test even with underwhelming predictions.

Finally, we set up two baselines for comparison with LLM performances. The first is a "random" baseline, which we obtain by performing random permutations of the individual answers in the ground-truth survey data.[20] This procedure yields a perfect global average prediction, but eliminates any link that may exist between sociodemographic properties and outcome answers. The second baseline is a linear regression model that, in standard social scientific fashion, predicts answers from sociodemographic variables based on the survey data. We use a multinomial logit to regress WVS answer distributions on the sociodemographic variables of the 687 subpopulations, and report cross-validated results for fair comparison.[21]

These baseline models serve several purposes. While they could be said to hold an unfair advantage, being based on the actual survey data, they do provide legitimate comparison points for evaluating the central promise of LLMs as substitutes for human subjects, which is precisely to remove the need for human data. The random baseline acts as a strict lower bound on quality: while absolute nEMD values are hard to interpret, any predictive model that performs worse than random permutations should be regarded as a poor representation of the social phenomena at hand. The linear model gives a more attainable lower bound: with around 100 parameters instead of the billions that make up LLMs, it is the first approximation that a social scientist would use to evaluate the link between sociodemographics and outcomes such as happiness or religious practice. It is a lower bound because more intricate models, involving interactions or nonlinear transformations, would likely yield better predictions. Additionally, the linear baseline provides an indication of how well the outcome variables are correlated with the given sociodemographic predictors, even in the presence of omitted variables. Good predictions of a simple data-based model would indicate that the task is feasible, that is, that the regularities in the variables at hand are sufficiently strong for it to be reasonable to ask LLMs to emulate them.

Table 2. nEMD of Average LLM Predictions with Respect to Average WVS Answer Distributions, by Model and Question.

	Happiness	Politics	Religion	Trust
GPT-4T	0.143	0.152	0.234	0.332
Llama	0.022	0.041	0.031	0.183
Mixtral	0.031	0.027	0.184	0.101

Note: nEMD = normalized Earth-Mover's Distances; LLM = large language models; WVS = World Values Survey.
Reading example: The overall average prediction of GPT4-turbo on the Happiness question has an nEMD distance of 0.143 to the overall average of ground-truth answer distributions.

Results

LLMs' Predictions Are Far from the Ground Truth

Can LLMs generate answers close to what a diverse sample of actual humans would? We first evaluate the accuracy of LLMs with descriptive statistics on the distance between their predictions and the ground truth. To measure this distance we use the normalized EMD (nEMD).

In some cases, the average responses of LLMs come quite close to the average answer distributions observed in the WVS data. Table 2 shows that the nEMD of average Mixtral and Llama responses fall in the "very good" range for the Happiness and Politics question, with distances lower than .05. GPT, on the other hand, yields average predictions that are much more distant from average WVS answers, with nEMD ranging from .143 to .332.

Comparing answer distributions at the fine-grained subpopulation level gives a more detailed picture. Considering all 687 subpopulations and four outcome variables together, the distance between the predictions of the three LLMs and the survey responses is usually high. Figure 2 displays the density of the nEMD, and Table 3 summarizes their distribution over quality classes. Only a small part of LLM predictions fall in the "very good" range (a distance lower than .05), and a large proportion of subpopulations fall in the "bad" to "very bad" range (.15 and more).

We also notice variation across models (Figure 2 and Table 3). On all questions, GPT-4 is the worst-performing, with 90 percent of predictions at best mediocre (nEMD $\geq$ .1). Llama fares better with 52.3 percent; and Mixtral, the best model in our experiment, still gets half of its

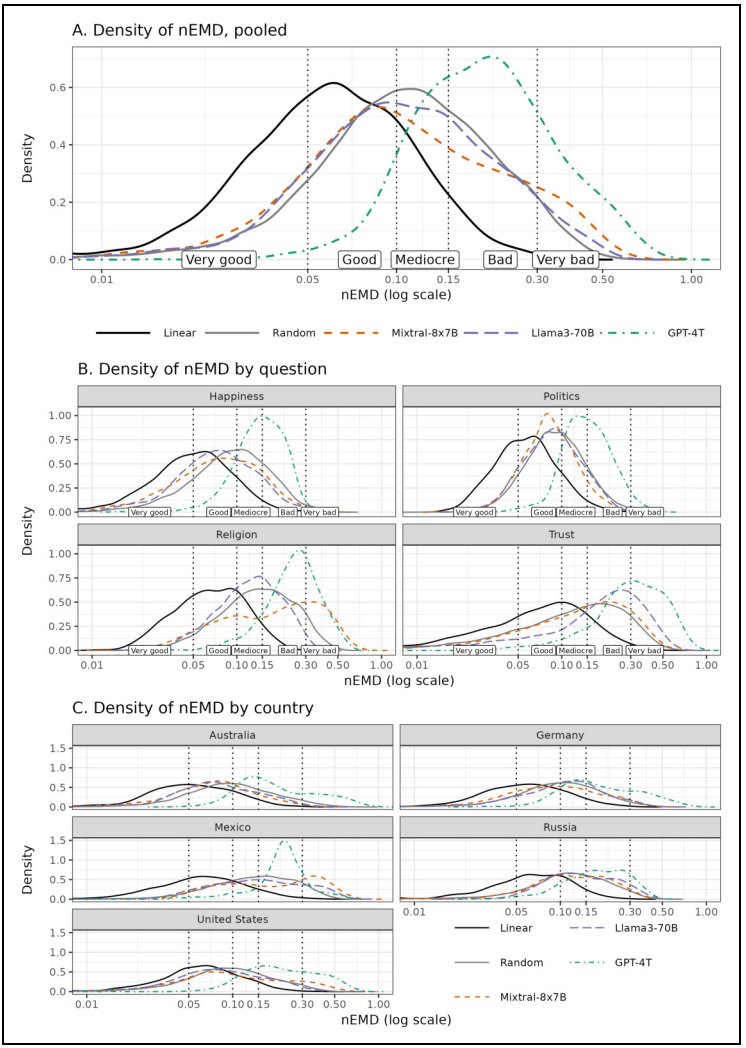

Figure 2. The overall quality of LLM predictions is low when compared to actual survey responses, and LLMs are worse predictors than a linear model. Panel A shows the density of the nEMD across subpopulations when all questions are pooled together and densities are stratified by method: most LLM predictions are "mediocre" or worse, irrespective of the LLM used. Panel B shows the same density, stratified by LLM and question. Panel C shows the same density, now stratified by LLM and country.
Note: nEMD = normalized Earth-Mover's Distances; LLM = large language models.

Table 3. Quality of LLM Predictions, Compared to the Linear Baseline and Random Permutations, Over all Subpopulations and Outcome Variables (in percent).

	Very good	Good	Mediocre	Bad	Very bad	Total
nEMD	(0–.05)	(.05–.10)	(.10–.15)	(.15–.30)	(.30–1)	
Baseline						
Linear (CV)	39.4	40.2	14.1	6.1	0.2	100.0
Random	13.7	31.8	23.3	26.2	5.1	100.0
LLM						
Mixtral	16.1	33.7	18.3	21.7	10.2	100.0
Llama	14.9	32.8	21.5	24.6	6.2	100.0
GPT-4T	1.1	8.9	22.3	45.5	22.3	100.0

Note: nEMD = normalized Earth-Mover's Distances; LLM = large language model; CV: cross-validated.

Reading example: Mixtral-8 × 7B achieves "very good" quality (nEMD lower than 0.05) on 16.1 percent of the subpopulations, compared to 13.7 percent for random permutations of the ground truth, and 39.4 percent for a cross-validated linear model.

predictions in the mediocre range or worse (50.2 percent). Notably, all LLMs fare significantly worse than the linear model (79.6 percent good or very good), and the best models perform only marginally better than random permutations of the ground-truth individual answers (45.5 percent good or very good), while the worst model, GPT, does not even reach the low standard of this random baseline.[22] This result challenges the representative hypothesis championed by part of the literature: at least for our questions, we find that these LLMs are untrustworthy with regard to predicting people's survey responses.

Are these results driven by some topics that are more difficult to predict than others? The overall picture of poor prediction holds when disaggregating the subpopulations along the four questions, as shown in Panel B of Figure 2. Political ideology does appear to have better predictions than trust and religious attendance, and so does happiness. However, even these questions have more than half of subpopulations in the "mediocre" category, or worse.

The pattern of poor prediction quality also holds when we group the subpopulations along the five country samples, as shown in Panel C of Figure 2. Few country differences arise and, contrary to our initial expectations based on the social bias literature, predictions are not always better for the United States.

LLMs Have No Systematic Social Bias

We then turn to the social bias hypothesis, and explore whether the error in prediction can be attributed to the LLMs being more representative of some social groups than others. To do so, we use a linear model that we call the *Social* model, regressing the LLM's prediction error (measured by nEMD distance) on the sociodemographic control variables.

$$\text{(Social model) } d_i = \alpha + \beta \times S_i + u_i$$
$$d_i = \log(1 + nEMD(WVS_i, LLM_i))$$
$$S_i \text{ is the vector of socio-demographic controls}$$
$$WVS_i,\ LLM_i \text{ are answer distributions}$$

Two predictions follow from the social bias hypothesis:

1. Sociodemographic characteristics are good predictors of the distance between LLM answers and the ground truth. This can be tested with goodness-of-fit indicators of the social model. The social bias hypothesis should lead to high R^2 values with this model.[23]
2. Sociodemographic characteristics are consistent predictors. This can be tested by comparing the signs of coefficients across questions. The social bias hypothesis should lead to significant coefficients of the same sign across all four questions for at least some predictors.

We fit one model for each LLM–question pair. We first consider the goodness-of-fit. Table 5 displays indicators for the three main LLMs. In a few models, the error in prediction is well predicted by sociodemographic variables. This is for instance the case of religious attendance by Mixtral ($R^2 \simeq .74$), or trust and politics by GPT ($R^2 \simeq .57$). On the eight other models we evaluated, the R^2 is quite low, with a median of .36. We conclude from these indicators that sociodemographics have a weak to moderate correlation with the LLMs' prediction error, and that social bias fails to predict the major part of this error.

We then look at coefficient stability across questions. Figure 3 displays the coefficients of all 24 dummy social predictors in the *Social* model, for Mixtral, for the four questions. It clearly shows that the association between sociodemographic variables and the distance to the ground truth is not consistent across questions. For the Mixtral model, the only predictor showing consistent and significantly non-null values is "Mexico," with a positive sign indicating a worse performance for Mexican data compared to

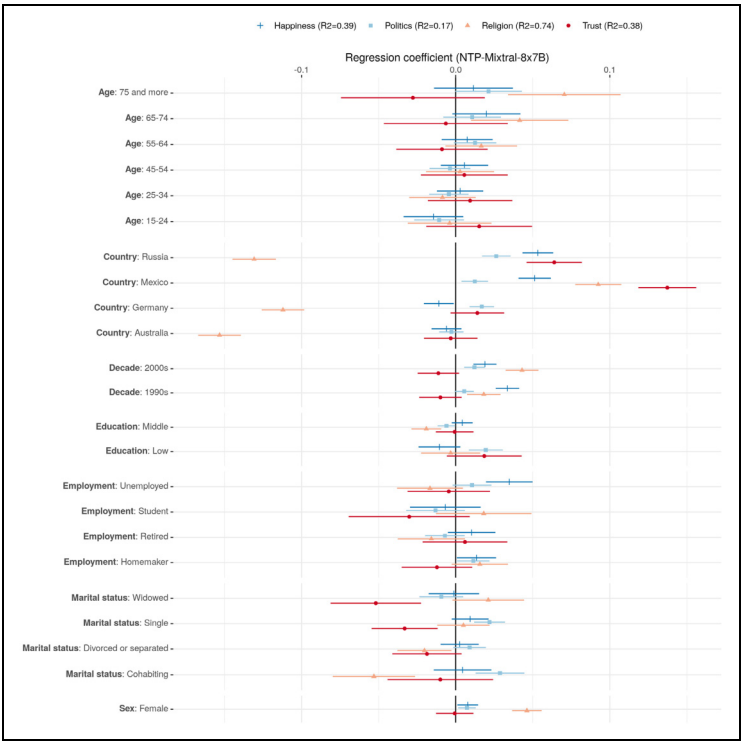

Figure 3. LLM performance exhibits hardly any bias for specific social groups. *Social* model for Mixtral-8 × 7B, estimated coefficients and 95 percent confidence intervals. Reference: 2010s, United States, age 35–44, high education, working, married male. *Reading: a coefficient is significant at the 5* percent *threshold if its confidence interval doesn't include the 0 vertical.*
Note: LLM = large language models.

the U.S. data. "Female" also has positive effects across questions—the language model predicts women's opinions less accurately than men's opinions—but one of them is not significant at the 5 percent level. The effects of other predictors are either not significant or do not have a consistent sign, even with only four questions. These results hold for all LLMs we tested, as shown in the Supplemental Figures S5 and S6.

In other words, we do not find any evidence of consistent bias in support of the social bias hypothesis. One practical implication is that we are unable to predict how answers to a new question might be

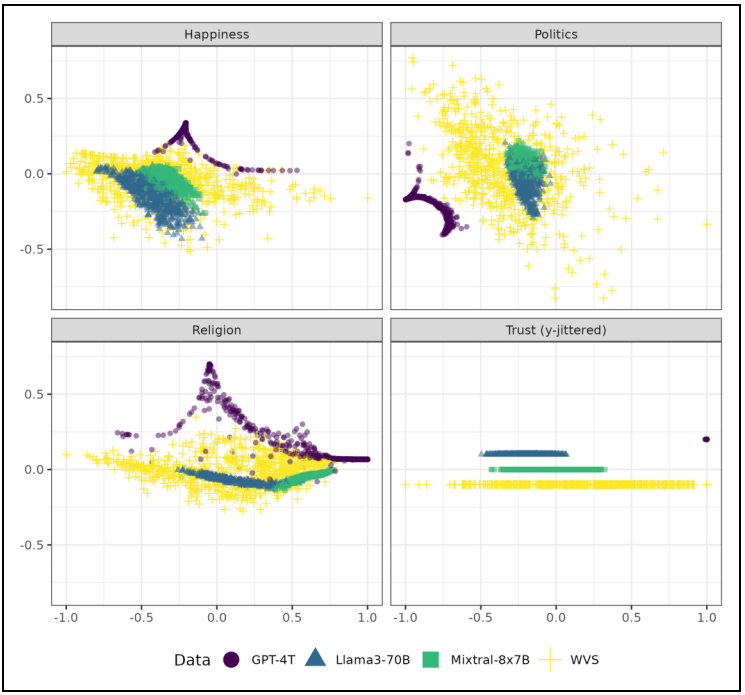

Figure 4. The space of answer distributions generated by LLMs (dark data points) is much narrower than the survey answers of human respondents (light data points). Multidimensional scaling on the pairwise normalized Earth-Movers' Distances.
Note: LLM = large language models.

biased, that is, whether married or single, high or low education, young or old people would be more accurately described if we generated LLM answers to a new question.

LLMs Have Low Adaptability

We then investigate the extent to which LLM answer patterns change with sociodemographic prompts, i.e., the social adaptability of each LLM. We first visualize the dispersion of WVS and LLM answer distributions with multidimensional scaling (MDS). MDS is a visualization

Table 4. Median Pairwise nEMD of Subpopulation Answer Distributions.

	Happiness	Politics	Religion	Trust
WVS	0.115	0.108	0.174	0.156
GPT-4T	0.006	0.021	0.131	0.000
Llama-3-70B	0.052	0.026	0.069	0.043
Mixtral-8 × 7B	0.033	0.019	0.032	0.063

Note: nEMD = normalized Earth-Mover's Distances; LLM = large language models; WVS = World Values Survey.
Reading example: the median nEMD between all pairs of subpopulations of WVS on Happiness is 0.115, while only 0.006 for GPT, indicating a lower dispersion of answer distributions generated by the LLM.

technique used for scaling observations defined by a distance matrix. For each question, we compute a single MDS on the pairwise nEMD distances on all subpopulations (pooled LLMs and ground truth).

In Figure 4, the *x*- and *y*-axes represent the coordinates of the MDS, and each point represents the answer pattern of a subpopulation, either WVS or LLM. In this space, the closer two points are, the lower their distance is, as measured by nEMD. The color distinguishes between "ground truth" WVS (light) and LLMs (darker shades). The visualization indicates that LLM answer patterns are much more concentrated than human ones: they occupy a limited subspace of answer space, especially compared to the high dispersion of human data.[24] This result of low LLM adaptability to sociodemographic prompts is consistently found across questions, models and generation strategies (see Supplemental Materials section C.2).

Table 4 offers further evidence of low LLM adaptability. It displays a dispersion measure of answer distributions over subpopulations, median pairwise nEMD. The lower values observed for LLMs indicate that their answer patterns are more concentrated than WVS ones, that is, less diverse across sociodemographic controls.

LLM Errors Are Driven by Machine Bias

We have established that language models predict answers that are far from the ground truth, that their responses are not consistently driven toward any given social group, and that their answer patterns do not vary much over sociodemographics. We propose to call this phenomenon "machine bias": each LLM gives an idiosyncratic answer pattern

to each question, limited to a narrow portion of the human answer space and socially inconsistent across questions.

In terms of LLM–WVS prediction error, this "machine bias" implies that a large portion of error variance can be explained by the distances between the (diverse) WVS answers and the central tendency of their (concentrated) LLM counterparts. In order to test this, we devise two new models, extending the first *Social* model described above. As a reminder, the *Social* model describes the error in LLM predictions as a function of socio-demographic properties of subpopulations. Succinctly, the new *Center* model replaces this with one simple predictor, the distance between the ground-truth answer patterns and the average pattern produced by the LLM. In other words, it tests the hypothesis that each LLM outputs a narrow set of answers. Finally, the *Full* model includes both sets of predictors.

$$
\begin{aligned}
\text{(Social) } d_i &= \alpha + \beta \times S_i + u_i \\
\text{(Center) } d_i &= \gamma + \delta \times c_i + v_i \\
\text{(Full) } d_i &= \kappa + \lambda \times S_i + \mu \times c_i + w_i \\
&\text{where} \\
d_i &= \log(1 + nEMD(WVS_i, LLM_i)) \\
c_i &= \log(1 + centerEMD_i) \\
&= \log(1 + nEMD(WVS_i, \overline{LLM}))
\end{aligned}
$$

S_i is the vector of sociodemographic controls; WVS_i and LLM_i are answer distributions; $\overline{LLM}$ is the average LLM answer distribution on a held-out sample; and u_i, v_i, and w_i are error terms.

In more detail, the *Center* and *Full* models evaluate the impact of low LLM adaptability on the LLMs' prediction error. Low adaptability, that is, the tendency of LLM answers to be similar across sociodemographics, implies a concentration around a central answer distribution. We operationalize this by computing, for each LLM–question pair, the $\overline{LLM}$ average answer distribution. This average is computed on a random subset of 20 subpopulations, a subset subsequently held out of the regression models so as to lift suspicions of endogeneity.

The distance between WVS answers and the $\overline{LLM}$ central tendency, that we call *centerEMD*, functions as a counterfactual: it represents the value that the LLM's prediction error would have if the LLM was completely unresponsive to variations in sociodemographics. In such a case, LLM prediction errors would be driven by the sociodemographic variations of the ground-

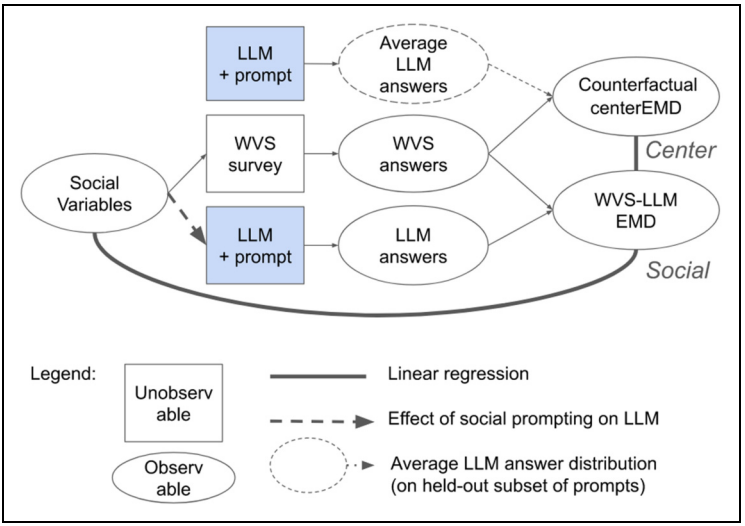

Figure 5. Data-generating process and analytical strategy.

truth WVS answer patterns only. Figure 5 illustrates this counterfactual reasoning, aimed at isolating the effect of social controls on the LLM answers.

The *Center* model regresses LLM prediction errors only on the counterfactual *centerEMD*, the *Full* model both on the counterfactual and on the social controls. The former measures how much of LLM prediction error can be attributed to low LLM dispersion alone, the latter the relative contributions of low dispersion and social controls. Similarly to the *Social* model, we estimate these two additional models on the four survey questions separately, applying a logarithmic transformation to the nEMD distances to better match their observed approximate log-normal distribution.

We first look at the goodness of fit of the three models. Table 5 presents two measures for each of these linear models, adjusted R^2 and the Bayesian Information Criterion (BIC). Under the machine bias hypothesis, we expect the R^2 of the *Center* model to be higher than that of the *Social* model, and to be generally very high, and the BIC to be symmetrically lower. Both measures account for the difference in model sizes (24 predictors for the *Social* model, 1 for the *Center* model, 25 for the *Full* model); the BIC is known to give a stronger penalty to models with more predictors.

Center models show markedly better predictive power than *Social* models in most cases (11 out of 12), and they often do so by a large margin. Comparing

Table 5. Goodness-of-fit Measures for the Three Linear Models of LLM–WVS Distance.

		Adjusted R^2			BIC		
Question	LLM	*Social*	*Center*	*Full*	*Social*	*Center*	*Full*
Happiness							
	GPT-4T	0.441	**0.828**	0.870	−2399	**−3306**	−3367
	Llama-3-70B	0.346	**0.667**	0.728	−2308	**−2878**	−2886
	Mixtral-8×7B	0.387	**0.829**	0.858	−2276	**−3250**	−3245
Politics							
	GPT-4T	0.560	**0.925**	0.949	−2257	**−3480**	−3588
	Llama-3-70B	0.260	**0.852**	0.909	−2313	**−3433**	−3610
	Mixtral-8×7B	0.172	**0.930**	0.948	−2371	**−4020**	−4084
Religion							
	GPT-4T	**0.364**	0.232	0.685	**−1621**	−1616	−2084
	Llama-3-70B	0.198	**0.531**	0.548	−1974	**−2452**	−2350
	Mixtral-8×7B	0.744	**0.951**	0.982	−1808	**−3026**	−3571
Trust							
	GPT-4T	0.581	**1.000**	1.000	−1360	**−9264**	−9326
	Llama-3-70B	0.444	**0.895**	0.946	−1377	**−2612**	−2925
	Mixtral-8×7B	0.382	**0.695**	0.774	−1494	**−2085**	−2157

Note: BIC: Bayesian Information Criterion; nEMD = normalized Earth-Mover's Distances; LLM = large language models; WVS = World Values Survey.
Reading example: For Happiness with GPT-4-Turbo, the linear model of all the social conditioning variables accounts for 44 percent of the variance in LLM–WVS nEMD distances. The model whose single predictor is the distance of WVS to LLM central tendency accounts for 83 percent of variance, and the full model for 87 percent (after adjustment for the number of regressors). The BIC for these models are −2399, −3306, and −3367 respectively: according to this criterion, the center-only model is better than the social-only model, and the full model is slightly better than the center-only one.
Bold font denotes best-fitting model between Social and Center.

the *Center* and *Full* models gives an indication of how much information the sociodemographic predictors contribute to the models. Comparisons favor the *Full* model by at least a small margin in all cases according to adjusted R^2, and all but two cases according to BIC, indicating that social prompting does play a role, albeit a small one, in the production of LLM error, even when controlling for machine bias. The *Full* model nonetheless performs better than the *Social* model in all cases, and does so by a large margin. These results provide significant support to what we call the machine bias hypothesis, where the low

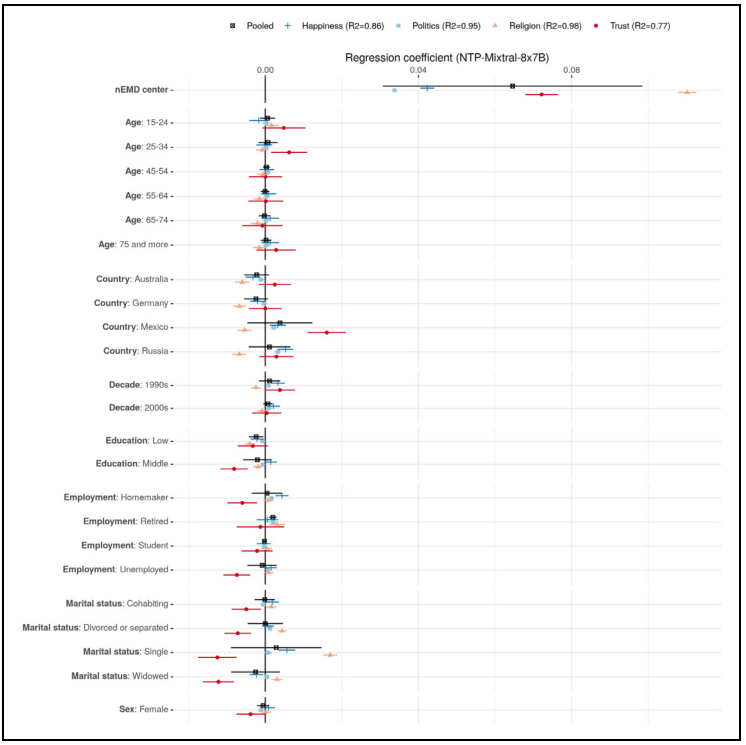

Figure 6. The effect of machine bias ($nEMD_{center}$) is substantially larger than all sociodemographic indicators: estimated coefficients of the *full* model for Mixtral, with normalized regressors functioning as variable importance measures. Error bars denote 95 percent confidence intervals, $N = 667$ subpopulations (619 for politics). Pooled coefficient based on random effects meta-analyses in black, individual question coefficients in gray.
Note: nEMD = normalized Earth-Mover's Distances.

diversity of LLM answer patterns is the main driver behind LLM error patterns, only marginally affected by socio-demographic variables.

Another indication of the predominance of this machine bias is found in the magnitudes of the normalized parameters of the *Full* model, as can be seen in Figure 6 for Mixtral, the coefficient for distance to the "center" is at least an order of magnitude greater than the social factor coefficients. This means that, for each model and question, the best predictor of LLM quality is the quality of a counterfactual socially agnostic LLM. In other

Table 6. F-Test of Nested Models, Social or Center Versus Full Model.

		Center versus *Full* (i.e., can we remove *Social*?)		*Social* versus *Full* (i.e., can we remove *Center*?)	
Question	LLM	*F*-statistic *df* (642, 23)	*p*-value	*F*-statistic *df* (642, 1)	*p*-value
Happiness					
	GPT-4T	10.33	3.20×10^{-31}	2122.43	1.04×10^{-205}
	Llama-3-70B	7.44	2.15×10^{-21}	900.97	2.35×10^{-124}
	Mixtral-8×7B	6.73	6.22×10^{-19}	2126.66	6.39×10^{-206}
Politics					
	GPT-4T	13.19	6.32×10^{-40}	4526.83	7.16×10^{-281}
	Llama-3-70B	17.77	2.12×10^{-53}	4261.88	1.01×10^{-273}
	Mixtral-8×7B	10.53	1.70×10^{-31}	8904.94	0.00
Religion					
	GPT-4T	42.54	9.02×10^{-113}	656.53	2.80×10^{-100}
	Llama-3-70B	2.08	2.27×10^{-3}	498.21	4.01×10^{-82}
	Mixtral-8×7B	51.07	2.18×10^{-128}	8471.41	0.00
Trust					
	GPT-4T	10.44	1.39×10^{-31}	99691485.00	0.00
	Llama-3-70B	27.94	3.01×10^{-81}	5964.56	0.00
	Mixtral-8×7B	11.00	1.93×10^{-33}	1111.75	3.17×10^{-142}

words, once the central tendency of the LLM is known, one barely needs to feed the LLM new social prompts in order to measure its prediction quality: computing the distance between the ground truth and the LLM's known central tendency can be a reliable proxy. Supplemental Figures S6 and S8 display similar results for GPT and Llama.

Finally, we leverage the nestedness of our models for a formal test of nullity of either all social parameters, or the counterfactual *centerEMD* parameter. This amounts to testing the *Full* model against the *Center* and *Social* models respectively, with a null hypothesis that the simpler model is a better fit. *F*-test results are presented in Table 6, and show that both social factors and distances to LLM averages contribute significantly to the *Full* model, albeit with *F* scores often two orders of magnitude higher for the exclusion of LLM central tendency. In other words, it can be costly to drop the social variables from the equation, but it is much more costly to drop the *Center* term.

Limitations

Our results are dependent on a series of choices. The first, most obvious one is that our prompting strategy might not have worked properly. Put otherwise, there may exist other, more efficient ways to steer language models (Lin 2024), in this case toward various social groups. While we cannot prove that an alternative prompting strategy would not lead to results confirming either the representative or the social bias hypotheses, we adopted the same strategy as some influential articles (Argyle et al. 2023; Santurkar et al. 2023). We also tried several variations around this prompting strategy, with similar conclusions (see Supplementary Materials section D2 for details).

Another potential cause for concern could be that the models we used may not be the best, or at least not the best suited for the task performed. It is true that in a rapidly changing landscape, it is often difficult to remain up to date. In fact, in the course of this research, we reran our experiments frequently, either due to new versions of models being released, or due to an entirely new model emerging as the new "state of the art" (see Supplementary Materials section B1). We only report the most recent version tested for each model. Still, we believe that our strategy yields solid results to date. For instance, we tested GPT-3 because it was the main benchmark mobilized by other scientists using LLMs to replicate opinions, but we also investigated GPT-3.5-Instruct and GPT-4-Turbo, only to get similar results.

We also tested powerful open-weight models such as Llama and Mixtral. The answer patterns changed from model to model, even after a simple

update. This is in itself a problem; the lack of transparency of proprietary models, and of open-weights models whose training code or data cannot be fully accessed limit the reproducibility and interpretability of research on language models (Ollion et al. 2024; Spirling 2023). However, none of these changes affected our main results: the overall pattern of a high distance to the ground truth, a low adaptability to and a weak correlation with sociodemographics, always remained the same. While we cannot rule out that future models may be able to mimic real populations, we feel confident writing that the current generation of LLMs cannot.

One might argue that the number of survey questions (four) and social contexts (five demographic variables, five countries, three time periods) we considered is low. This is certainly true, but our choice was theoretically grounded. We selected questions and control variables that are all well documented in sociological research. The seven control variables are strong predictors of the outcome variables, as evidenced by the high accuracy of the linear baseline model, ensuring that they should be sufficient for guiding LLM predictions. The outcome questions are diverse, in contrast with studies focused on politics, and we did not expect the inconsistency in biases. It is not impossible that, given more outcomes, a consistent social bias could be found, and we think that this question could be more widely tested using the framework proposed in this article. Finally, we did not add or remove any question, predictor, country, or period during the research, and tested multiple models and generation techniques, all leading to the same conclusions (see Supplementary Materials section D).

Discussion

The puzzle at the center of current debates on LLMs is whether recent advances provide them with sufficient accuracy to replicate human behavior and attitudes. The focus of this article is on the particular case of using LLMs to simulate human responses to surveys. We proposed a formal test of the two main hypotheses which sit at opposite ends of the debate.

Our results confirm neither of the two arguments. Against the "representative" hypothesis, we find that opinions and attitudes of actual social groups are on the whole poorly predicted. Over half of all estimates are at a substantial distance from the ground truth, and no LLM performs significantly better than random guesses. This adds to a growing body of evidence that casts doubt on the idea that the current models could be prompted to mimic certain social groups, such as citizens of a country or

more fine-grained subpopulations. Against the "social bias" hypothesis, we find that socio-demographics are not only weak predictors of LLM errors, but also inconsistent across questions.

Can LLMs Replace Research Subjects?

The first implication of our study concerns the use of silicon samples in science. The appeal is obvious. Should the representative hypothesis hold, they would considerably reduce the costs of conducting research. The results presented in this article, however, should warn us against an overly enthusiastic or rapid use of these technologies. To date, LLMs cannot be used to replace research subjects, at least not without further specifications nor without taking precautions in the use of the results.

This does not mean that LLMs are either useless for social science on other tasks (Bail 2024), or altogether incapable of capturing social regularities. Strimling, Krueger, and Karlsson (2024) show for instance that GPT-4 is able to output the dominant view on certain moral topics with a reasonable success rate, at least at the country level. Sometimes, the demands placed on the model are more limited: when asked to imitate what is considered a universal reaction to a signal, LLM simulations may be informative.[25]

Passing additional information to the model may also help them attain better performances. For instance, Kim and Lee (2023) fine-tuned an LLM with survey responses to increase the quality of their output. They showed that the models were able to correctly impute data or retrodict information that was missing for various years of the General Social Survey. This is a far cry from producing accurate answers for a set of diverse populations on questions that were never asked before, but it may be one path forward for fruitful uses of these models.

Beyond Social Bias, Investigating Machine Bias

Our results call for a clarification, and they open a new path of investigation. The clarification concerns the concept of bias. At the moment, the term is widely used, a sign of the enthusiasm mixed with fear that surrounds the use of generative AI. It has also become particularly vague, as a number of researchers have already pointed out (Navigli, Conia, and Ross 2023). One contribution of this study is to advocate for a precise definition of "social bias" in LLMs' responses. We contend that in order to establish the existence of such a bias, we need to demonstrate that the

model consistently favors the position of a specific social group. In other words, we need to clearly distinguish between bias (a simple deviation from the truth) and social bias.

The fact that we find no evidence of consistent social bias in our experiments calls for further investigations into the sources of LLM biases. They are generally attributed to the unbalanced training data. While this argument is intuitively convincing, and is supported by considerable evidence outside of the silicon sampling literature, it cannot account for our results. We know little about the composition of the training corpora, except that English is largely dominant. Yet our models sometimes replicate the views of non-English-speaking, and therefore underrepresented, countries.[26] More generally, we find that while all LLM responses are biased, they also point towards different subpopulations with each model and question. Beyond the training data, this machine bias may have other sources. Some have already been investigated in computational linguistics, such as the choice of architecture or the evaluation tasks on which the models were optimized (Hovy and Prabhumoye 2021). Others may yet come to light.

What can LLMs Know?

The idea of using language models to replace human respondents relies on an intuition, namely that they are able to successfully encode, and retrieve, social regularities. Having been trained on massive text corpora (mostly) produced by humans, LLMs would implicitly learn to simulate human thought and social patterns, insofar as they are expressed in the corpora. This intuition is motivated by the existence of unexpected but potent "emergent capabilities" that LLMs started to display as their models and training corpora grew larger (Schaeffer, Miranda, and Koyejo 2023; Zoph et al. 2022). Just a few years ago, not many would have believed word-prediction algorithms to be capable of carrying out such fundamentally human tasks as constructing sensible arguments, programming software, or making decisions and justifying them.

These capabilities are undeniable, and one would be hard-pressed to consider generative language models as a passing fad. Yet, despite their impressive achievements, they are still notoriously unreliable in some respects. Most prominently, they are prone to producing "hallucinated" information, outputting answers that are factually wrong, often with unsettling apparent self-confidence. This stems from the fact that information is stored in LLMs not in the form of a hard-coded database, but rather as a byproduct of probabilistic word prediction. "Knowledge," for LLMs, is

distributed throughout the billions of weights that produce their next-word probabilities, and that encode patterns and associations found in the data they were fed.[27]

Interrogating LLMs as if they were humans in a survey, what the literature calls "silicon sampling," is even more ambitious than asking them for factual knowledge. It is obviously easier for a model to reproduce for example, historical dates and names that frequently occur together in a large corpus than it is to simulate survey answers that are consistent with scientific knowledge about social regularities. Firstly, neither explicit social-scientific formulations of such regularities nor raw interview data are likely to occur frequently, if at all, in the (largely undisclosed) LLM training corpora. Secondly, even if some regularities had been successfully embedded in a model's weights, simply interrogating it with a context–question pair might not be the most efficient way to retrieve this information, if only because different reformulations of the same question are known to produce different answers (Berglund et al. 2024). Lastly, the very need for social science research stems from the fact that its discoveries are often contrary to common sense. To be sure, some frequent positive associations (e.g., "Republican" and "Trump") can help probabilistic models retrodict past electoral behavior to some extent. Beyond such simple tasks, however, expecting pretrained LLMs to successfully emulate complex human behaviors or personas is likely to be a risky bet on their "emergent capabilities."[28]

Conclusion

Alchemists of the Middle Ages believed that, in a nearby future, they would be able to conjure miniature humans out of inert matter. These artificial beings, which they called "homunculi," were seen as having a number of special qualities, such as exceptional mastery of the arts. Produced *in silico*, they would also be easily replicated. All that was needed to create them were a few mundane ingredients and, of course, craftsmanship.

Is the idea that LLMs can create miniature humans the modern-day version of the belief in homunculi? It certainly appears so, if by this we mean a maximalist position, according to which contemporary LLMs are able to simulate nontrivial social phenomena. Experimental results, our own and many of those cited in this article, indicate that there is reason to remain careful. In the specific case of survey answers, we find that current LLMs are plagued by machine bias: an inability to adapt to social variations resulting in poor predictions that cannot be accounted for by the unrepresentative nature of their training data.

Our conclusion does not suggest that the use of LLMs for social science research, and beyond, should be dismissed entirely. Provided we can better understand how they work, and avoid the pitfalls of both anthropomorphism and prophetism (Narayanan and Kapoor 2024), it is likely that LLMs will become useful tools in the day-to-day work of social science researchers.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: this work is supported by two grants from the Agence Nationale de la Recherche (Labex Ecodec/ANR-11-LABX-0047, TSIA Pantagruel/ANR-23-IAS1-0001), a Hi! Paris Collaborative Project research grant, and funding from the Swedish Vetenskapsrådet ("Mining for Meaning").

ORCID iDs

Julien Boelaert https://orcid.org/0000-0001-8675-1857
Samuel Coavoux https://orcid.org/0000-0001-7991-3555
Étienne Ollion https://orcid.org/0000-0003-3099-5240
Ivaylo Petev https://orcid.org/0000-0002-1563-655X
Patrick Präg https://orcid.org/0000-0001-6175-8470

Data and Code Availability Statement

The data and code used for this article are available online (Boelaert et al., 2025) https://doi.org/10.17605/OSF.IO/YZECT.

Supplemental Material

Supplemental material for this article is available online.

Notes

1. The alphabetical order matches the authors' contributions. Previous versions of this study were presented at the "Using LLMs and Text-as-Data in Political Science Research" workshop in Barcelona, and the "Generative Artificial Intelligence and Sociology" workshop at Yale. We thank the participants for their useful feedback. We are also grateful to Delia Baldassari, Gaël Le Mens, Charles Rahal, Felix

Elwert, the editors of the special issue and the anonymous reviewers for insightful comments.

We acknowledge the national infrastructure Huma-Num for providing us access to their servers, and Gérald Foliot who helped us navigate their intricate architecture. We are also grateful to Annina Claesson for excellent copyediting.

2. In a now classic paper, Argyle et al. (2023: 337) wrote "we propose that these (large-scale generative language models) can be used as surrogates for human respondents in a variety of social science tasks."
3. One company claims it can produce "synthetic boosters to gain reliability in undersampled groups," another one "draws on the growing body of research demonstrating the capacity of AI to simulate surveys, experiments and other empirical studies," and a third one promises "to create accurate synthetic interviews and surveys" (all citations were collected in the Fall of 2024).
4. It is interesting to note that the representative and the social bias hypotheses share an underlying assumption, namely that these models accurately mimic at least some human attitudes and behaviors. They disagree on the extent to which LLMs reflect social diversity: can they mimic all social groups (representative), or only a few (social bias)?
5. In LLM jargon, "zero-shot" refers to the practice of asking the LLM to generate answers without giving it examples of correct answers first. Giving such prior examples is called "few-shot." Yet another way to make LLMs learn from prior data is to fine-tune them on a training dataset, which is computationally more costly.
6. Dominguez-Olmedo et al. (2024), on the contrary, report high entropy of categorical LLM answers, that tend to be close to a uniform distribution.
7. Note that high adaptability is compatible with low entropy, and vice versa. For instance, if an LLM outputs highly skewed but distinct answer distributions for different subpopulations, it would display low entropy within populations, but high adaptability.

   The notion of adaptability can be related to the "steerability" investigated by Santurkar et al. (2023).
8. We ran the analyses with prompts in English as well as in the vernacular language of Mexico, Germany, and Russia to check whether, in addition to period, country, and question, the linguistic dimension would alter the prediction quality of LLMs. It did not. To focus on our main result, we only present the results for English-language prompts.
9. The question about political ideology was not included in the 2006 wave for Russia.
10. For a robustness check on alternative prompting strategies, see Supplemental Materials section D.2.

11. For observations where some sociodemographics were missing in the WVS, the question–answer pairs of these variables were dropped from the prompt.
12. Of these models, GPT-4-turbo is the only one fine-tuned for chat, as no foundational model is available for this class. Further experiments reported in Supplemental Materials section C.1 indicate that models fine-tuned for instruction or chat tend to perform worse than pretrained foundational models on this task.
13. The "compliance" of the LLM can be assessed at each generation by measuring the portion of total probability given to the valid answers. In our experiments, we found this to be very high (above 98%, see Table S4 in the Supplemental Materials), except in the case of instruct-tuned LLMs on "sensitive" questions such as politics and religion. This result is reassuring regarding the critique of the NTP method by X. Wang et al. (2024).
14. The deterministic nature of NTP removes the need for setting a temperature level or other generation hyper-parameters. We found, however, NTP not to be deterministic for GPT-4-Turbo, with sizeable variations of answer probabilities for several runs of an identical prompt, with worrying implications for robustness and reproducibility.
15. Consider a sample of 100 observations in WVS with the same sociodemographic properties, and therefore the same prompt. Supposing the LLM could predict their exact answer distribution, then its one-to-one answer accuracy would randomly vary from 0 to 100%, depending on the order in which WVS and LLM observations were compared.
16. The combination of the five stratifying variables within each country-year wave, missing values included, results in a total of 13,905 unique profiles, for an average of around two observations per profile. In order to strike a balance between the robustness of the answer distributions and a sufficient number of final subpopulations, we aggregate the small initial subpopulations using the following greedy iterative algorithm: at each step, until all subpopulations have at least 20 observations, the smallest group is merged with its least-populated closest sociodemographic group.
17. The political question being absent from the 2006 Russian data, the corresponding 48 subpopulations were dropped from the analyses of this question.
18. Robustness checks using two alternative distance measures (the Jensen–Shannon divergence and the total variation distance) did not affect the conclusions of this article. Neither of these two distances is perfectly correlated with the EMD on our data: 0.3 for Jensen–Shannon, 0.6 for total variation. The rationale for using EMD is that is the only one that takes into account the ordered nature of our survey answers, that is, the fact that "Very often" should be considered closer to "Often" than to "Never."
19. Note that if all WVS answers are in the central position, the normalized EMD to the LLM prediction is necessarily lower than 0.5. Despite this boundedness, we

empirically found the normalized EMD to be more suitable to comparisons across questions than an alternative standardization where each EMD is divided by its maximum value given the distribution of WVS data.

20. We generate 20 independent random permutations for each outcome question, and report their aggregated prediction quality.
21. Specifically, we use leave-one-out cross-validation, in which each observation (i.e., subpopulation) is predicted by a model trained on all observations except itself. For N observations, there are thus N models, each trained on $N - 1$ observations. Cross-validation is used to measure, rather than just a fit on the training data, the model's generalization capability, that is, its ability to predict data it hasn't been exposed to yet. We use the simplest possible specification, with no interactions between the seven sociodemographic predictors.
22. Additionally, the good performance of the linear baseline indicates that the conditioning variables are strong predictors of the outcome answers in the ground truth data.
23. Note that if there was no strong link between answers and controls in the ground-truth data, the R^2 of the Social model would mechanically be small; it would also render the social bias hypothesis irrelevant. This is not the case in our data, as the control variables are rather good predictors of the WVS answers (see the linear baseline model *supra*).
24. The extreme in this regard is attained on the binary Trust question by GPT-4, a model that seems to have learned never to trust people. More generally, out of the 13 models we tested, only the four instruct models (among which GPT-4) were consistent with Bisbee et al. (2024)'s result of low entropy, where each prompt leads to answers very concentrated on a single level. Foundational models showed more balanced answer distributions, where several answer levels have equivalent probabilities.
25. Horton (2023) rightfully noted that "one advantage economists have in using LLMs is that they tend to pose questions that place few demands on the sample. We do not think of demand curves sloping downward as a 'Western, Rich, Industrialized, and Democratic' phenomenon but rather as a result of rational goal-seeking that nearly all humans engage in" (Horton 2023: 6).
26. GPT 3.5, for instance, is closer to Russia on the politics question, and Llama-3-8B mostly echoes the answers of Mexicans to the happiness question (see Supplementary Materials section E4).
27. Furthermore, during and after training, LLMs are selected according to their performance on a series of natural language and reasoning tasks, much less on factual knowledge capabilities.
28. Moreover, the concept of emergent capabilities is prone to fuzzy definitions, as noted by Rogers and Luccioni (2024).

References

Aher, Gati, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies." *arXiv*. doi: 10.48550/arXiv.2208.10264.

Alvero, A. J., Jinsook Lee, Alejandra Regla-Vargas, René F. Kizilcec, Thorsten Joachims, and Anthony Lising Antonio. 2024. "Large Language Models, Social Demography, and Hegemony. Comparing Authorship in Human and Synthetic Text." *Journal of Big Data* 11(138):1–28. doi: 10.1186/s40537-024-00986-7

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31(3):337–51. doi: 10.1017/pan.2023.2.

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

Atari, Mohammad, Mona J. Xue, Peter S. Park, Damián Ezequiel Blasi, and Joseph Henrich. 2023. "Which Humans?" *PsyArXiv preprint*. doi: 10.31234/osf.io/5b26t.

Bail, Christopher A. 2024. "Can Generative AI Improve Social Science?" *PNAS* 121(21):e2314021121. doi: 10.1073/pnas.2314021121

Barrie, Christopher, Elli Palaiologou, and Petter Törnberg. 2024. "Prompt Stability Scoring for Text Annotation with Large Language Models." *arXiv Preprint*. doi: 10.48550/arXiv.2407.02039.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021 "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" Pp. 610–23 in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*. New York, NY, USA: Association for Computing Machinery.

Berglund, Lukas, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. "The Reversal Curse: LLMs Trained on 'A Is B' Fail to Learn 'B Is A'." *arXiv Preprint*. doi: 10.48550/arXiv.2309.12288.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis* 32(4):401–416. doi: 10.1017/pan.2024.5

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." Pp. 4356-4364 in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, edited by Daniel D. Lee,

Masahi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett. Red Hook, NY: Curran.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond. 2025. "Generative AI at Work." *Quarterly Journal of Economics*: 1–54. doi: 10.1093/qje/qjae044

Buolamwini, Joy and Timnit Gebru. 2018. "Gender Shades. Intersectional Accuracy Disparities in Commercial Gender Classification" Edited by S. A. Friedler and C. Wilson." *Proceedings of Machine Learning Research* 81:77–91.

Cao, Yong, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023 "Assessing Cross-Cultural Alignment Between ChatGPT and Human Societies: An Empirical Study." Pp. 53–67 in *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (c3nlp)*, edited by S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, and L. Benotti. Dubrovnik, Croatia: Association for Computational Linguistics.

Chen, Irene Y., Peter Szolovits, and Marzyeh Ghassemi. 2019. "Can AI Help Reduce Disparities in General Medical and Mental Health Care?" *AMA Journal of Ethics* 21(2):E167–79. doi: 10.1001/amajethics.2019.167.

Daikeler, Jessica, Michael Bošnjak, and Katja Lozar Manfreda. 2019. "Web Versus Other Survey Modes. An Updated and Extended Meta-Analysis Comparing Response Rates." *Journal of Survey Statistics and Methodology* 8(3):513–39. doi: 10.1093/jssam/smz008.

Dillion, Danica, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. "Can AI Language Models Replace Human Participants?" *Trends in Cognitive Sciences* 27(7): 597–600. doi: 10.1016/j.tics.2023.04.008.

Dominguez-Olmedo, Ricardo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. "Questioning the Survey Responses of Large Language Models." *arXiv preprint*. doi: 10.48550/arXiv.2306.07951.

Dutwin, David and Trent D. Buskirk. 2021. "Telephone Sample Surveys: dearly Beloved or Nearly Departed? Trends in Survey Errors in the Era of Declining Response Rates." *Journal of Survey Statistics and Methodology* 9(3):353–80. doi: 10.1093/jssam/smz044.

Gallegos, Isabel O., Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. "Bias and Fairness in Large Language Models: A Survey." *Computational Linguistics* 50(3):1097–179. doi: 10.1162/coli_a_00524.

Glass, Jennifer, Robin W. Simon, and Matthew A. Andersson. 2016. "Parenthood and Happiness. Effects of Work–Family Reconciliation Policies in 22 OECD Countries." *American Journal of Sociology* 122(3):886–929. doi: 10.1086/688892

Hartmann, Jochen, Jasper Schwenzow, and Maximilian Witte. 2023. "The Political Ideology of Conversational Ai: Converging Evidence on ChatGPT's

Pro-Environmental, Left-Libertarian Orientation." *arXiv Preprint*. doi: 10.48550/arXiv.2301.01768.

Heyde, Leah von der, Anna-Carolina Haensch, and Alexander Wenz. 2023. "Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion." *SocArXiv preprint*. doi: 10.31235/osf.io/8je9g.

Horton, John. 2023. "Large Language Models as Simulated Economic Agents. What Can We Learn from Homo Silicus?" *NBER Working Paper* 31122. doi: 10.3386/w31122.

Hovy, Dirk and Shrimai Prabhumoye. 2021. "Five Sources of Bias in Natural Language Processing." *Language and Linguistics Compass* 15(8):e12432. doi: 10.1111/lnc3.12432

Inglehart, Ronald, Christian Haerpfer, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen, 2022. *World Values Survey (WVS). All Rounds—Country-Pooled Datafile. Vienna: worldvaluessurvey.org, 4.0 ed.* doi: 10.14281/18241.17

Johnson, Rebecca L., Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. "The Ghost in the Machine Has an American Accent. Value Conflict in GPT-3." *arXiv preprint*. doi: 10.48550/arXiv.2203.07785.

Kim, Junsol and Byungkyu Lee. 2023. "AI-Augmented Surveys. Leveraging Large Language Models and Surveys for Opinion Prediction." *arXiv preprint*. doi: 10.48550/arXiv.2305.09620.

Kozlowski, Austin C., Hyunku Kwon, and James A. Evans. 2024. "In Silico Sociology. Forecasting Covid-19 Polarization with Large Language Models." *arXiv Preprint*. doi: 10.48550/arXiv.2407.11190.

Lin, Zhicheng. 2024. "How to Write Effective Prompts for Large Language Models." *Nature Human Behavior* 8:611–15. doi: 10.1038/s41562-024-01847-2

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys (CSUR)* 54(6):1–35. doi: 10.1145/3457607

Motoki, Fabio, Valdemar Pinho Neto, and Victor Rodrigues. 2024. "More Human Than Human. Measuring ChatGPT Political Bias." *Public Choice* 198(1):3–23. doi: 10.1007/s11127-023-01097-2

Nangia, Nikita, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020 "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models." Pp. 1953-67 in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* edited by B. Webber, T. Cohn, Y. He, and Y. Liu. Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154

Narayanan, Arvind and Sayash Kapoor. 2024. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton, NJ: Princeton University Press.

Navigli, Roberto, Simone Conia, and Björn Ross. 2023. "Biases in Large Language Models: Origins, Inventory, and Discussion." *J. Data and Information Quality* 15(2):1-21. doi: 10.1145/3597307

Ollion, Etienne, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2024. "The Dangers of Using Proprietary LLMs for Research." *Nature Machine Intelligence* 6:4–5. doi: 10.1038/s42256-023-00783-6

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Pearson, Helen. 2015. "Massive UK Baby Study Cancelled." *Nature* 526(7575):620–21. doi: 10.1038/526620a

Plaza-del-Arco, Flor Miriam, Debora Nozza, and Dirk Hovy. 2024 "Wisdom of Instruction-Tuned Language Model Crowds. Exploring Model Label Variation." Pp. 19–30 in *Proceedings of the 3rd Workshop on Perspectivist Approaches to nlp (Nlperspectives) @ Lrec-Coling 2024*, edited by G. Abercrombie, V. Basile, D. Bernadi, S. Dudy, S. Frenda, L. Havens, and S. Tonelli. Torino, Italia: ELRA; ICCL.

Rogers, Anna and Sasha Luccioni. 2024. "Position: Key Claims in LLM Research Have a Long Tail of Footnotes." *arXiv*. doi: 10.48550/arXiv.2308.07120

Röttger, Paul, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024 "Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models." Pp. 15295–311 in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand: Association for Computational Linguistics.

Rozado, David. 2024. "The Political Preferences of LLMs." *PLOS ONE* 19(7):1–15. doi: 10.1371/journal.pone.0306621

Rutinowski, Jérôme, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. 2024. "The Self-Perception and Political Biases of ChatGPT" Edited by S. G. Fashoto." *Human Behavior and Emerging Technologies* 2024(7115633):1–9. doi: 10.1155/2024/7115633.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. ""Whose Opinions Do Language Models Reflect?" Edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett." *Proceedings of Machine Learning Research* 202:29971–30004.

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 2023 "Are Emergent Abilities of Large Language Models a Mirage?" Pp. 55565–81 in *Advances in Neural*

Information Processing Systems. 36, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Curran Associates, Inc.

Schilke, Oliver, Martin Reimann, and Karen S. Cook. 2021. "Trust in Social Relations." *Annual Review of Sociology* 47:239–59. doi: 10.1146/annurev-soc-082120-082850.

Schnabel, Tobias, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. "Recommendations as Treatments: Debiasing Learning and Evaluation." *arXiv*. doi:10.48550/arXiv.1602.05352

Shankar, Shreya, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. "No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World." *arXiv*.

Spirling, Arthur. 2023. "Why Open-Source Generative AI Models Are an Ethical Way Forward for Science." *Nature* 616(7957):413. doi: 10.1038/d41586-023-01295-4

Strimling, Pontus, Joel Krueger, and Simon Karlsson. 2024. "GPT-4's One-Dimensional Mapping of Morality: How the Accuracy of Country-Estimates Depends on Moral Domain." *arXiv preprint*. doi: 10.48550/arXiv.2407.16886.

Voas, David and Mark Chaves. 2016. "Is the United States a Counterexample to the Secularization Thesis?" *American Journal of Sociology* 121(5):1517–56. doi: 10.1086/684202

Wang, Xinpeng, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. ""My Answer Is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models." Pp. 7407–16 in *Findings of the Association for Computational Linguistics ACL 2024*, edited by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand; virtual meeting: Association for Computational Linguistics.

Wang, Angelina, Jamie Morgenstern, and John P. Dickerson. 2024. "Large Language Models Should Not Replace Human Participants Because They Can Misportray and Flatten Identity Groups." *arXiv preprint*. doi: 10.48550/arXiv.2402.01908.

Williams, Douglas and J. Michael Brick. 2018. "Trends in US Face-to-Face Household Survey Nonresponse and Level of Effort." *Journal of Survey Statistics and Methodology* 6(2):186–211. doi: 10.1093/jssam/smx019

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics* 50(1):1-55. doi: 10.1162/coli_a_00502

Zoph, Barret, Colin Raffel, Dale Schuurmans, Dani Yogatama, Denny Zhou, Don Metzler, Ed H. Chi, Jason Wei, Jeff Dean, Liam B. Fedus, Maarten Paul Bosma, Oriol Vinyals, Percy Liang, Sebastian Borgeaud, Tatsunori

B. Hashimoto, and Yi Tay. 2022. "Emergent Abilities of Large Language Models." *arXiv*. doi: 10.48550/arXiv.2206.07682

Author Biographies

Julien Boelaert is an assistant professor of political science at the Centre d'études et de recherches administratives, politiques et sociales (CERAPS), University of Lille. His research focuses on political sociology and computational methods.

Samuel Coavoux is an assistant professor of sociology at the Center for Research in Economics and Statistics (CREST) and at École nationale de la statistique et de l'administration économique (ENSAE). His research focuses on digital uses and cultural inequalities.

Étienne Ollion is a senior researcher at CNRS and a professor of sociology at l'École polytechnique. His work focuses on the transformations of political fields in Europe, as well as on how to integrate machine learning and LLMs into social scientific research. Recent work includes *The Candidates. Amateurs and Professionals in French Politics* (Oxford University Press, 2024), and publications in outlets such as *Nature Machine Intelligence* and *Sociological Methods and Research*.

Ivaylo Petev is a CNRS research fellow at the Center for Research in Economics and Statistics (CREST) and Lecturer in Sociology at Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE). His research explores questions related to inequalities in consumption and lifestyles, environmental practices and discrimination from historical and comparative perspectives. His work has appeared in *Socius*, *American Sociological Review*, and *Economics and Statistics*.

Patrick Präg is an associate professor of sociology at the Center for Research in Economics and Statistics (CREST) and at École nationale de la statistique et de l'administration économique (ENSAE). His research focuses on social cohesion and social inequality. His work has been published in *Social Forces*, *Demography*, and *European Societies*.