


The Problems of LLM-generated Data in Social Science Research

Luca Rossi*  ^a

Katherine Harrison  ^c

Irina Shklovski  ^{b, c}

^a Department of Digital Design, IT University of Copenhagen (Denmark)

^b Department of Computer Science, Department of Communication, University of Copenhagen (Denmark)

^c Department of Thematic Studies – Gender Studies, Linköping University (Sweden)

Submitted: May 17, 2024 – Revised version: September 6, 2024

Accepted: September 23, 2024 – Published: October 30, 2024

Abstract

Beyond being used as fast and cheap annotators for otherwise complex classification tasks, LLMs have seen a growing adoption for generating synthetic data for social science and design research. Researchers have used LLM-generated data for data augmentation and prototyping, as well as for direct analysis where LLMs acted as proxies for real human subjects. LLM-based synthetic data build on fundamentally different epistemological assumptions than previous synthetically generated data and are justified by a different set of considerations. In this essay, we explore the various ways in which LLMs have been used to generate research data and consider the underlying epistemological (and accompanying methodological) assumptions. We challenge some of the assumptions made about LLM-generated data, and we highlight the main challenges that social sciences and humanities need to address if they want to adopt LLMs as synthetic data generators.

Keywords: LLM; synthetic data; social science; research methods.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation.

*  lucr@itu.dk

1 Introduction

Despite the novelty of the technology, Large Language Models (LLMs) have seen nascent adoption by social scientists for research purposes in part due to the widespread public availability of such tools made possible by commercial offers. Within the social sciences we observe two approaches to LLM-produced textual content that seem to dominate. First, there are many studies that have creatively applied social science research methods to study algorithmic systems themselves (for an overview see Moats & Seaver, 2019; and articles in this Symposium). In particular, with the advent of LLMs we see a range of applications of social science methods from psychology experiments (Almeida et al., 2024) to variations on ethnography (Demuro & Gurney, 2024) as alternative ways to understand the capacities and limits of these technologies. Second, there are a number of studies that have used LLMs to augment existing social science methods either in data analysis or data generation (Møller et al., 2024). While all of these deserve discussion, in this essay we focus on the use of LLMs to produce — or experiment with producing — synthetic data as research material for social science questions. In particular, we are concerned with studies that seek to use such LLM-generated data to model or make inferences about how people might act, react, or respond in a variety of situations.

In a widely cited article, Grossmann et al. (2023) argue that the capacity of LLMs for “simulating human-like responses and behaviors offers opportunities to test theories and hypotheses about human behavior at great scale and speed” (p. 1108). The authors go as far as to imagine that “LLMs may supplant human participants for data collection” (p. 1109). This enthusiasm seems to be shared by an increasing number of scholars (e.g., Jansen et al., 2023; Aher et al., 2023; Argyle et al., 2023; Hämäläinen et al., 2023; Törnberg et al., 2023). At the same time, many scholars question some of the assumptions underlying these studies and suggest caution in deploying LLMs as social science data-generation mechanisms (e.g., von der Heyde et al., 2023; Bisbee et al., 2023; Agnew et al., 2024).

If LLMs could generate data as if they were “participants”, answering questions, making decisions or arguing for those, this would certainly represent a seismic shift for many social sciences that have often been struggling with hard-to-find participants, expensive data collection and questionable convenience samples. Yet while Grossmann et al. (2023) might claim that LLMs are now able to produce language in a “contextually aware and semantically accurate fashion” (p. 1108) current evidence within NLP research suggests that this is not quite the case just yet (Cui et al., 2023). While LLMs might represent a tempting opportunity for social science research given the ease and apparent facility in language production in response to carefully worded prompts, the use of such data for generating insights about people requires serious critical consideration.

In what follows, we review the extant literature and the emergent debate on the topic and discuss what we see as the main concerns with the use of LLMs as a source of social science research data.

First, we will discuss what type of synthetic data is actually produced by LLMs and in what way it differs from other approaches to producing synthetic data. Second, we will explore the implications of the cautions Grossmann et al. (2023) note — the fidelity of training data and considerations of its representativeness, the problem of LLM biases, the challenges of transformer-type models and their propensity for hallucinations, the emergent art of prompt engineering and the idea of benchmark selection. Third, we will consider what kind of knowledge can be gained from the analysis of LLM-produced data if we were to imagine that the problems of benchmarking, transparency in model training, and data fidelity can be addressed

and what sort of legitimacy this knowledge might claim. Ultimately, we reflect on the proposition recently put forward by Bail (2024), that researchers could work towards the creation of open-source LLMs specifically trained and deployed for social research, by highlighting what we see as the major challenges that would need to be addressed.

2 Particularities of LLM-generated Data

LLM-produced data can be seen as a type of generated synthetic data. Synthetic data are datasets generated using purpose-built mathematical models or algorithms (Jordon et al., 2022) instead of being extracted from existing digital systems or produced through particular forms of data collection. The idea of synthetic data has a significant history in statistics and particularly within governmental and public institutions that need to make population data available for analysis while ensuring confidentiality (Abowd & Vilhuber, 2008) where population datasets have typically been treated through data augmentation and statistical disclosure control techniques (Raghunathan, 2021). Synthetic simulation data also has a long history in areas such as computer vision, which typically uses model specifications to generate datasets (Nikolenko, 2021). The advent of greater computational capacities and more complex machine learning models made synthetic data augmentation and generation faster, cheaper, less complex, and increasingly popular (Jordon et al., 2022).

Typically, synthetic data are seen to address any or all of three primary data challenges: data scarcity, data privacy, and data bias (Van der Schaar & Qian, 2023). While there is no doubt that generating synthetic data addresses the issue of scarcity by producing more data given a set of specifications, whether these data are of sufficient quality to be useful and whether they are able to address the challenges of privacy and bias depends on context (Abowd & Vilhuber, 2008; Belgodere et al., 2023; Figueira & Vaz, 2022; Jacobsen, 2023). With the advent of generative AI models in the development of synthetic data pipelines, we see the development of new frameworks proposed to evaluate the quality of the generated datasets. For example, Eigenschink et al. (2023) propose five assessment criteria for synthetically produced datasets: representativeness, novelty, realism, diversity, and coherence. They argue that high-quality synthetic data produced via generative AI should be able to capture population-level properties of the original data, create novel data points that are realistic (given what we know of the original data), and show internal diversity while maintaining coherency with the original data. They also note that the importance of the individual criteria varies significantly across domains and that the ways in which such criteria should be tested differ. Without delving too deeply into the specific framework, it is clear that the focus is on the ability of the synthetic data to reproduce key characteristics of the original data.

The development and rapid adoption of LLMs have led to the use of these systems for a broad range of data generation tasks. Whitney & Norman (2024) distinguish between generated, augmented, and procedurally created synthetic datasets differentiating between them “based on how derivative of a real-world training dataset they are” (p. 4). Procedurally created synthetic data rely on purpose-built models that create data along a set of explicitly pre-specified parameters, while generated data, such as what LLMs produce, arise from a model abstracting from a training dataset in response to a particular input. Here, the training dataset is crucial for ensuring that the resulting dataset is usefully similar to the data we might expect to collect from people. LLM-driven systems, however, are not built with faithful data generation as a goal. Rather than aiming to produce data that resemble a given original dataset, LLMs have been trained to predict the occurrence of the next most likely letter, word, or group of words

given the linguistic patterns of a text. Trained on ever larger amounts of data, LLMs are able to mimic human-produced content (Jakesch et al., 2023) showing emergent abilities to perform tasks that were not explicitly present in the training data¹ (Radford et al., 2019; Wei et al., 2022). There is no effort here to adhere to particular pre-existing patterns in the data.

Given the ease with which current LLM implementations can generate human-like text on a near-infinite number of topics, it was not a huge leap to imagine the use of these technologies for generating data for research purposes. Seen as a convenient type of data, such LLM-generated datasets are typically obtained given sufficient effort in prompt engineering and, occasionally, the use of different available LLM implementations for comparison (Dillion et al., 2023, Horton, 2023). It is possible to classify LLMs as a type of synthetic data generator (Jordon et al., 2022). Keeping in mind that the goal of synthetic data generators is to produce data that resemble particular aspects of real data while addressing issues such as privacy, lack of diversity or data scarcity, one might ask if these goals are achieved by LLMs.

On the surface, LLMs can certainly generate data that mimics human-produced data, thus potentially resembling aspects of real data. Yet there is one important caveat to consider. Synthetic data is typically evaluated for utility and fidelity — how useful they are for a particular task and how well they resemble a real dataset given parameters important to the task. Here different types of fidelity may be considered but the important question is what is necessary for the task, but they require some capacity to compare the data we intend to mimic or augment and the synthetic output. At times, fidelity issues can be quite insidious, as Johnson & Hajisharif (2024) demonstrate a lack of intersectional fidelity in their tests with GAN-produced synthetic census data. How are we to evaluate the fidelity of an LLM-produced dataset, especially given the lack of access to the models generating the data or their training data? How might we evaluate differences in these data and their implications? Despite continuous advances, LLMs continue to display a lack of internal consistency in output. For example, Atil et al. (2024) have recently reported how, when systematically studied, LLMs show a lack of stability even when the input is the same. This was observed despite all the hyper-parameters being set to maximize the deterministic nature of the model. Importantly, stability not only varied between different models (both commercial and open source) but also within the same model as a function of the task.

Questions of utility and fidelity, of course, will vary depending on the particular applications of LLM-produced data, which we will address in the rest of the essay.

3 Applications of LLM-produced Data in Social Science Research

While clear-cut divisions are inherently problematic, here we identify what we see as major streams in LLM adoption for data production in the context of different types of social science research. Organizing the existing methodological experimentation around three directions allows us to see the existing similarities and differences between these approaches as well as the shared underlying assumptions about “what” LLMs can do and why. It is worth noting that while LLMs have been used to generate a variety of data we will generally treat this as textual data for two main reasons. First, the underlying training data of LLMs are essentially textual in nature. Second, even when LLMs are used to produce numeric values (e.g., producing a

1. Several researchers have questioned the concept of emergent capabilities (see Schaeffer et al., 2024). The actual nature of these capabilities as well as their origin is irrelevant for the point of the specific article.

number that expresses agreement or disagreement on a scale) that is achieved through a textual prompt that asks the model to express the output as a number that would, otherwise, be textual.

When researchers use LLMs to produce synthetic research data, they are leveraging the underlying large amount of training data that characterizes these models as a proxy for individual data or data about specific groups and populations. This underlying, often unspoken, premise holds regardless of the specific research design and the specific LLM used. Within this perspective, there is no difference between using GPT-4, LLAMA-3 or Claude 3.5. What appears to be revolutionary is that LLM technology has reached a scale that allows emergent and unprecedented capabilities (Grossmann et al., 2023), irrespective of specific implementations. Nevertheless, the vast majority of research in this overview relies on OpenAI's GPT-3.5 and GPT-4 models. This is not due to any explicit analysis of the specific capabilities of GPT versus alternative models but rather due to the easy access provided by OpenAI's well-developed set of APIs.

3.1 LLMs as an Improved Version of Agent-Based Modeling (ABM)

The excitement about LLM-produced data hinges on the model's capacity to simulate human-produced text (Grossmann et al., 2023; Bail, 2024). Controlled through carefully developed prompts, such text production has been used to augment or even replace other forms of agent-based simulation by several scholars (Park et al., 2023; Törnberg et al., 2023; Horton, 2023). While agent-based modeling has a long history, scholars in this domain have long struggled to overcome the limitations of ABM approaches: the necessary abstraction and simplification of the modeled context and the lack of capacity of these models to capture human discourse (Törnberg et al., 2023). Given LLMs' ability to produce text that reflects realistic human reasoning, such simulations would seem to address these major shortcomings of ABM.

In this context, LLMs have been used to create "personas" (Törnberg et al., 2023) through specific prompts defining the relevant personality traits for each manifestation. These agents were then used to create role-play situations following the prompted guidelines. The result showed a higher level of emergent behavior when compared to mechanistic ABMs (Törnberg et al., 2023). Park et al. (2023) have created an architecture based on ChatGPT to generate computational software agents that present what they term "believable simulations of human behavior." Such social simulacra (Park et al., 2022) are used to explore real-world scenarios with increased nuance, not available to more traditional ABM approaches (Wu et al., 2023), achieved by ensuring backward and forward continuity (Argyle et al., 2023) as well as extended memory (Park et al., 2023).

While the development of ABMs can be quite complex, authors point out that it is possible to generate autonomous goal-oriented agents by using LLMs with well-designed prompts quickly and at little cost (Phelps & Russel, 2023). This direction has generated a great amount of enthusiasm and has led to the creation of ad-hoc solutions where less technical researchers can deploy LLM-based social simulations (Rossetti et al., 2024).

Despite the excitement, there are some cautions in deploying these approaches. While some researchers find the results of such explorations convincing (Törnberg et al., 2023) or believable (Park et al., 2023), others note that there are limitations to how well such models are able to replicate human behavior in simulations of well-known contexts such as the iterated Prisoner's Dilemma (Phelps & Russel, 2023).

Nevertheless, simulations produced through ABM or LLM-based efforts do not need to be entirely faithful to particulars of human behavior. After all, following George Box's famous

maxim, models can be useful even though all of them are wrong (Box, 1976). Where implementations of LLMs for modeling social contexts are used for insights into how people might act in a variety of situations, some additional caution is warranted. It is precisely because LLMs generate text, we notice an interesting trend towards personification (Jones et al., 2023) of these systems in the interpretation of results. In their exploration of how people relate to GPT-3 output, Jones et al. (2023) note that personification seems a common response, defining this as the tendency to seek a human-like intentionality behind the output. For example, Törnberg et al. (2023) present an interesting effort to simulate how different designs of social media platforms might affect the resulting toxicity of the posted content. They use LLMs to generate personas that then interact by producing simulated messages given prompts. They measure the toxicity of the resulting text and suggest particular designs as potentially more successful. However, in their interpretation, they seem to personify the simulated agents by noting that the agents were “[...] responding to the posts from the other side that trigger or upset them” (Törnberg et al., 2023, p. 6). Of course, LLMs can not be triggered or become upset, regardless of whether these systems are simulating a persona or simply producing text in response to a prompt. LLMs, after all, don’t have emotions but such personification is common in interaction with systems that produce text (Jones et al., 2023). Törnberg et al. (2023) seem to rely on such personification as a causal explanation for the evidence of increasing toxicity in the simulation potentially over-interpreting or oversimplifying the implications of their data.

Interpretation, of course, is the linchpin of any social science research and the question remains how to interpret LLM-generated output in this context. ABM researchers readily admit the limits and oversimplifications of their models (Phelps & Russel, 2023). Yet as Box (1976) reminds us, knowing how and why our models are wrong is what enables us to make them useful. There is no doubt that LLM-generated output is “wrong” in the Box sense, but how and why are difficult questions to answer. Thus interpretation of results may rely on personification and naive comparison to the researcher’s prior knowledge of contexts under study without any real relationship to what the output actually represents.

3.2 LLMs as Humans in the Bottle

A second stream of research uses LLMs to substitute research participants in what would traditionally be an experimental setting (Horton, 2023; Breum et al., 2023; Dillion et al., 2023). In this context, the role-playing ability of LLMs together with the ability to act according to specific instructions are used to generate particular interactions often between two instances of the same model. For example, treating LLMs as “implicit computation models of humans,” Horton (2023) draws on classical experiments in behavioral economics to demonstrate how the use of LLMs to simulate socio-economic decision-making and outcomes can move beyond theoretical economics as a way to generate insights that could be tested using more expensive methods of research with people. Horton readily acknowledges that LLMs are just as wrong as mathematical models of economic behavior but demonstrates how they can provide useful insight. This approach has also been used to study whether LLMs can reproduce dynamics of persuasion typical of human social systems (Breum et al., 2023) or if they can replicate well-known economic and social psychological behaviors (Aher et al., 2023). Where some of this research turns social science methods to explore the limits and possibilities of LLMs, these studies also explore the potential of such approaches for advancing social science research in general.

One question this research explores is whether LLMs can “faithfully” reproduce human dynamics through text production (Aher et al., 2023). Some researchers focus on comparing

the output of LLMs with the results of well-known psychological or economic experiments. This research attempts to make an argument for exactly how well such “implicit computational models of humans” (Horton, 2023, p. 2) can perform, in order to assess how reliable these models might be for new experimental efforts (Aher et al., 2023). As such, the criteria used to evaluate the resulting data quality — and ultimately the ability of the model to act “as a proxy for a human subject” — are largely based on the ability of the model to reproduce outcomes in well-known, previously published papers. For example, Horton (2023) uses LLMs to simulate outcomes of a decision-making scenario of allocating the federal budget between highway and car safety programs, originally presented in a well-known paper by Samuelson and Zekhauser (1988). Results appear to show that the more advanced GPT-3 Davinci model can replicate the status quo bias demonstrated in the original paper.

Aher et al. (2023) propose the term “Turing Experiments (TE)” as a means of evaluating AI systems “in terms of its use in simulating human behavior in the context of a specific experiment” (p. 1). They replicate, among others, the famous controversial shock experiment designed by Milgram in 1963, where subjects were asked to shock the victim (an actor in another room) with an increasingly high voltage. The experiment was originally intended to demonstrate how far people are willing to go to conform to authority demands in the face of causing harm and pain to someone. While ideally, simulations, such as those presented by Horton (2023) or Aher et al. (2023), ought to be zero-shot, the fact that LLMs have been trained on the vast corpora of Internet data generally means that these data are likely to include prior descriptions of these famous experiments. To mitigate this factor, Aher et al. (2023) augmented the original experiment in ways that they argued maintained the integrity of the results. They compare the level of compliance observed by Milgram and reported in the 1963 publication with the level of “compliance” simulated by the LLM, noting the similarity between simulated and experimentally observed outcomes.

The idea here is that the similarity of LLMs’ outcomes to published experimental data demonstrates how faithfully a model is capable of reproducing human behavior. This provides a legitimate argument for the use of these systems for validating new hypotheses about human behavior, especially where more traditional modes of data collection can be difficult or prohibitively expensive. Part of the problem with this argument is the fundamental assumption that prior experimental results are representative of human responses — an assumption that has been repeatedly called into question, especially around classic social psychology experiments of conformity conducted by Milgram & Ash (Greenwood, 2018; Henrich et al., 2010). The capacity of these models to reproduce such experiments is likely more reflective of a collective Western conviction that these experiments represent human behavior, rather than reflecting or representing human behavior. The famous psychology experiments were intended to demonstrate that our own beliefs and stories about why we do what we do are faulty. The question then is how should LLMs’ output be interpreted correctly in such studies.

3.3 LLMs as Respondents to Surveys or Interviews

While this is the least common of the three streams of research and experimentation that we have identified, it is also the most problematic. The use of LLMs in assisting survey research spans the gamut from generating survey questions, pre-testing survey instruments, or analyzing data and summarizing findings (Jansen et al., 2023). Some research, however, has explored the potential of LLMs to generate data that is then analyzed. In this section, we share examples of how such data has been tested for predicting human responses in fields as varied as political

theory, market research and design.

Although there are no studies yet that attempt to use LLM-produced data to make strong claims about human responses, several researchers are exploring this possibility. Argyle et al. (2023) presented one of the first efforts to demonstrate that LLMs can be used to generate data that replicates known distributions of particular response patterns in what they, similar to Horton's (2023) "homo silicus", call "silicone samples". They make the assumption that LLM output is based on underlying "human-like concept associations" where, "given basic human demographic background information" the output can model "underlying patterns between concepts, ideas, and attitudes that mirror those recorded from humans with matching backgrounds" (Argyle et al., 2023, p. 4). While such a statement is in agreement with the sentiment voiced by Grossmann et al. (2023), recent NLP research demonstrates that this is an overstatement of current LLM capacities. For example, transformer-based language models continue to have trouble generalizing beyond common linguistic constructions (Cui et al., 2023).

Similarly, Brand et al. (2023) explore the capacity of LLMs to respond to survey questions in a way that is consistent with economic theories and known consumer behavior patterns. Motoki et al. (2024) deploy several well-known survey instruments about organizational behavior and compare LLM-generated responses to published papers, noting that despite some limitations the outcomes do replicate human behavior and can potentially be used to validate survey instruments. In contrast, von der Heyde et al. (2023) generate LLM-based personas based on German voting data and show that the LLM-generated outcomes tend to be more biased and inaccurately predict voter choices. There is an emergent debate in the field where several studies have demonstrated that LLM-generated data tends to be significantly unrepresentative, arguing that perhaps such models are unfit for research applications (Simmons & Savinov, 2024; Bisbee et al., 2023; Santurkar et al., 2023).

Going beyond survey responses, Hämäläinen et al. (2023) explore whether LLMs can be productively used for qualitative research, specifically in design and user-experience research. They generate interview responses using persona-based prompts and compare the outcomes to published interview data. While there is an agreement that LLMs may not be particularly useful for predicting human responses to a range of cues, it is argued that such social simulations may nevertheless be useful for design purposes (Park et al., 2023; Hämäläinen et al., 2023). Designers have long used techniques such as developing personas and imagining responses to particular interactions with technology (Salminen et al., 2022) loosely based on research with potential users. Thus it is not a far cry to imagine how LLMs could be used for a similar purpose. Further, design research has frequently struggled with the problems of representation and inclusiveness — where user research focused on easily accessible people thus failing to address edge users and lacking diversity in samples (Sin et al., 2021; Elsayed-Ali et al., 2023). Here again, LLMs may offer a seemingly reasonable alternative, especially given the fact that engineering prompts is perceived as easier and cheaper than recruiting people for user research (Hämäläinen et al., 2023).

Whether generating survey or interview responses, researchers argue that LLM-generated data could be useful as it is not only cheaper and quicker to produce, but it can also potentially address sample diversity challenges (Aher et al., 2023; Argyle et al., 2023; Bail, 2024). In a recent scoping review of the efforts to use LLM-generated data for various types of social research, Agnew et al. (2024) caution that the substitution of human subjects with "homo silicus" comes into conflict with core research ethics values of representation and inclusion. They argue that study participants have important discretionary powers when participating in research, such as

opting out, resisting or being able to point out misconceptions on the part of researchers. The use of LLM-generated data instead of human subjects then would shift these powers, making the resulting research inherently exclusionary. This would exacerbate already existing issues in user research, as scholars repeatedly point out LLMs tend to produce exaggerations of “stereotypical response patterns” (Simmons & Savinov, 2024; Bisbee et al., 2023) and reflect some opinions over others (Santurkar et al., 2023).

4 The Challenge of Representativeness, Privacy, Bias and Hallucinations

As scholarly excitement grows around the capacity to produce increasingly varied types of LLM-generated data, we return to the typical challenges such data are expected to address: data scarcity, privacy concerns and regulations, and lack of diversity and data bias. The papers we reviewed differed substantially in how they discussed these concerns.

The vast majority of the papers we reviewed were clearly motivated by the problem of scarcity. Results are often praised in light of the “low cost and high speed” of LLM data generation (Hämäläinen et al., 2023; Törnberg et al., 2023; Argyle et al., 2023). Agnew et al. (2024) also identify scarcity as the most common reason. As is often the case with social research, scarcity in this context is due to cost. For the most part, people are not exactly scarce — not in the way that medical images of patients affected by a rare disease can be — but they can be expensive or complicated to engage. As a result, a number of authors are enthusiastic about the possibility to scale research in social and behavioral science, where it has notoriously relied on small and unrepresentative samples (Bail, 2024; Grossmann et al., 2023; Horton, 2023). There is no doubt that LLM-produced data can surely come in any volume necessary and at a very low cost, yet it is not clear whether such scaling is, in fact, defensible or useful.

When considering the challenge of privacy and regulatory limitations, none of the papers attend to the issue, although human participants do require a growing level of privacy protection and this directly translates both in ethical limits as well as into augmented costs for data collection, processing, and storage. This is not unexpected. Existing research shows that the actual risk that LLM-based synthetic data poses via the generation of non-maliciously prompted data seems quite low (Yan et al., 2024) and it is fair to assume that LLM-based synthetic data would not fall under the protection of regulations such as GDPR and would not require complex reviews from research ethics committees.

It is the issue of data bias, diversity, and representativeness that is discussed extensively across the reviewed research. After all, for LLM-generated textual data to be viable for social science research, the capacity to produce data that is representative of populations of interest is key. What seems to be the bottom line for much of the existing research is well exemplified by Argyle et al. (2023) when they argue that “algorithmic bias” in LLMs should be treated not as a macro-level property to be corrected, but as a feature that allows the model to produce outputs that reflect expected biases in the population and different subgroups. This argument builds on the idea that, since LLMs are trained on massive amounts of online data, the data will be able to capture fine details of the social system and of the several populations in it. This assumption is often paired with the assumed ability of LLMs to be conditioned, through prompting or fine-tuning, to assume specific points of view. In this way, LLMs are able to “extract” responses that faithfully represent actual subgroups or demographics from their massive amount of training data.

Yet there are many papers that document how LLMs tend to fail to generate output that is representative of various population subgroups (Bisbee et al., 2023; Simmons & Savinov,

2024; von der Heyde et al., 2023; Cao et al. 2023). These apparently contradictory results are not surprising at this early stage. Given that the sheer amount of data needed to achieve the performance reached by recent models is hardly obtainable through curated datasets, it is difficult to know exactly what the specific model ingested as data as well as what information about specific subgroups and with what level of reliability could be extracted from it. Such capacity to create what would essentially amount to data segmentation by sub-groups appears to be one of the key arguments in favor of LLM-generated data (Argyle et al., 2023; Aher et al., 2023).

“Bias as a feature to be exploited” is a cornerstone of the idea of algorithmic fidelity. Scholars argue that biased training data and its incorporation into the model is what gives the model the ability to faithfully reproduce social groups (Argyle et al., 2023). At the same time, since their large-scale commercial deployment, biases in LLMs’ outputs have been at the center of public attention (Gordon, 2023) as well as academic research (Fang et al., 2024). So much so that we have witnessed several attempts by commercial companies such as Google and OpenAI to mitigate model bias in their final output, often with mixed results (Goodman & Sandoval, 2024). Even when accepting the idea of algorithmic fidelity as unproblematic, researchers’ interests and platforms’ commercial plans do not seem aligned and research into the level of bias that is actually present in the final outputs of commercial models shows contradictory results (Tjuatja et al., 2023).

There are two fundamental questions — still largely unanswered — that suggest a careful approach to the idea of algorithmic fidelity and its consequent concept of algorithmic or “silicone” sampling (Argyle et al., 2023). First, what is the actual amount of bias that LLMs can reproduce? Second, what is the relation between the training data and emergent behaviors displayed by the models?

4.1 The Problem of Bias and Representativeness

Questions of bias and representativeness have spurred several studies (Bisbee et al., 2023; Simmons & Savinov, 2024; von der Heyde et al., 2023). For example, Tjuatja et al. (2023) evaluated whether nine LLMs exhibit human-like response biases in survey questionnaires. Following Törnberg et al. (2023) and Aher et al. (2023), this work leverages a framework widely used in social psychology that aims to elicit bias by changing the wording of prompts. The results demonstrated that LLMs’ output is not aligned with the expected human behavior such as a “significant change in the opposite direction of known human biases, and a significant change to non-bias perturbations” (Tjuatja et al., 2023, p. 2). These observations echo research by Santurkar et al. (2023) reporting substantial differences between the views reflected by several LLMs and those of many US demographic groups, noticeable even when the model was specifically prompted to represent a particular group.

In addition to showing poor bias-reproduction the work from Tjuatja et al. (2023) also showed that LLMs that used Reinforced Learning Human Feedback (RLHF) resulted in fewer changes to question modifications as a result of response biases. Reinforced Learning Human Feedback is a specific technique that allows the model to be trained on human-feedback rather than just on data alone. This is largely used to mitigate known biases and unwanted behaviors. While the application of RLHF may result in better “products” for the general user with models that are overall more harmless and helpful (Sun, 2023), it contradicts the assertion that the inherent bias of LLMs is what affords its representativeness (Argyle et al., 2023). While the adoption of vanilla models — that have not gone through the process of RLHF — showed

some benefit, the number of researchers in the social sciences who can realistically use LLMs outside of the commercial offer, is, at the moment, quite modest.

The issue of representation gets even thornier if we consider the capacity (or lack thereof) of LLMs to address cultural diversity in human populations (Cao et al., 2023). The use of these models runs the risk of “value lock-in” (Weidinger et al., 2022) as LLMs are not able to respond to subtle changes in normative positions and opinions in the population over time. Agnew et al. (2024) point out that the use of LLMs supports notions of representation in research only in a very weak sense, unresponsive to changes in opinions, views, and preferences. As a result, studies using LLM-generated data run the risk of misrepresentation of smaller, potentially more vulnerable populations is high, essentially reproducing age-old data colonialism problems of social research (Couldry & Mejias, 2019).

4.2 The Problem of Emergent Behaviors

The second question that demands careful consideration is the tendency of LLMs towards hallucination and emergent behaviors. Transformer-based models have a well-documented tendency to hallucinate, typically defined as the production of factually incorrect yet convincing information (McKenna et al., 2023). Since in the context of LLM-based data generation the goal is not to retrieve specific information from the training data, it might seem that the problem of hallucination is not relevant to the task at hand (and this might explain why it is never mentioned in the research papers we have reviewed). Nevertheless, we would argue otherwise. Recent research from McKenna et al. (2023) shows how sentence memorization and statistical patterns in the training data are major causes of hallucinations. In both cases hallucinations are not caused by emergent properties but by “overreliance” on the sentences or the statistical patterns that have been learned from the training data. This has three possibly important consequences for data generation. First, hallucinated responses would be perfectly “believable” but, since they do not refer to any factual information, they will be harder to identify. Second, the ability of LLMs models to be effectively conditioned to reasoning outside of its training data can be limited. Third, this ability might not be equal for all the possible sub-populations researchers might want to study. This expectation that LLMs should be *segmentable*, able to reproduce multiple sub-population, is a key element in the approaches that use LLMs supporting ABMs. Here (see Törnberg et al., 2023) LLMs are explicitly asked to role-play different positions on a specific issue. We call this expectation segmentability and it is worth noticing that even if the model should preserve algorithmic fidelity to the training data, this does not imply that the model would be able to be segmented and produce data representative of various population sub-groups. This needs to be evaluated on a case-by-case basis suggesting problems of replicability and legitimacy of the resulting insights.

5 The Art and Challenge of Prompt Engineering and Evaluation

Social science research relies on robust methodological descriptions for evaluating research output and, in some fields, for ensuring replicability of results. With LLM-generated textual data, the methodological descriptions typically focus on prompt engineering as many papers argue that “proper conditioning” (Argyle, 2023) is key to ensuring fidelity of this kind of research. In many cases (with the notable exception of Park et al. 2022, 2023) prompt engineering is described as a tuning process necessary to achieve the best outputs/responses from the LLM,

rather than a process with possibly profound consequences for the resulting data. Horton (2023) provides a good example of how prompts are often simply “listed”.

Listing the prompts used in the research process seems to speak more to the problem of transparency of research procedures and replicability than to the problem of data production. If prompt engineering is a matter of replicability of the results, this means that the selected prompt becomes the way to unlock the model’s ability to generate the desired data or the desired population. With the same prompting, the same or similar data would be produced again in the future. Yet, when investigating this specific assumption, Bisbee et al. (2023) found that generated data varies significantly both for small changes in the wording of the prompts as well as for the same prompt but asked at different moments in time. Similarly Atil et al. (2024) have reported overall instability of the output even when the conditions for deterministic behaviors are met.

If repeated prompts do not assure replicability of the research, then we have to consider how that should be documented. Lack of replicability can have many causes, from the hallucinatory nature of LLMs to the commercial nature of available platforms — platforms that are constantly updated and upgraded to offer an improved commercial product that does not need to be backward consistent. This has implications for how such data may need to be interpreted and what kinds of insights might be warranted. After all, differences due to emergent behavior have different implications to differences due to changes implemented at the platform level for commercial reasons.

Prompts lead us to consider evaluation and benchmarking. The papers we reviewed above offer different approaches to evaluating the resulting datasets for usefulness, fidelity, “faithfulness”, or “believability”. Where evaluations would typically cleave close to the purpose of data, they also need to be systematic and replicable (thus becoming consistent and robust research instruments). Current implementations run the gamut (thus Törnberg’s essay [2024] in this special issue) but they seem to often get reduced to measures of believability — does this output *look* like human output — which does not address the issue of usefulness given the assumptions of surveys or interviews and given the fact that people are just notoriously bad at distinguishing human and AI output even for older version of LLMs (Köbis & Mossink, 2021). Nor do they consider how normative ideas of what counts as “human” may be embedded in and reproduced through use of such evaluative measures (Rhee, 2018). Where believability is useful for creating non-player characters in computer games, because their goal is only to be “believable” in interactions with players, it isn’t a great measure of utility for making inferences about human responses. Just because the produced content is “believable” does not mean it has epistemic legitimacy.

This acknowledgement of different standards of “believability” draws attention to the importance the context of use and epistemological standpoint. For some social scientists, the idea that LLMs could produce “believable” material seems quite alien to the experience of, for example, conducting fieldwork to acquire relatively small amounts of qualitative data about lived experiences. For others, the material produced by an LLM may be sufficiently “believable” to be useful in a simulation. Underlying these differences are epistemological assumptions about how knowledge can and should be produced in order to say something useful about the world. Using “believability” as a way of assessing the material also tends to obscure darker questions: assessing “believability” requires a baseline in which humanness is quantified and measured, and can be used to compare outputs from LLMs. This quantification process reproduces highly problematic norms about who counts as human and why (D’Ignazio & Klein, 2023; & Gebru, 2018; Rhee, 2018).

The measure of faithfulness then may seem a better form of evaluation for some branches of the social sciences at least. Here we see replications of prior psychological or economics experiments with the idea that if LLM output aligns with what we know about how people respond to the known situation, then output produced in response to novel situations will be similarly aligned. Once again we run into several problems. The fields of psychology and economics have been going through a crisis of replicability — where scholars seem to be unable to replicate old and established experimental evidence with new studies with people. What does it say then if LLMs replicate the canon, even as it is challenged by studies with people? Many psychology and economics studies, as well as large-scale surveys in political science and demography, have been criticized for studying an unattainable ideal of the average person, statistically derived but non-existent in practice. Based on the results obtained by Bisbee et al. (2023) and by von der Heyde et al., (2023), LLMs might be in a similar situation, they can produce a set of averages — unattainable in practice.

As Horton (2023) notes, sure all models are wrong, LLMs included, but that does not mean we can't use them for thinking about what questions to ask and how to ask them. Yet it is worth reflecting on what questions might emerge given the particular notions of the average person embedded in LLMs and what kinds of questions might be left out. After all, if LLMs were to somehow produce data that might lead to fundamentally new questions, would it not by definition fail the test of faithfulness?

6 The Question of Legitimacy and Situatedness of Knowledge

While much current work acknowledges the limitations of the LLM-generated data and explores using currently available technology, there is also substantial agreement that we can expect the quality of LLM-generated textual data to improve as more complex models come online, model architectures evolve and data curation methods become ever more sophisticated (Hämäläinen et al., 2023; Törnberg et al., 2023; Park et al., 2023). This, however, does not address the epistemological issues that we have considered in this paper. Stepping back from specific technical challenges, there remain broader epistemological questions which are provoked by the increasing interest in LLM-generated data across the social sciences. Here we focus on two areas of concern: legitimacy and situated knowledge.

First, several of the papers we have analyzed build their justification on a specific version of the data scarcity argument. Data is scarce for many reasons due to the costs and challenges associated with recruiting people for research studies. LLMs promise an infinite number of quasi-human participants, lowering that cost by making quasi-human data abundant. While the cost of data is indeed a serious barrier for many researchers, it is not clear to what extent this can justify the use of LLMs without a thoughtful assessment of its epistemic legitimacy. Data scarcity is also just as likely to indicate that the research participants are unwilling to take part or have never been considered as “valid” participants before. In such cases, there may be no existing data with which to compare, or the data may be considered highly problematic. In such cases, LLM-generated data may exacerbate existing data inequalities. Following Agnew et al. (2024) critique, using such data to solve a data scarcity problem risks misrepresentation or ignores the “real” question of *why* there is no data. The research community was tempted, in the not-too-distant past, to assume that large quantities of digital data could be used as good proxies for complex social phenomena, only to find out that this was not the case (Jungherr et al., 2012).

Second, qualitative scholars have long struggled to claim the validity of their insights, especially in fields dominated by alternative epistemological positions where quantitatively produced knowledge with marginally defensible claims to generalizability was seen as the only legitimate sort. There is much excitement about the capacity to scale prior experiments and studies on limited samples through the use of LLM-generated data (Agnew et al., 2024; Bail, 2024). This stems from the underlying assumption that such data can be seen as more representative of population groups that the researchers wish to study, also known as fixing the diversity problem. Insights derived from such simulations then would only need limited substantiation in “the real world” as it were. There is nothing inherently wrong with simulations, but the challenge here is in understanding how much, and in what ways will LLM-generated outcomes differ from human answers, and in identifying how and in what ways these models may be wrong, especially if we have limited prior data with which to compare against. The danger, as we see it, is in the well-documented tendency of LLMs to produce output that reduces already expected diversity when compared to prior studies (von der Heyde et al., 2023), essentially replicating the status quo, because this is strongly embedded in the training data.

LLM transformer architectures innovate beyond the problem of the more traditional machine learning models, which in their reliance on past data for making predictions are by definition “always fighting the last war again” (Groves, 2015). Yet they too are constrained by whatever reality the training data represent — the models, after all, can only inhabit a reality described in that training data and no other. Despite the vast amount of data used for training OpenAI’s GPT models, the gargantuan effort to clean training data and make them less toxic (Perrigo, 2023) speak to the desperate lack of quality or representativeness of these data. Yes, these data are the largest and “the best we got” but there is a good reason why unvarnished and uncorrected models are not made available — the mirror they hold up to humanity is profoundly terrible (Finkelstein, 2008). The models that are made available for consumption are adjusted, cleaned up, made palatable and “value-aligned” resulting in output that might create an imaginary generic average person, but who that person is, is difficult to assess. While simulating human social systems with LLMs provides intriguing insights into the models themselves, what such output might reveal about us more generally, is a question that requires cautious consideration.

How might we “situate” the knowledge produced using LLM-generated data then? The papers we cite default to lists of prompts and details of the technical setup, but arguably also require deeper considerations of where the models are wrong and what is missing, tempering the excitement with the possibility of sweeping statements about “human” behavior. What counts as “data” influences how “data” are understood, collected and processed, and are intimately connected with establishing and validating the boundaries of “proper” knowledge production (Haraway, 1988; Kitchin & Lauriault, 2018). LLM-generated data require deep considerations of fidelity — both intersectional and otherwise (Johnson & Hajisharif, 2024) — as well as of positionality, paying attention to what kinds of knowledge are made possible with the use of these data and which are foreclosed.

7 Data Must Be Cooked with Care

The idea of a computational assistant that could precisely and flexibly analyze vast amounts of complex data with quasi-qualitative skills is undoubtedly tempting for researchers, who have often struggled to adapt their methods to the growing amount and complexity of the data at their disposal. LLMs show such remarkable analytical skills as a result of the unprecedentedly

large amount of data used in their training phase, but it is a leap to simply assume that these training data represent a viable proxy for the social reality behind the model. While still in its infancy, compared to other applications of LLMs, the idea of LLMs as data generators is intriguing, given the range of scholarly struggles with the complexities of data access. Some types of data may be more abundant, but they may not be easier to obtain or use.

Data is a complex and contested term, yet it has come to define the digital world we inhabit. Early debates around big data contested notions of raw data (Helmond, 2014), pointing out that data are never raw, out there, merely waiting to be collected (Bowker, 2008). Rather, data are always made, created, cooked as it were and if we are to acknowledge this, data ought to be cooked with care (Bowker, 2013). More importantly, when it comes to scientific practice, different epistemologies and methodologies “cook” data differently — seeing some methods of data generation as more legitimate than others. In many ways, synthetic data generation offers a way to create tidy, well-appointed datasets that are ultimately made specifically for this or that purpose, without the problems of cleaning messy data. Such control is one of the attractive qualities of synthetic data for many (Savage, 2023). What sort of cooking happens when generating data using LLMs? This is a much more difficult question to answer given myriad assumptions about training data, prompt processing, and emergent behaviors that must be made. Arguably, LLMs offer ease of generating data, but they provide far less control over the recipe, compared to any other approach to generating similar data from human participants.

In this essay we have looked at different examples of using LLMs for data production within the social sciences. We have discussed how LLM-generated data are similar to what we generally define as synthetic data but also where it differs. LLM-generated data aim to solve the problems of scarcity and privacy. LLMs’ ability to produce large amounts of seemingly realistic data, as well as their ability to role-play various demographics with a very tenuous identifiability with the underlying training data perfectly address these needs. When it comes to bias, studies proposing the use of LLM-generated data take a different approach compared to other types of synthetic data. Rather than seeing bias as a problem that should be measured, quantified, and potentially addressed, LLM-based approaches attempt to embrace it, leveraging it, either implicitly or explicitly, as algorithmic fidelity. This difference has interesting consequences and originates from the difference in goals. While most of the recent interest towards synthetic data is driven by the need to feed more and more high-quality data into AI models to improve their performance, LLM-generated data is the output of an AI model that could potentially be used directly for research or prototyping activities. This has potential but we argue that current approaches overlook a number of important issues.

8 The Future of Data?

As we have highlighted in the sections above, the idea of using LLM-generated data for research in the social sciences is relatively new and its robustness is still disputed. The inherent complexity of LLMs, as well as their fast-paced evolution suggest caution when researchers make assumptions about the models’ algorithmic fidelity or about their actual ability to be conditioned to represent various segments of the population. These are, after all, products not developed for research. Regardless of their commercial nature, which introduces additional complexity due to misaligned goals between tech companies and academic researchers, LLMs have not been developed as proxies of society. They have not been fed growing amounts of data aiming at improving their ability to represent a societal digital twin. Au contraire, many of the current trends (Saracco, 2023) that we see in the actual development of LLMs seem to suggest that the

future will not be more data and more algorithmic fidelity but smaller models, trained with smaller amounts of data and fewer parameters that will still be able to score similar results in reasoning tasks.

While today LLMs can surely prove useful to produce data for prototyping research or testing initial hypotheses, their reliability should constantly be questioned and confirmed. Of course, the results we have discussed here leave open the possibility that future LLMs could be specifically designed, trained and developed for research applications. This possibility has recently been proposed by Bail (2024). Imagining large-scale LLMs developed and dedicated to research is not simple and requires a substantial change in the way social scientists approach their research tools, but it could also open up unprecedented opportunities. As many authors have noticed, what is potentially revolutionary are LLM-based models trained on large amounts of data, rather than the specific commercial implementation. Commercial solutions, while currently more advanced than open-source alternatives, come with many of the problems and limitations that we have discussed above (from unknown safeguards to fine-tuning and opaque training data). Open-source LLMs are better for ethical reasons (Spirling, 2023) and they may, at least in theory, offer better transparency, and improved control and could be based on ad-hoc training data. Yet if research-oriented open-source LLMs might be the future of LLMs for social research, it is probably a good idea to reiterate some of the key challenges they will have to face: representativeness, segmentability and data curation.

Representativeness: As many authors who have proposed the use of LLMs for data generation argue, bias that derives from biased data should not be considered a problem but as a feature of the system. Given biased training data, from a research point of view, one might want that bias to be transferred to the model's outputs. The challenge is that this is not what has been observed. While LLMs seem to be able to faithfully reproduce biases and leaning at the level of large groups they systematically fail at representing smaller groups and minorities. Algorithmic fidelity, in other words, is not stable when the system is prompted to represent certain parts of the overall population. What is more, while some types of differences may be reproduced, there is always the danger of what Johnson & Hajisharif (2024) term "intersectional hallucination" where the inherent LLM bias and built-in attempts to mitigate it might result in strange demographic configurations.

Segmentability: Directly building on the assumption of bias as a way to faithfully represent the underlying data, there is the idea that LLMs can be segmented and conditioned to represent specific sub-populations. The extent to which this is true is still unclear. While prompting and fine tuning have shown some ability to condition the results, limits have also been observed as well as a considerable amount of inconsistency even with stable prompting.

Data curation: While LLMs require, by definition, a large amount of training data, complete lack of control over what constitutes training data is problematic both for ethical and legal reasons (Rahman & Santacana, 2023). With the progress shown by smaller models (Saracco, 2023), research LLMs should carefully consider to what extent curated training data is a possibility and what would be the consequences. Over the years researchers working with *hard-to-get* data have developed considerable experience with projects of data donation (Araujo et al., 2022). This experience could be leveraged to coordinate massive collaborative efforts that would select training data not because it is available or accessible but because it has been deemed relevant. The limits and consequences of such an approach are, clearly, unknown but the ethical and legal risks of the alternatives might end up being too large for non-profit research institutions.

As things are right now these problems have been scarcely investigated and solid ways to

measure them and their impact on the LLMs' ability to work as data-generation tools for social scientists have not been proposed. This should probably be a key part of any research agenda that leads to the actual development and deployment of LLMs as data generators for social sciences.

References

- Abowd, J.M., & Vilhuber, L. (2008). How Protective Are Synthetic Data? In J. Domingo-Ferrer & Y. Saygin (Eds.), *Privacy in Statistical Databases* (pp. 239–246). New York, NY: Springer. https://doi.org/10.1007/978-3-540-87471-3_20
- Agnew, W., Bergman, A.S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K.R. (2024). The Illusion of Artificial Inclusion. *arXiv*, 2401.08572. <https://doi.org/10.48550/arXiv.2401.08572>
- Aher, G.V., Arriaga, R.I., & Kalai, A.T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 337–371). <https://proceedings.mlr.press/v202/aher23a.html>
- Almeida, G.F., Nunes, J.L., Engelmann, N., Wiegmann, A., & de Araújo, M. (2024). Exploring the Psychology of LLMs' Moral and Legal Reasoning. *Artificial Intelligence*, 333, 104145. <https://doi.org/10.1016/j.artint.2024.104145>
- Araujo, T., Ausloos, J., van Atteveldt, W., Loecherbach, F., Moeller, J., Ohme, J., Trilling, D., van de Velde, B., de Vreese, C., & Welbers, K. (2022). OSD2F: An Open-source Data Donation Framework. *Computational Communication Research*, 4(2), 372–387. <https://doi.org/10.5117/CCR2022.2.001.ARAU>
- Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3), 337–351. <https://doi.org/10.1017/pan.2023.2>
- Atil, B., Chittams, A., Fu, L., Ture, F., Xu, L., & Baldwin, B. (2024). LLM Stability: A Detailed Analysis with Some Surprises. *arXiv*, 2408.04667. <https://doi.org/10.48550/arXiv.2408.04667>
- Bail, C.A. (2024). Can Generative AI Improve Social Science?. *Proceedings of the National Academy of Sciences, PNAS*, 121(21), e2314021121. <https://doi.org/10.1073/pnas.2314021121>
- Belgodere, B., Dognin, P., Ivankay, A., Melnyk, I., Mroueh, Y., Mojsilovic, A., Navratil, J., Nitsure, A., Padhi, I., Rigotti, M., Ross, J., Schiff, Y., Vedpathak, R., & Young, R.A. (2023). Auditing and Generating Synthetic Data with Controllable Trust Trade-offs. *arXiv*, 2304.10819. <https://doi.org/10.48550/arXiv.2304.10819>
- Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2023). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *SocArXiv*, May 4. <https://doi.org/10.31235/osf.io/5ecfa>

- Brand, J., Israeli, A., & Ngwe, D. (2023). Using LLMs for Market Research (Harvard Business School Marketing Unit Working Paper No. 23-062). *Social Science Research Network*. <https://doi.org/10.2139/ssrn.4395751>
- Breum, S.M., Egdal, D.V., Mortensen, V.G., Møller, A.G., & Aiello, L.M. (2023). The Persuasive Power of Large Language Models. *arXiv*, 2312.15523. <https://doi.org/10.48550/arXiv.2312.15523>
- Box, G.E.P. (1976). Science and Statistics. *Journal of the American Statistical Association*, 71(356), 791–799. <https://doi.org/10.1080/01621459.1976.10480949>
- Bowker, G.C. (2008). *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.
- Bowker, G.C. (2013). Data Flakes: An Afterword to “Raw Data” Is an Oxymoron. In L. Gitelman (Ed.), *“Raw Data” Is an Oxymoron* (pp. 167–172). Cambridge, MA: MIT Press.
- Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S.A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency*, PMLR, 81, 77–91. <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>
- Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, & L. Benotti (Eds.), *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP* (pp. 53–67). Singapore: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.c3nlp-1.7>
- Couldry, N., & Mejias, U.A. (2019). Data Colonialism: Rethinking Big Data’s Relation to the Contemporary Subject. *Television & New Media*, 20(4), 336–349. <https://doi.org/10.1177/1527476418796632>
- Cui, R., Lee, S., Hershcovich, D., & Søgaard, A. (2023). What Does the Failure to Reason with “Respectively” in Zero/Few-Shot Settings Tell Us about Language Models?. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Vol. 1* (pp. 8786–8800). Singapore: Association for Computational Linguistics.
- Demuro, E., & Gurney, L. (2024). Artificial Intelligence and the Ethnographic Encounter: Transhuman Language Ontologies, or What It Means “To Write like a Human, Think like a Machine”. *Language & Communication*, 96, 1–12. <https://doi.org/10.1016/j.langcom.2024.02.002>
- D’Ignazio, C., & Klein, L.F. (2023). *Data Feminism*. Cambridge, MA: MIT Press.
- Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI Language Models Replace Human Participants? *Trends in Cognitive Sciences*, 27(7), 597–600. <https://doi.org/10.1016/j.tics.2023.04.008>
- Eigenschink, P., Reutterer, T., Vamosi, S., Vamosi, R., Sun, C., & Kalcher, K. (2023). Deep Generative Models for Synthetic Data: A Survey. *IEEE Access*, 11, 47304–47320. <https://doi.org/10.1109/ACCESS.2023.3275134>

- Elsayed-Ali, S., Bonsignore, E., & Chan, J. (2023). Exploring Challenges to Inclusion in Participatory Design From the Perspectives of Global North Practitioners. In J. Nichols (Ed.), *Proceedings of the ACM on Human-Computer Interaction* (p. 7). New York, NY: Association for Computing Machinery <https://doi.org/10.1145/3579606>
- Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., & Zhao, X. (2024). Bias of AI-generated Content: An Examination of News Produced by Large Language Models. *Scientific Reports*, 14(1), 5224, 1–20. <https://doi.org/10.1038/s41598-024-55686-2>
- Figueira, A., & Vaz, B. (2022). Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics*, 10(15), 1–41. <https://doi.org/10.3390/math10152733>
- Finkelstein, S. (2008). Google, Links, and Popularity versus Authority. In J. Turow & L. Tsui (Eds.), *The Hyperlinked Society: Questioning Connections in the Digital Age* (pp. 104–120). Ann Arbor, MI: University of Michigan Press.
- Goodman, J.D., & Sandoval, E. (2024). Google Chatbot’s A.I. Images Put People of Color in Nazi-era Uniforms. *The New York Times*, 22 February. <https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html>
- Gordon, R. (2023). Large Language Models Are Biased. Can Logic Help Save Them? *MIT News*, 3 March. <https://news.mit.edu/2023/large-language-models-are-biased-can-logic-help-save-them-0303>
- Greenwood, J. (2018). How Would People Behave in Milgram’s Experiment Today. *Behavioral Scientist*, 24 July. <https://behavioralscientist.org/how-would-people-behave-in-milgrams-experiment-today>
- Grossmann, I., Feinberg, M., Parker, D.C., Christakis, N.A., Tetlock, P.E., & Cunningham, W.A. (2023). AI and the Transformation of Social Science Research. *Science*, 380(6650), 1108–1109. <https://doi.org/10.1126/science.ad11778>
- Groves, C. (2015). Logic of Choice or Logic of Care? Uncertainty, Technological Mediation and Responsible Innovation. *NanoEthics*, 9(3), 321–333. <https://doi.org/10.1007/s11569-015-0238-x>
- Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study. In A. Schmidt, K. Väänänen, T. Goyal, P.O. Kristensson, A. Peters, S. Mueller, J.R. Williamson, & M.L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–19). New York, NY: Association for Computing Machinery. <https://doi.org/10.1145/3544548.3580688>
- Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3), 575–599. <https://doi.org/10.2307/3178066>
- Helmond, A. (2014). “Raw Data” Is an Oxymoron. *Information, Communication & Society*, 17(9), 1171–1173. <https://doi.org/10.1080/1369118X.2014.920042>
- Henrich, J., Heine, S.J., & Norenzayan, A. (2010). The Weirdest People in the World? *Behavioral and Brain Sciences*, 33(2–3), 61–83. <https://doi.org/10.1017/S0140525X0999152X>

- Horton, J.J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? (Working Paper 31122). *National Bureau of Economic Research*. <https://doi.org/10.3386/w31122>
- Jacobsen, B.N. (2023). Machine Learning and the Politics of Synthetic Data. *Big Data & Society*, 10(1), 20539517221145372, 1–12. <https://doi.org/10.1177/20539517221145372>
- Jakesch, M., Hancock, J.T., & Naaman, M. (2023). Human Heuristics for AI-Generated Language Are Flawed. *Proceedings of the National Academy of Sciences, PNAS*, 120(11), e2208839120, 1–7. <https://doi.org/10.1073/pnas.220883912>
- Jansen, B.J., Jung, S., & Salminen, J. (2023). Employing Large Language Models in Survey Research. *Natural Language Processing Journal*, 4, 100020, 1–7. <https://doi.org/10.1016/j.nlp.2023.100020>
- Johnson, E., & Hajisharif, S. (2024). The Intersectional Hallucinations of Synthetic Data. *AI & Society*, 1–3. <https://doi.org/10.1007/s00146-024-02017-8>
- Jones, C.R., Trott, S., & Bergen, B. (2023). Epitome: Experimental Protocol Inventory for Theory of Mind Evaluation. *Proceedings of the First Workshop on Theory of Mind in Communicating Agents, PMLR*, 202. <https://openreview.net/pdf?id=e5Yky8Fvj>
- Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S.N., & Weller, A. (2022). Synthetic Data. What, Why and How? *arXiv*, 2205.03257. <https://doi.org/10.48550/arXiv.2205.03257>
- Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the Pirate Party Won the German Election of 2009 or the Trouble with Predictions: A Response to Tumasjan, A., Sprenger, T.O., Sander, P.G., & Welpe, I.M. “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment”. *Social Science Computer Review*, 30(2), 229–234. <https://doi.org/10.1177/0894439311404119>
- Kitchin, R., & Lauriault, T.P. (2018). Toward Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work. In J. Thatcher, J. Eckert & A. Shears (Eds.), *Thinking Big Data in Geography: New Regimes, New Research* (pp. 3–20). Lincoln, NE: University of Nebraska Press.
- Köbis, N., & Mossink, L.D. (2021). Artificial Intelligence versus Maya Angelou: Experimental Evidence that People Cannot Differentiate AI-generated from Human-written Poetry. *Computers in Human Behavior*, 114, 106553, 1–13. <https://doi.org/10.1016/j.chb.2020.106553>
- McKenna, N., Li, T., Cheng, L., Hosseini, M., Johnson, M., & Steedman, M. (2023). Sources of Hallucination by Large Language Models on Inference Tasks. *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 2758–2774). Singapore: Association for Computational Linguistics <https://doi.org/10.18653/v1/2023.findings-emnlp.182>
- Moats, D., & Seaver, N. (2019). “You Social Scientists Love Mind Games”: Experimenting in the “Divide” between Data Science and Critical Algorithm Studies. *Big Data & Society*, 6(1), 2053951719833404, 1–11. <https://doi.org/10.1177/2053951719833404>
- Møller, A.G., Pera, A., Dalsgaard, J., & Aiello, L. (2024). The Parrot Dilemma: Human-labeled vs. LLM-augmented Data in Classification Tasks. In Y. Graham, & M. Purver (Eds.), *Pro-*

- ceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 179–192). Singapore: Association for Computational Linguistics. <https://aclanthology.org/2024.eacl-short.17/>
- Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More Human than Human: Measuring ChatGPT Political Bias. *Public Choice*, 198(1–2), 3–23. <https://doi.org/10.1007/s11127-023-01097-2>
- Nikolenko, S.I. (2021). *Synthetic Data for Deep Learning*. Cham: Springer International Publishing.
- Park, J.S., Popowski, L., Cai, C., Morris, M.R., Liang, P., & Bernstein, M.S. (2022). Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In M. Agrawala, J.O. Wobbrock, E. Adar, & V. Setlur (Eds.), *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–18). New York, NY: Association for Computing Machinery. <https://doi.org/10.1145/3526113.3545616>
- Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., & Bernstein, M.S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. In S. Follmer, J. Han, J. Steimle, & N. Henry Riche (Eds.), *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–22). New York, NY: Association for Computing Machinery. <https://doi.org/10.1145/3586183.3606763>
- Perrigo, B. (2023). OpenAI Used Kenyan Workers on Less Than \$2 Per Hour: Exclusive. *Time*, 18 January. <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Phelps, S., & Russell, Y.I. (2023). Investigating Emergent Goal-Like Behaviour in Large Language Models Using Experimental Economics. *arXiv*, 2305.07970. <https://doi.org/10.48550/arXiv.2305.07970>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models Are Unsupervised Multitask Learners. *OpenAI*, 1(8), 1–24. <https://api.semanticscholar.org/CorpusID:160025533>
- Raghunathan, T.E. (2021). Synthetic Data. *Annual Review of Statistics and Its Application*, 8(1), 129–140. <https://doi.org/10.1146/annurev-statistics-040720-031848>
- Rahman, N., & Santacana, E. (2023). Beyond Fair Use: Legal Risk Evaluation for Training LLMs on Copyrighted Text. *Proceedings of the 40th International Conference on Machine Learning* (pp. 1–5). <https://blog.genlaw.org/CameraReady/57.pdf>
- Rhee, J., (2018). *The Robotic Imaginary: The Human and the Price of Dehumanized Labor*. Minneapolis, MN: University of Minnesota Press.
- Rossetti, G., Stella, M., Cazabet, R., Abramski, K., Cau, E., Citraro, S., Failla, A., Improta, R., Morini, V., & Pansanella, V. (2024). Y Social: An LLM-powered Social Media Digital Twin. *arXiv*, 2408.00818. <https://doi.org/10.48550/arXiv.2408.00818>
- Salminen, J., Guan, K.W., Jung, S.G., & Jansen, B. (2022). Use Cases for Design Personas: A Systematic Review and New Frontiers. In S. Barbosa, C. Lampe, C. Appert, D.A. Shamma, S. Drucker, J. Williamson, & K. Yatani (Eds.), *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–21). New York, NY: Association for Computing Machinery. <https://doi.org/10.1145/3491102.3517589>

- Samuelson, W., & Zeckhauser, R. (1988). Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty*, 1, 7–59. <https://doi.org/10.1007/BF00055564>
- Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning*, PMLR, 202, 29971–30004. <https://proceedings.mlr.press/v202/santurkar23a.html>
- Saracco, R. (2023). How Much Bigger Can/Should LLMs Become?. *IEEE Future Directions*, April 24. Retrieved from <https://cmte.ieee.org/futuredirections/2023/04/24/how-much-bigger-can-should-llms-become/>
- Savage, N. (2023). Synthetic Data Could Be Better than Real Data. *Nature*, d41586-023-01445-01448, 27 April. <https://www.nature.com/articles/d41586-023-01445-8>
- Schaeffer, R., Miranda, B., & Koyejo, S. (2024). Are Emergent Abilities of Large Language Models a Mirage? *Proceedings of the 37th International Conference on Neural Information Processing Systems* (pp. 55565–55581). <https://doi.org/10.48550/arXiv.2304.15004>
- Simmons, G., & Savinov, V. (2024). Assessing Generalization for Subpopulation Representative Modeling via In-Context Learning. *arXiv*, 2402.07368. <https://doi.org/10.48550/ARXIV.2402.07368>
- Sin, J., Franz, R.L., Munteanu, C., & Barbosa Neves, B. (2021). Digital Design Marginalization: New Perspectives on Designing Inclusive Interfaces. In Y. Kitamura & A. Quigley (Eds.), *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). <https://doi.org/10.1145/3411764.3445180>
- Spirling, A. (2023). Why Open-source Generative AI Models Are an Ethical Way Forward for Science. *Nature*, 616((7957)), 413–413. <https://doi.org/10.1038/d41586-023-01295-4>
- Sun, H. (2023). Reinforcement Learning in the Era of LLMs: What Is Essential? What Is Needed? An RL Perspective on RLHF, Prompting, and Beyond. *arXiv*, 2310.06147. <https://doi.org/10.48550/arXiv.2310.06147>
- Tjautja, L., Chen, V., Wu, S.T., Talwalkar, A., & Neubig, G. (2023). Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *arXiv*, 2311.04076. <https://doi.org/10.48550/ARXIV.2311.04076>
- Törnberg, P., Valeeva, D., Uitermark, J., & Bail, C. (2023). Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms. *arXiv*, 2310.05984. <https://doi.org/10.48550/arXiv.2310.05984>
- Törnberg, P. (2024). Best Practices for Text Annotation with Large Language Models. *Sociologica*, 18(2), 67–85. <https://doi.org/10.6092/issn.1971-8853/19461>
- van der Schaar, M., & Qian, Z. (2023). *AAAI Lab for Innovative Uses of Synthetic Data*. Association for the Advancement of Artificial Intelligence. https://www.vanderschaar-lab.com/wp-content/uploads/2022/08/AAAI_Synthetic-Data-Tutorial.pdf

- von der Heyde, L., Haensch, A.-C., & Wenz, A. (2023). Assessing Bias in LLM-Generated Synthetic Datasets: The Case of German Voter Behavior. *SocArXiv*. <https://EconPapers.epec.org/RePEc:osf:socarx:9718s>
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *arXiv*, 2206.07682. <https://doi.org/10.48550/arXiv.2206.07682>
- Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L.A., Rimell, L., Isaac, W., ... Gabriel, I. (2022). Taxonomy of Risks Posed by Language Models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). <https://doi.org/10.1145/3531146.3533088>
- Whitney, C.D., & Norman, J. (2024). Real Risks of Fake Data: Synthetic Data, Diversity-washing and Consent Circumvention. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1733–1744). New York, NY: Association for Computing Machinery. <https://doi.org/10.1145/3630106.3659002>
- Wu, Z., Peng, R., Han, X., Zheng, S., Zhang, Y., & Xiao, C. (2023). Smart Agent-Based Modeling: On the Use of Large Language Models in Computer Simulations (arXiv: 2311.06330). *arXiv*. <https://doi.org/10.48550/arXiv.2311.06330>
- Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., & Cheng, X. (2024). On Protecting the Data Privacy of Large Language Models (LLMs): A Survey. *arXiv*, 2403.05156. <https://doi.org/10.48550/ARXIV.2403.05156>

Luca Rossi – NERDS research group, Department of Digital Design, IT University of Copenhagen (Denmark)

ORCID: <https://orcid.org/0000-0002-3629-2039> | ✉ lucr@itu.dk

🔗 <https://pure.itu.dk/da/persons/luca-rossi>

Luca Rossi is an Associate Professor of Digital Media and Networks at the Department of Digital Design of IT, University of Copenhagen (Denmark). He coordinates the Human Centered Data Science research group, and he is member of the Networks Data and Society (NERDS) research group. He teaches Network analysis and Digital Data Analysis.

Katherine Harrison – Department of Thematic Studies – Gender Studies, Linköping University (Sweden)

ORCID: <https://orcid.org/0000-0002-8325-4051>

🔗 <https://liu.se/en/employee/katha38>

Katherine Harrison, Ph.D., is an Associate Professor in Gender Studies at Linköping University (Sweden). Her research sits at the intersection of Science & Technology Studies, media studies, and feminist theory, bringing critical perspectives on knowledge production to studies of different digital technologies.

Irina Shklovski – Department of Computer Science, Department of Communication, University of Copenhagen (Denmark); Department of Thematic Studies – Gender Studies, Linköping University (Sweden)

ORCID: <https://orcid.org/0000-0003-1874-0958>

🔗 <https://researchprofiles.ku.dk/en/persons/irina-shklovski>

Irina Shklovski is a Professor of Communication and Computing in the Department of Computer Science and the Department of Communication at the University of Copenhagen (Denmark). She holds a WASP-HS visiting professorship at Linköping University (Sweden). Her research areas include speculative AI futures, AI ethics, data quality, synthetic data, explainability, privacy, and creepy technologies.