Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline

Shivani Kapania*
Carnegie Mellon University
Pittsburgh, USA
skapania@andrew.cmu.edu

Stephanie Ballard
Microsoft
Redmond, USA
stephballard@microsoft.com

Alex Kessler
Microsoft
Redmond, USA
alexkessler@microsoft.com

Jennifer Wortman Vaughan
Microsoft
New York City, USA
jenn@microsoft.com

Abstract

Alongside the growth of generative AI, we are witnessing a surge in the use of synthetic data across all stages of the AI development pipeline. It is now common practice for researchers and practitioners to use one large generative model (which we refer to as an *auxiliary model*) to generate synthetic data that is used to train or evaluate another, reconfiguring AI workflows and reshaping the very nature of data. While scholars have raised concerns over the risks of synthetic data, policy guidance and best practices for its responsible use have not kept up with these rapidly evolving industry trends, in part because we lack a clear picture of current practices and challenges. Our work aims to address this gap. Through 29 interviews with AI practitioners and responsible AI experts, we examine the expanding role of synthetic data in AI development. Our findings reveal how auxiliary models are now widely used across the AI development pipeline. Practitioners describe synthetic data as crucial for addressing data scarcity and providing a competitive edge, noting that evaluation of generative AI systems at scale would be infeasible without auxiliary models. However, they face challenges controlling the outputs of auxiliary models, generating data that accurately depict underrepresented groups, and scaling data validation practices that are based primarily on manual inspection. We detail general limitations of and ethical considerations for synthetic data and conclude with a proposal of concrete steps towards the development of best practices for its responsible use.

ACM Reference Format:
Shivani Kapania, Stephanie Ballard, Alex Kessler, and Jennifer Wortman Vaughan. 2025. Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3715275.3732005

*This work was done when the author was an intern at Microsoft Research.

FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732005

1 Introduction

Data is a critical building block of AI systems [33]. As the tech industry shifts attention and resources towards building large-scale, data-hungry generative models, traditional data sources are being exhausted, giving rise to an intense search for more data [64, 81]. However, developing high-quality datasets remains complex and resource-intensive, with substantial costs, time, and expertise required for data collection, annotation, and curation [83]. Increasingly, *synthetic data*—artificial data generated using models or algorithmic techniques [76]—is seen as an appealing alternative [69]. OpenAI [3], Apple [32, 44], Microsoft [2], Google [91], Meta [30], and IBM [43] have all reported using synthetic data in their AI development pipelines or advertised the ability of their models to create synthetic data.

Synthetic data is not new. Explorations into data simulation to address data scarcity date back at least to the 1970s [69]. The idea began to gain momentum with the introduction of practical methods for data generation, like generative adversarial networks [28] and variational autoencoders [51]. Throughout much of the past decade, the use of synthetic data was viewed as a way to mitigate concerns about fairness, bias, and privacy. Proponents argued that privacy concerns could be addressed by replacing traditional datasets with differentially private, synthetic alternatives [1, 16, 49, 105] and that model biases could be mitigated by augmenting traditional datasets with simulations of data from underrepresented groups or rare scenarios [46, 85]. The latter approach was especially popular for applications like facial recognition, where collecting diverse data can be prohibitively challenging [7, 53]. Practitioners have since applied synthetic data in applications within the robotics [42, 56], automotive [106], finance [5], and healthcare [63, 71] industries, among others.

Now, the widespread availability of large, off-the-shelf generative models has given practitioners the ability to automate the creation of synthetic data without requiring extensive domain-specific expertise or custom-built tools. This has made synthetic data more accessible and rapidly expanded the scope of its use in AI development [48, 58]. It is increasingly common for researchers and practitioners to use one large generative model (which we refer to as an *auxiliary model* throughout this paper) to generate synthetic data that is used to train or evaluate another (the *primary model*). At the same time, the nature of data is evolving. Auxiliary models

are used not only to expand training datasets [60], create benchmarks [36], and develop test cases [23, 77], but also to generate scores or labels for model outputs [20, 110], a task previously performed by human annotators or with simple automated approaches. And since auxiliary models can dynamically generate responses to a primary model's outputs in real-time, synthetic data no longer needs to be static but can be generated interactively as needed.

Despite concerns that have been raised over the risks of synthetic data [4, 96], policy guidance and best practices for its responsible use have not kept pace with shifting trends. Much of the policy guidance is still centered on privacy, with more widespread applications of model-generated data just starting to receive attention. For instance, Singapore's Proposed Guide to Synthetic Data, released in July 2024, focuses on the benefits and risks of synthetic data as a privacy-enhancing technology [76]. The EU's General-Purpose AI Code of Practice touches only briefly on synthetic data, with a recommendation added in the November 2024 draft to document "a description of the methods, if any, used to synthetically generate training data" and "the name(s) of any AI model(s) or system(s) used to synthetically generate training data" [70].

One challenge in creating comprehensive guidance on synthetic data is a lack of understanding of practitioners' current workflows. With the landscape rapidly evolving and the use cases so varied, we do not have a clear picture of the full spectrum of practitioners' motivations for using synthetic data, the properties they hope it will satisfy, or how (or even if) they validate the data they create. To address this gap, we explore the following research questions:

- **RQ1:** What are practitioners' motivations for using synthetic data in the AI development pipeline?
- **RQ2:** What are practitioners' current practices, desiderata, and challenges when generating synthetic data?
- **RQ3:** What are practitioners' current practices and challenges when validating synthetic data?
- **RQ4:** What limitations and ethical considerations arise with the use of synthetic data?

We draw on a two-phase interview study. We first conducted semi-structured interviews with 19 practitioners actively engaged in the development of generative AI models or systems to understand how they use synthetic data across the development pipeline. To further explore the limitations and ethical considerations of the practices identified, we conducted an additional 10 interviews with subject matter experts experienced in responsible AI aspects of dataset creation or model evaluation, scaffolding the discussion with three vignettes based on use cases observed in phase 1.

While synthetic data is a broad concept, in order to focus on new and evolving practices, we scope our study to the use of data (including labels and scores of outputs) produced by auxiliary generative models for use in the AI development pipeline. While we use terms like 'real' or 'human-generated' data for distinction throughout this paper, we acknowledge that synthetic data is also shaped by human design and intervention [45, 54] and that, in practice, there is a spectrum between 'real' and 'synthetic' data.

Our interviews reveal how auxiliary models are now embedded in nearly all stages of the AI development pipeline, with synthetic data increasingly substituted for traditional human-generated data and labels. Practitioners perceive synthetic data as a promising and indispensable tool to address data scarcity challenges, resource constraints, and high data collection costs across a range of critical tasks, including evaluating harms, diversifying datasets to reduce bias, preserving privacy when working with sensitive data, and scoring model outputs. However, practitioners faced challenges in understanding and controlling the outputs of auxiliary models and generating data that accurately depicts members of underrepresented groups. Although practitioners acknowledged the importance of data validation, they struggled to define what makes synthetic data 'good' and reported that most validation currently takes the form of 'spot-checking' or 'eyeballing.' Responsible AI experts noted additional limitations, including risks of stereotyping and the removal of avenues for data subjects to exercise agency over their data. We reflect on these findings and their implications for the evolution of data practices in section 5 and conclude by highlighting opportunities for future FAccT research to support the responsible use of synthetic data.

Our paper makes three main contributions:

- We provide an empirical account of the ubiquitous and inconsistent integration of synthetic data into modern AI development pipelines.
- We highlight limitations and ethical considerations of synthetic data use and propose avenues to study its impact.
- We provide considerations for the responsible use of synthetic data aimed at both practitioners and policymakers.

2 Related Work

In this section, we provide an overview of the uses of synthetic data in modern AI development pipelines, draw attention to critiques of synthetic data, and engage with scholarship examining industry practices around AI development and data production.

2.1 Synthetic Data Usage in Modern AI Development

Advances in generative AI have led to a rapid expansion of the use of synthetic data. We briefly review use cases from the literature and direct readers to the surveys of Jordon et al. [48] and Liu et al. [58] for a comprehensive overview.

Within the training stage of the AI pipeline, synthetic data is used to improve generative model capabilities. For instance, researchers have designed synthetic datasets geared to teaching models mathematical reasoning [60, 108] and computer programming [41]. Synthetic language-image datasets have been designed to improve the performance of multi-modal applications [57], while datasets of synthetic multilingual question-answer pairs have been used to improve cross-lingual performance [80, 84]. Synthetic data is also commonly used for knowledge distillation, which aims to transfer knowledge from a large model to a smaller one [29, 39]. Across these varied use cases, synthetic data is positioned as a way to augment real-world datasets and introduce controlled variability [58].

Within the evaluation stage, researchers have generated synthetic test cases to evaluate models for characteristics like factual consistency [23] and harmful behavior [77], and have published synthetic benchmark datasets, for instance, for detecting harmful output [36]. Synthetic test cases can take the form of static model prompts or, since auxiliary models can generate responses on the fly,

interactive simulations of a user's interactions with a model or system. Beyond creating test cases, auxiliary models are increasingly used to evaluate or score outputs from a primary model [20, 26], an idea colloquially known as 'LLM-as-a-judge' [13]. We view this as a form of synthetic data as the scores produced by the auxiliary model take the place of labels that often would have been collected from human evaluators.

2.2 Critical Studies of Synthetic Data

Scholars have challenged the assumption that data offers an unmediated reflection of reality, revealing instead its deeply constructed nature. As Gitelman [27] argued, data is never 'raw.'

The abstraction introduced by synthetic data creates further distance from the material realities seemingly represented in datasets, raising the question of whether and to what extent synthetic data is a reflection of reality.

Researchers have invited a critical examination of how synthetic data may introduce new risks and exacerbate existing ones in ways that may evade scrutiny [45]. Susser and Seeman [90] caution against positioning synthetic data as free from ethical scrutiny while enabling its role in accelerating unchecked AI development. Helm et al. [38] argue that synthetic data operates as a discursive device, legitimizing a shift from data collection to data generation without addressing the ethical and epistemological harms intrinsic to the models that produce this data. Synthetic data's ostensible detachment from real-world individuals can mitigate privacy concerns, but it simultaneously forecloses meaningful engagement with the people and communities most affected by AI systems [89]. By severing relationships between data subjects and the systems that act upon them, synthetic data erodes pathways for accountability, effectively shielding its use from critical oversight [24] and adding layers of opacity that obscure how synthetic data influences downstream decision-making. Furthermore, the promise of synthetic data as a fix-all solution can undermine democratic approaches to data governance, silencing opportunities for public participation and reflexive deliberation [96]. To interrogate these risks and weigh them against potential benefits, there is a need to examine both the conditions of data production and the auxiliary models that are used to generate this data [101].

2.3 AI Development Practices

The logics, methods, and techniques that define the field of AI are inseparable from its values, social norms, and organizational imperatives [21]. Decisions made throughout the AI development pipeline are embedded in broader social, cultural, and organizational contexts [93]. To identify points or intervention or improvements, it is crucial to understand the context and challenges of AI development. A growing line of research within the FAccT, CSCW, and broader HCI communities aims to shed light on the practices, challenges, and needs of practitioners developing AI systems. This line of work has explored practices around responsible technology development [18, 19, 40, 61, 62, 75, 79, 100, 107] and examined data practices specifically [15, 34, 37, 50, 67, 82, 103, 109], exposing the challenges faced by practitioners throughout data curation [34], exploratory data analysis [103], annotator selection [50], and data documentation [37]. The role of organizational factors has been a common theme [19, 61, 62, 97]. Researchers have emphasized the discretionary choices that shape data work, such as how tasks are formulated, what data is collected and annotated, how data quality is measured, which errors are acceptable, and what is communicated to stakeholders [74]. Muller et al. [67] examine how trade-offs and compromises are made when intervening with data, while Sambasivan et al. [83] report how contextual factors and incentives for prioritizing model development over careful data development result in downstream harms [52].

In concurrent research, Qian et al. [78] studied practices within Google around the adoption of large language models for data curation. They find that data workers emphasize the importance of high-quality data, but have trouble defining data quality for generative outputs without clear ground-truth references. They describe an emerging dataset hierarchy where the practice is shifting from the usual human-labeled 'golden datasets' to model-generated 'silver datasets' and 'super-golden datasets' that are created and rigorously evaluated by product managers, policy makers, engineers, and other experts. Our study complements this work by exploring a wider range of use cases for synthetic data across multiple organizations to understand practitioners' motivations for using synthetic data; their current practices, desiderata, and challenges when generating and evaluating synthetic data; and limitations and ethical considerations that arise.

3 Methods

To investigate our research questions, we developed a two-phase interview study, conducting semi-structured interviews with a total of 29 participants from 14 organizations across the United States. The first phase focused on understanding AI practitioners' dataset, model, and system development and evaluation practices for generative AI. We asked participants to describe their AI pipeline for a particular project, including the types and sources of data they used. We then asked specifically about their use of synthetic data and labels, investigating their motivations, how they generate the data, the validation processes they follow, the challenges they encounter, and their perceptions of any limitations.

To further explore the limitations and ethical considerations of the practices identified as well as paths forward, the second phase of interviews focused on understanding responsible AI (RAI) experts' perspectives on the sociotechnical dimensions of synthetic data use. After asking about participants' areas of focus within RAI, we asked them to reflect on the ways synthetic data intersects with privacy, fairness, consent, and other RAI concerns, discussing whether synthetic data could help mitigate existing challenges and whether it introduces new risks and trade-offs. To scaffold our discussions, we distilled the use cases of synthetic data mentioned in phase 1 interviews into three hypothetical scenarios: using synthetic data to fine-tune models, create evaluation datasets, or simulate user interactions. For each scenario, participants reflected on the practical and ethical implications of the practice, including any concerns and trade-offs involved in adopting synthetic data over traditional methods. Refer to Appendix A for study materials, including the vignettes.

All interviews were conducted over video conferencing software in English between May and August 2024. Each session lasted 40–60

minutes, and participants were compensated with a gift card of $75 USD. We recorded field notes and video recordings, which were transcribed for analysis. See Section 7 for a discussion of research ethics considerations.

Recruitment. We recruited participants until we reached saturation through advertisements on social networks such as X (formerly Twitter) and LinkedIn, direct emails to contacts in our professional networks, and snowball sampling. For phase 1, we recruited 19 participants from industry and academia who were actively engaged in the development of generative AI models or products. We refer to these participants as *practitioners*. Most practitioners focused on product development and research, in roles such as applied scientist, software engineer, linguist, or manager. For phase 2, we recruited 10 participants who were focused on RAI research or development and had experience in RAI aspects of dataset creation or model evaluation, whom we refer to in this paper as *RAI experts*. These participants worked in technology companies, non-profits, and academia. Table 1 includes more details on participants.

Analysis. We followed a reflexive thematic analysis approach inspired by Braun and Clarke [10, 11]. Our analysis was an inductive and iterative process guided by our research questions. The first author read each interview transcript multiple times to get familiar with the data, then extracted codes that aligned with the concepts in our research questions. The entire research team met regularly to discuss ambiguities and to define themes based on our initial codes. We surfaced 762 first-level codes. As we generated themes from the codes, we also identified categories with a description and examples of each category. In the early stages of our codes-to-themes process, we generated nine domain categories (stages of use, motivations, objectives, synthetic data generation techniques, validating synthetic data, challenges, limitations, trade-offs, and assumptions). These categories were also iteratively refined through meeting, diverging, and synthesizing into four top-level categories organized by our research questions and presented in our findings.

Limitations. Because of the sensitive nature of our study and the current competitive environment in AI, recruiting practitioners to discuss their data practices was challenging. We were limited in the number of participants we could recruit, the breadth of organizations they represented, and the global diversity of those participants. In particular, all practitioners were based in the United States. Additionally, most practitioners in our study focused on language-based technologies and, therefore, text data. The three exceptions to this focused on multimodal data (text and images), image data, and biological sequence data, respectively. These factors potentially skewed the data practices we observed and challenges highlighted in our findings. Additionally, while part of our original motivation for recruiting RAI experts was to aggregate guidance for synthetic data use, we found that concrete best practices have yet to emerge. We therefore pivoted this discussion to steps the community can take to develop best practices in light of our findings.

4 Findings

We first examine how practitioners use auxiliary generative models to produce synthetic data and their motivations for doing so (RQ1). We next describe their current practices, desiderata, and challenges when generating (RQ2) and validating (RQ3) synthetic data. Finally, we discuss general limitations of and ethical considerations for synthetic data (RQ4).

4.1 The Expanding Role of Synthetic Data

Our interviews reveal a reconfiguration of AI development workflows, with auxiliary models now embedded in nearly all stages of the AI development pipeline and synthetic data increasingly substituted for traditional, human-generated data and labels. Practitioners articulated a wide range of motivations for incorporating synthetic data, including its potential ability to scale, perceived performance improvements, and increased controllability.

4.1.1 Use Cases Across the AI Pipeline. Practitioners described a breadth of use cases for synthetic data, closely aligned with those proposed in the literature (Section 2.1). Within the training stage of the pipeline, practitioners frequently used auxiliary models to produce large datasets or augment existing datasets for pre-training new primary generative models, particularly in domains where data scarcity or sensitivity limited access to training datasets. Participants explained how synthetic data enabled them to create specialized corpora, such as synthetic legal documents or medical records, which would otherwise remain inaccessible due to privacy and/or compliance requirements. Synthetic data was also used to fine-tune existing primary models for specialized tasks. P15, for example, generated instruction-following datasets to align model outputs with safety requirements, while P12 used a larger auxiliary model as an 'oracle' to perform knowledge distillation [cf. 29] to train a smaller, more efficient model.

In the evaluation stage, practitioners turned to synthetic data to assess both the performance and safety of the primary model. Participants described how auxiliary models were used to generate diverse test cases to evaluate the primary model's ability to handle a range of input types, including cases that might not be well-represented in real-world datasets. For emerging tasks that lacked established datasets or evaluation metrics, synthetic data promised a solution for creating custom benchmarks. Sometimes, synthetic evaluation data consisted of fixed queries, while sometimes, practitioners used auxiliary models for simulating user interactions dynamically, such as sequences of queries or dialogues. Such simulated interactions were integrated into automated red-teaming efforts in which auxiliary models generated adversarial inputs designed to test the primary model's adherence to specific safety principles.

Beyond generating inputs for evaluation, auxiliary models were also increasingly used to evaluate the quality of the primary model's outputs by producing ratings or scores (as in LLM-as-a-judge [110]), a task that was previously typically carried out by human annotators or more conventional automated metrics (e.g., [22, 73]). For example, P11 was using an auxiliary model for red-teaming purposes: specifically, to evaluate whether the primary model was producing harmful outputs. They explained how the model *"output obviously needs to be categorized, and that is done using [auxiliary] models. So we have kind of like scorers [...] across different harm categories that sort the outputs, which I think could count as synthetic data insofar as it's making a judgment on the output of the [primary model]."* In some cases, the scoring prompts and evaluation rubrics themselves were developed using an auxiliary model. Initially, several participants

Practitioners (n = 19)		RAI Experts (n = 10)	
Industry (n = *15*)	Researcher (n = 7)	Industry (n = *3*)	Researcher (n = 4)
Academia (n = *4*)	Applied Scientist (n = 4)	Academia (n = *6*)	Professor (n = 3)
	ML Scientist (n = 2)	Non-profit (n = *1*)	Data Operations Lead (n = 1)
	Product Manager (n = 2)		Policy Researcher (n = 1)
	Linguist (n = 1)		Program Manager (n = 1)
	Red-teaming Lead (n = 1)		
	Research Manager (n = 1)		
	Software Engineer (n = 1)		

Table 1: Summary of participants' roles. Participants spanned a total of 14 organizations across the United States. In reporting our findings, we use identifiers beginning with 'P' for practitioners and 'E' for responsible AI experts.

did not consider their use of auxiliary models for scoring model outputs as a form of synthetic data (or 'data' at all), but on further reflection, they came to identify these labels as 'synthetic evaluation data.'

4.1.2 Perceived Promises of Synthetic Data. Practitioners brought up a variety of reasons why they expected synthetic data to be beneficial, which we describe here. However, as we explore in Sections 4.2.3 and 4.3.2 when reflecting on the challenges of generating and validating synthetic data, these promised benefits were not always realized in practice.

In describing motivations for their use of synthetic data, practitioners emphasized that auxiliary models enabled them to overcome data and resource constraints and thus significantly accelerate their development and evaluation processes.

The relatively lower cost and time requirements of using synthetic data emerged as key factors driving its appeal. For example, P15 described how synthetic data allowed them to produce large volumes of expert-level math or science training data without needing to recruit mathematicians or scientists. Participants emphasized that model training is better resourced than model evaluation, making synthetic data an especially attractive option for evaluation purposes. P3, who specialized in developing an automated red-teaming pipeline, provided a compelling comparison: Manual red-teaming often required hiring cybersecurity experts, each spending up to forty-five minutes interacting with the system to generate a single sample. In contrast, automated red-teaming incurred substantially lower costs—primarily initial setup and model inference expenses. The cost differential made a "*cheap and dirty*" automated approach far more attractive.

The use of human annotators, meanwhile, was limited by the growing needs across teams, especially for evaluation. P5 and P11 noted that tight timelines and organizational demands required generating thousands of prompts and testing models within just a few weeks, with dozens of teams in P5's organization conducting weekly model evaluations. They noted that scoring large volumes of model outputs weekly was infeasible with human reviewers.

Participants also embraced synthetic data for its potential to boost performance and provide a competitive edge. P17's decision to adopt synthetic data was inspired by another team within their company that had successfully used synthetic data for evaluation. As they explained: "*We had a pretty good idea going into our own testing that this type of prompt was appropriate, or was going to be fairly accurate, hopefully, if we did things correctly.*" The anticipated performance gains also helped some participants justify its use to upper management. Several participants emphasized the competitive advantage of synthetic data offered in the tech industry, particularly for pre-training new, primary foundation models. P18 underscored how synthetic data would facilitate differentiation in model development, even if the approaches for generating it were publicly known: "*Even if I tell you exactly how I generated my synthetic data, I will bet all my money that there's no way you can generate the same thing... we're not going to have the same model. And so it's an easy way to differentiate between companies and models, especially when we all start with the same boat, which is web data.*"

Practitioners were particularly drawn to the parameterized and (ostensibly) controlled nature of synthetic data generation, which was perceived as a stark contrast to the unpredictability of real-world data collection and labeling processes. For example, P1, who was developing a benchmark for hate speech and toxicity detection, noted that publicly available datasets frequently exhibited skewed representations, favoring particular identity groups or containing disproportionately high levels of toxic content compared to non-toxic examples. These imbalances limited practitioners' ability to effectively evaluate their primary models. Synthetic data, then, provided a promising avenue to address these shortcomings by enabling more 'controlled' data generation. As P18 put it, with synthetic data, "*you have full control of what you're generating.*" However, as we describe in section 4.2, 'full control' was not always achievable in practice.

Practitioners also viewed synthetic data as a way to simplify compliance requirements related to data protection. P15 explained how generating synthetic data allowed them to avoid the "*lengthier processes*" involved in securing data rights. Similarly, P4 noted that "*feedback from real customers has to be deleted after 30 days and is subject to GDPR, data subject requests, and things like that, but the synthetic version does not have those same encumbrances.*" While highlighting this as an advantage, practitioners acknowledged the "*gray area*" with this approach, noting that *"legal issues"* with regards to synthetic datasets are far from resolved.

4.2 Generating Synthetic Data

We next explore current practices, desiderata, and challenges when *generating* synthetic data (RQ2).

4.2.1 Desiderata for Synthetic Data. The two desiderata most commonly identified by practitioners were diversity and resemblance to 'natural' or 'real' data. Diversity was often described in terms of capturing a range of scenarios, input types, or demographic characteristics, which participants considered important to enable primary models to generalize effectively across varied real-world contexts. Natural-looking data was described as closely resembling human-generated content, but without replicating the training data of the

auxiliary model. For instance, in the context of generating data for automated red-teaming, P3 explained that the average user would not make an explicitly harmful query like "*how to make a bomb*." Instead, more frequent "*naturally occurring queries*" might trigger more subtle forms of harm relating to, for instance, interpersonal bias in the workplace.

4.2.2 Approaches to Data Generation. Practitioners chose different approaches to synthetic data generation, with distinct trade-offs between control and flexibility. Some preferred a constrained approach that limited the scope of generation to maintain more oversight over outputs. Most commonly, this involved having the auxiliary model generate variations (*e.g.,* paraphrases) of manually curated seed examples. P13, for example, generated synonyms for a list of successful red-teaming prompts, arguing that this structured approach supported the rigor required for controlled experimentation by allowing them to monitor how language shifts impacted the primary model's responses. Another common constrained approach was to design structured templates that dictate the format the auxiliary model's output should take, for instance, steering the model output toward specific topics or demographic groups.

Other practitioners were willing to sacrifice some control for the flexibility afforded by more open-ended approaches. One such approach was to generate synthetic data by having the auxiliary model take on a simulated persona reflecting specific demographics or scenarios. Practitioners reasoned that this approach allowed for exploration across a broader set of possible outputs, offering more diversity and resemblance to natural data. While there was a risk of introducing lower-quality irrelevant data, practitioners adopting open-ended approaches would iteratively rewrite the prompt for the auxiliary model if the generated data did not match their expectations.

Different techniques were needed to generate the (synthetic) scores or ratings used in evaluation. The most common approach was to specify annotation guidelines for the auxiliary model to follow, similar to those that would be given to human raters. P10 believed this approach is slowly emerging as the gold standard for evaluating model-generated summaries: "*It's not perfect, but for better or worse, it's the [method] most calibrated with human judgments that exists. You give instructions on how GPT-4 should evaluate the summary, present the source document and the generated summary, and then GPT-4 assigns a score from 1 to 7 based on the rubric provided.*"

Participants described an ad-hoc approach to selecting which auxiliary model to use for data generation, with choices driven by immediate needs and availability rather than systematic criteria. Many relied on models readily accessible within their organizations or those considered "*state-of-the-art.*" P9 explained "*we've always operated on the idea that the newest model is going to be the best and the most helpful for our customers.*" P4 described choosing "*the most powerful LLM that [their team] feels is economically reasonable to prompt.*" Teams often matched models to task complexity, as P18 noted, using GPT-3.5 for "*straightforward tasks, such as cleaning up typos or translating text*," while reserving GPT-4 quotas for more complex activities, such as generating math problems.

Typical large, off-the-shelf models undergo alignment processes intended to ensure helpful and safe responses [cf. 72]. By design, this would limit the model's capacity to generate responses containing stereotypes or potentially offensive content. P5 recalled trying to use an 'aligned' auxiliary model for safety testing. When the primary model outputted "*I'm sorry, I cannot generate stereotypes,*" the auxiliary model would affirm with responses like "*Great, you should not,*" derailing the evaluation. For this reason, when working on safety-related tasks, such as fine-tuning a primary model to generate safe responses or evaluating for harms, participants often used what they believed were 'unaligned' auxiliary models.

4.2.3 Challenges with Synthetic Data Generation. Participants identified several challenges in generating high-quality synthetic data. Most could be traced to the limitations of the auxiliary models themselves since, as P16 emphasized, synthetic data inevitably reflects the capabilities and limitations of the auxiliary model. Participants described auxiliary models as brittle and highly sensitive to minor changes in prompts. P6 referred to prompt engineering as *"more of an art than a science,"* noting that small adjustments (e.g., reordering words, changing punctuation, or rephrasing sentences) would significantly impact quality. Practitioners found themselves engaged in constant experimentation, attempting to (prompt) "*engineer away*" odd behaviors or inconsistencies.

While natural-looking data was desirable, some practitioners questioned whether synthetic data distributions could closely resemble real data. Many were skeptical, and P7 cautioned that "*one would hope that the distributions are quite similar, but they're not going to be exactly the same on the real and synthetic data.*"

While controllability was viewed as a major benefit of synthetic data, practitioners often struggled to understand or control the outputs of the auxiliary models. P12, who used an auxiliary model for model distillation, acknowledged how its opacity left their team uncertain about whether the primary model was learning the correct function, noting that "*it is hard to say exactly what's going on here.*" P5 encountered a similar issue when using an auxiliary model to create socio-cultural identity prompts to seed safety testing for their primary model. They intended the simulator to generate responses reflecting specific stereotypes for testing purposes but discovered that "*all the responses were about tacos.*" Upon manual inspection, their debugging revealed how "*it was an issue in the pipeline where a single few-shot example that happened to mention Mexican food and tacos [were] being injected and overriding some other stuff.*" And P13, the practitioner who had generated synonyms for a list of successful red-teaming prompts, acknowledged that even with this targeted method, they couldn't fully control the model's semantic interpretations: "*If you just think of the word 'red,' in addition to giving us colors, [the auxiliary model] gave us this more conceptual kind of bloody. It could've gone in a political direction with that. It could've gone in a racist direction with that... That might be in the [data], and we currently don't know.*"

Finally, while synthetic data was thought to have the potential to address gaps in representation by generating data for underrepresented groups, many practitioners described how synthetic data failed to accurately depict smaller populations. P8 noted that this increased the risk of privacy violations by producing outputs that could potentially identify individuals within these groups. For this reason, E20 was apprehensive about any use of synthetic data in the evaluation stage; if synthetic test cases do not adequately reflect

minority groups, the resulting metrics may provide a misleading sense of confidence in the primary model's safety or performance.

4.3 Validating Synthetic Data

We now explore practitioners' current practices and challenges when *validating* synthetic data (RQ3).

4.3.1 Approaches to Data Validation. Practitioners described the validation of synthetic data as "*subtle work*" (P18) that was subjective and required them to rely on their intuition. Manual inspection—also referred to as "*spot-checking*" or "*eyeballing*" by participants—emerged as the most common validation method. P3 described a periodic monitoring routine in which they inspected examples every few weeks to "*make sure the data is as it should be, and that things are going smoothly.*" During these inspections, they often discovered cases where the model "*was not following the prompt correctly and things don't make sense.*" Similarly, P5, who focused on measuring harms, shared their process of manually reviewing datasets to check for alignment with instructions, explaining "*I would generate a dataset and then scroll through it, asking, 'Does this make sense? Is it following the instructions we've given it?'*" Practitioners acknowledged that this approach made it challenging to maintain consistency, especially when the data was inspected by different team members.

Alternative methods beyond manual inspection were rarer. P4, whose work involved simulating user data, described how they "*don't yet have a methodology for evaluation*" of synthetic data. Instead, they focused on the usefulness of the data in downstream applications. A few, including P15 and P18, conducted small-scale studies with human experts (e.g., doctors or trust and safety professionals) to validate synthetic data for specialized tasks. These participants also performed ablation experiments by generating small synthetic datasets and evaluating their impact on the primary model's performance when used for training. P7 described how their team used the internal confidence score of the auxiliary model in validating its own output, noting the circular logic. They set thresholds to classify outputs, treating synthetic data with confidence scores above 90% as high quality and those around 30% as low quality.

4.3.2 Challenges in Data Validation. Participants described validation of synthetic data as a consistent bottleneck. Ironically, the qualities that made synthetic data attractive–such as its scale, diversity, and the ability to simulate data for rare scenarios–also made it exceptionally challenging to validate since, as described above, validation was typically manual.

Certain types of data were especially challenging to inspect manually. P8, who worked in medical imaging, described the challenges of confirming whether synthetic images accurately reflected specific medical conditions, particularly those with subtle visual markers. They collaborated with a resident doctor to assess the realism of their data. Practitioners working in high-stakes domains, like harmful content detection or red-teaming, encountered similar challenges.

Another significant challenge was defining what qualifies as 'good' synthetic data, especially in contexts where no real data exists to guide conceptualization. As described above, practitioners sought data that appeared diverse and natural-looking. Yet, as participants noted, it was difficult to articulate what these criteria mean. For example, P5, who focused on establishing safety evaluations for generative AI products, described how their workflow lacked user data to help gauge the diversity and distribution of model outputs. Practitioners often defaulted to defining diversity as the generation of non-redundant examples, such as P16, who described how "*the main way [they] defined diversity was how many unique examples could be generated. If [they] call the generator model a thousand times, how many redundant examples were generated?*" However, simply avoiding redundancy did not guarantee that the data covered a meaningful range of variation. P3, who focused on detecting stereotypes in primary model outputs, found it challenging to define the types of stereotypes they aimed to capture and operationalize these criteria within prompts. And P17 struggled to validate whether the synthetic data accurately reflected underrepresented groups, noting the lack of clear indicators to confirm that the auxiliary model was following instructions or capturing the intended range of diversity.

Lastly, practitioners highlighted that their companies prioritized scaling data generation to keep pace with model development demands, relegating validation to a lower priority. P3 expressed this tension: "*Honestly, people just don't want to do [the validation] right now because of the urgency associated with everything.*" While this participant recognized the risks of insufficient validation, they felt pressured to move forward to meet internal deadlines, hoping to address data quality issues later if time and resources allowed.

4.4 General Limitations and Ethical Considerations of Synthetic Data

Beyond the challenges specific to generating and validating synthetic data, practitioners and RAI experts identified several broader limitations and ethical considerations with the use of synthetic data.

4.4.1 'Chaining' Auxiliary Models. Practitioners described using the same or similar auxiliary models to produce both training and evaluation data, or to generate evaluation data and score the primary model's outputs. This practice of 'chaining' auxiliary models raised concerns about the risk of amplifying biases at different stages of the AI development pipeline. P10 gave a hypothetical example where the auxiliary model was the same as the model being evaluated: "*GPT-4 as a scorer and asking it to evaluate a GPT-4 based prediction and comparing that to a LLaMA prediction. There's a high chance GPT-4-based prediction will be scored higher because your scoring module is also GPT-4.*" E24 reinforced how the cumulative effect of seemingly *"minor problems or morally neutral outcomes"* could become significant when repeated on a massive scale: "*You've done something that maybe is not that bad, but you have done it 10,000 times instead of 100 times, and you have also done it in a way where you are not really sure what the gaps are and what the dangers are.*" They argued that chaining models creates feedback loops that mask issues that would emerge through more diverse evaluation methods.

Using generative models across multiple stages also introduced the risk of creating distribution shifts that are not easily identifiable without real data to compare against. Both practitioners and RAI experts expressed concerns about the long-term risk of 'model collapse' [cf. 86] where the repeated use of synthetic data at the

training stage causes models to deviate from real-world data distributions, leading to cascading errors and degraded performance over time. Opinions on this risk varied: some participants believed that combining high-quality synthetic data with real data could sustain model development without severe degradation, while others viewed this distribution shift as a slippery slope that resulted in compounding errors that threatened the integrity of models over time. The lack of consensus around this topic was described as a challenge to both understanding the problem and formulating appropriate solutions.

4.4.2 Removed Avenues for Exercising Agency. RAI experts brought up that the use of synthetic data undermines stakeholders' ability to exert meaningful agency over AI systems. They emphasized that synthetic data creates distance between individuals and the data that was originally derived from their activities, decisions, or behaviors. This layer of abstraction erodes opportunities to contest the ways that datasets are used. As E24 noted, "*the link between a person and their data and a training dataset*" is arguably one of the most effective opportunities for participation or redress.

Several RAI experts remarked on the extractive nature of generating synthetic data using auxiliary models that may have been developed through the expropriation of the work of data subjects. E27 linked these concerns to ongoing debates in the digital humanities, where artists, writers, and creators have called for ethical engagement with data derived from their work, critiquing practices that commodify their intellectual and creative output [cf. 47]. E28 suggested that the purported privacy-preserving benefits of synthetic data often come at the expense of others, explaining that for auxiliary models to generate data about a specific group, they require relevant information in the training data: *"so how good this [data is] is a matter of how well it is violating someone else's privacy, right?"*

4.4.3 Stereotyping and Cultural Hegemony. Participants highlighted risks of using identity-related prompting (i.e., including a series of adjectives to describe a simulated user's identity) when generating data. E25 noted that this practice could cause an auxiliary model to either produce stereotypical patterns of behavior or generate overly generic responses due to its alignment process. P6 and E26 both noted that synthetic data intended to reflect identity often lacked nuance and depth, in part because auxiliary models rely on data *about* a group rather than data generated *by* members of the group [cf. 94]. As P6 put it: "*Most of that data is in a tone or a stance, or a perspective of being about these different groups of people, and not by these groups of people, and so there's a lot of cases where we want to generate data as a certain demographic, or as a certain marginalized group, and it almost feels like a caricature of that group.*"

Taking this argument one step further, E22 argued that this misrepresentation of minority groups stems from generated data often defaulting to normative assumptions embedded within the auxiliary model. This raises a broader concern that the use of synthetic data reinforces cultural hegemony—that is, the dominance of one cultural framework (often Western) over others—by embedding dominant values in AI systems, perpetuating a worldview that marginalizes alternative cultural perspectives. Addressing such concerns would demand a fundamental reevaluation of the social and cultural assumptions underlying synthetic data generation.

4.4.4 Organizational Pressure to Scale. While practitioners acknowledged that synthetic data has its limitations, many expressed a sense of resignation toward addressing them. Practitioners described how they were navigating a fast-paced and competitive AI landscape where organizational priorities often emphasized scaling data generation to meet the demands of model development and evaluation. While practitioners recognized that synthetic data could introduce biases or fail to capture the complexity of real-world data, they frequently proceeded despite these risks, with the intention of addressing potential issues later. P18 recognized this recurring tension: "*Frankly, this is a very busy space. I would love to take things slower, but it's just– I don't know. We've had multiple failed attempts to do better documentation of our synthetic data, have better practices, which did not go very well with how fast we worked honestly.*"

RAI experts shared concerns about a growing complacency toward addressing these limitations. E21 expressed a worry that the "*ease of use of synthetic data will so far outweigh, according to the calculus of certain groups, the harms that it does and they're just like 'Oh, we'll just sort of know this is a limitation.'*" Most participants noticed a tendency to rely on partial measures and surface-level fixes (e.g., prompt variation) rather than developing substantive longer-term solutions or foregoing the use of synthetic data.

5 Discussion

The ubiquitous use of auxiliary models reflects practitioners' optimistic perceptions that synthetic data can be a solution for challenges like data scarcity, privacy concerns, resource constraints, and high data collection costs. However, the anticipated benefits of synthetic data were seldom fully realized in practice. The troubling prevalence of logics of scale and automation [35] signal an industry-wide prioritization of efficiency that positions synthetic data as the most practical—even if imperfect—option. Indeed, many practitioners acknowledged unresolved challenges with their use of synthetic data and warned against the danger of normalizing these practices without critically assessing their implications for specific use cases and domains. Practitioners described how the pressure to meet immediate demands often outweighed the desire for rigorous validation. The scale and opacity of synthetic data created through auxiliary models further compound these problems as data issues become more challenging to trace, deferring accountability [97] and leaving systems more prone to perpetuating harm. Thus, the scale of synthetic data emerges as both a promise and a peril.

While our findings highlight significant concerns about the role of auxiliary models in development practices, we resist framing synthetic data as inherently problematic. An AI development pipeline that excludes synthetic data may involve other challenges, for instance, in meeting the ever-increasing demands of evaluation for robust model development across a range of harm areas and specialized domains, as required by policymakers or industry standards. We contend that the impacts of synthetic data are multifaceted and plural, with both potential benefits and risks, and invite critical, ongoing investigation by the FAccT community.

5.1 Interrogating Impacts to the AI Supply Chain

Our findings capture the perspectives of practitioners embedded within systems of institutional power. This 'studying up' approach [8, 68] may fall short in surfacing how synthetic data use reconfigures other aspects of the AI supply chain [99]. Below, we outline opportunities to investigate impacts on model deployers, data workers, and model subjects.

Building AI from scratch requires vast amounts of high-quality data that raises the barrier for entry for new companies seeking to develop applications [98]. One might argue that synthetic data could contribute to the democratization of model development since smaller companies no longer need to have access to large-scale, costly, hard-to-obtain datasets to train their models [55]. In theory, synthetic data could counteract the concentration of power within the hands of a few incumbents by allowing startups or other entities to compete against large tech companies.

Nevertheless, there is a paradox in this democratization narrative. The auxiliary models typically used to generate synthetic data are almost exclusively developed by the same tech giants that already dominate the AI industry [98]. Smaller companies or researchers who wish to train a model themselves are likely to rely on these auxiliary models offered as-a-service to produce synthetic data [59]. This results in a consolidation of power as big tech companies monopolize the tools and frameworks [12] that smaller companies must rely on for generating synthetic data [6]. As synthetic data pipelines become entangled with proprietary model APIs, usage restrictions, and tiered access to compute, the ability to experiment or contest how synthetic data is generated is increasingly constrained.

We must also consider how these emerging practices impact the data supply chain. In recent years, data work has shifted from independent platform workers to private annotation firms specializing in labeling services [95]. These firms often operate within opaque supply chains that hide the outsourcing of work to the Global South [65]. Participants noted that synthetic data was often used to avoid the need to engage with data workers, but training and testing infrastructures that use auxiliary models still require human input. The effects of the shift on labor infrastructures remain underexplored. How do these emerging practices impact job opportunities in data work? How does the increasing reliance on synthetic data potentially reshape data workers' roles from creators to 'custodians of data quality'? How might synthetic data practices provide opportunities for data workers' upward mobility, and under what conditions would such shifts be equitable? These questions require further investigation.

On the other hand, auxiliary model use offers potential opportunities to mitigate some of the harms associated with traditional data annotation, particularly in cases where workers are exposed to toxic or harmful content. Participants mentioned how automating the evaluation of certain types of harms (**e.g.,** those involving violent, graphic content) and relying on annotators for subtle or complex forms of harm, could reduce the need for human annotators to engage directly with harmful material that has been linked to psychological distress and trauma [87].

Finally, the use of synthetic data for training and evaluating models has significant implications for model subjects. Studies have shown that distribution shifts caused by synthetic data during training can lead to reductions in model performance and fairness, and to increased representational harms for minoritized groups [104]. If auxiliary models are *also* used in the evaluation, they might fail to detect critical issues, particularly those that disproportionately affect underrepresented groups or rare phenomena. This risk of degraded model performance during training, which is likely unnoticed in evaluation, also highlights the urgency of advancing research in this area. What mechanisms could then be developed to scrutinize and mitigate the ways in which general-purpose models propagate biases into synthetic datasets?

5.2 Toward Considerations for Responsible Use of Synthetic Data

Our interviews reflect a particular moment in time when practices and norms around the use of synthetic data in both industry and research contexts are rapidly evolving. The viewpoints expressed by different participants were often in tension with each other or with the growing synthetic data literature [78]. Some participants argued that synthetic data is well-suited for use in training due to its tolerance for noise but is too risky to use in evaluation, while others expressed the opposite viewpoint. Some praised the use of synthetic data because it gave them "*full control*," while others recounted difficulties controlling data generation in practice. Some proposed that auxiliary models can be used to simulate data from members of underrepresented groups, a viewpoint some literature supports [31], while others argued that data generated in this way would be merely a "*caricature*" of the group, in line with other scholarship [94]. These disagreements reflect the contested nature of our collective understanding of synthetic data's benefits, limitations, and appropriate use. Because of this, rather than offering prescriptive recommendations, we focus on foregrounding key considerations that researchers and practitioners must navigate as they engage with synthetic data.

Practitioners should carefully consider what constitutes an appropriate application of synthetic data. Deciding whether to use synthetic data, and in what capacity, involves assessing the intended purpose or use case and the domain of application, and balancing trade-offs between resource availability and the 'right' methodology for data generation and validation. For example, in a 'high-risk' domain (like training a model for a medical application), it is arguably more crucial to clearly operationalize the desiderata (diversity, for instance) and to systematically validate that the data meets those characteristics. In 'low-risk' domains, there is perhaps more tolerance for noise or fuzzy data. Future research could investigate which dimensions of a use case or domain are relevant in determining the appropriate scope of use for synthetic data. Of course, what counts as high-risk or low-risk is itself a value-laden assumption that must be interrogated.

Practitioners should carefully consider auxiliary model capabilities and limitations when selecting which model to use. Synthetic data practices are heavily entangled with the selection of auxiliary models. The harms of synthetic data are a result of the assumptions and design decisions made during the creation of synthetic data *and* the limitations of the auxiliary models. Each model has a different risk profile [92], which makes auxiliary model

selection a consequential decision. Despite this, participants in our study often defaulted to selecting what they perceived as the state-of-the-art model. This tendency reflects a broader pattern in AI practices, where newer models are assumed to outperform older ones [102], even when their suitability for particular contexts remains untested. Practitioners could benefit from a more deliberate approach to model selection, where they experiment with whether a particular model aligns with their specific needs.

In addition to careful model selection, there is a need to make synthetic data validation more systematic. Participants in this study relied on 'eye-balling' or 'spot-checking' as their primary validation method. While manual review offers a direct way to identify errors, participants frequently reported it as inconsistent, insufficient, or deprioritized over the demands of data generation. First, practitioners need to invest more time and resources in synthetic data validation. Second, practitioners should develop a methodology that ideally combines multiple forms of data validation (e.g., manual verification *and* calibration experiments). There would be several considerations to keep in mind. For example, instead of randomly sampling data, one could prioritize defining (including with the absence of real-world data) and creating underrepresented or edge-case examples. Similarly, to guide validation efforts, practitioners could develop a rubric that articulates the criteria against which they compare the synthetic data. Yet, how can such qualities be effectively measured, and how might reliance on proxy metrics introduce additional risks? We call for research and experimentation to develop effective validation practices for synthetic data.

Researchers should also support the development of artifacts and processes for synthetic data documentation. Data practices among our participants were highly iterative and typically involved multiple rounds of generation and validation, interleaved together, making synthetic datasets fluid and iterative objects that change constantly. However, rarely did participants prioritize documentation of data or process, either due to organizational priorities or the perception that code or prompts served as sufficient documentation for synthetic data. While there are plenty of resources for data and model documentation [9, 14, 25, 66], there is almost no existing guidance on what to document for synthetic data that is created and used in such myriad ways. The EU's General Purpose AI Code of Practice [70] recommends documenting the methods used to generate synthetic training data, but does not yet get into specifics or touch on use cases beyond training. What aspects of the synthetic generation process should be recorded? For example, how should revisions to prompts, examples of rejected data, or the reasoning behind design decisions be captured? Practitioners should also consider documenting their choice of the auxiliary model as well as a justification. Beyond recording data provenance, there are broader questions about synthetic data documentation that warrant attention: Who are the intended audiences, and what purposes should synthetic data documentation serve? How should documentation practices be adapted to account for the experimental and iterative process that practitioners described?

Lastly, AI workflows that incorporate auxiliary models must prioritize meaningful community engagement. Synthetic data inherently reduces opportunities for direct participation, as it is neither created nor annotated by individuals or communities. Often, an underlying motivation for using synthetic data is to "take the people out" [88], whether to address privacy concerns, meet compliance requirements, or minimize exposure to harmful content. While these intentions may serve some goals, they also exclude data subjects and data workers from the pipeline and restrict opportunities to question or shape datasets. To increase participation, participants in our study proposed a consultation model where experts or community members provide input on use cases, evaluation outcomes, or critical decisions [17]. However, when is it appropriate to consult only experts, and when should affected communities play a central role? The answer likely depends on the specific use case and the degree of transformation involved. For example, small or targeted modifications to existing datasets might raise different participatory needs compared to entirely synthetic datasets. The timing of participation is equally significant. Should stakeholders contribute during the generation process, shaping decisions about what data is created and how? Or is it more effective to involve them during validation? These decisions shape the accountability of synthetic data practices to the communities they may impact.

In highlighting these considerations, we recognize how meaningful change requires rethinking incentives for different stakeholders across the AI supply chain. First, regulation could establish reporting requirements for both auxiliary model producers and downstream developers. For model producers, this could involve disclosing evaluations of their models' capacity to generate synthetic data in specific domains, along with guidance on the intended scope of use. Developers using synthetic data could be expected to document prompts, generation parameters, and any modifications made to the outputs. Furthermore, demonstrating that this kind of reporting leads to better outcomes (e.g., product usage or increased user trust) would establish clear incentives for these evaluations.

6 Conclusion

In this paper, we illustrate the expanding role of synthetic data in AI development. Through interviews with AI practitioners and RAI experts, we highlight the motivations and promises of synthetic data, its various use cases across the AI development pipeline, current practices and challenges around its generation and validation, its limitations, and ethical considerations for its use. We discuss the implications of our findings for the AI supply chain. Finally, we propose several considerations for improving synthetic data practices. We note that while synthetic data is valued precisely for its efficiency, practices like rigorous validation, documentation, or participatory methods all require human intervention, increasing costs and extending timelines. To arrive at adoptable practices, we must reconcile competing demands between the need for rigor and the desire to maintain efficiency in line with organizational priorities and constraints.

7 End Matter

Ethical considerations. We took several measures to protect our research participants. Our study design was reviewed and approved by our organization's Institutional Review Board. Participation was voluntary. Informed consent was obtained electronically and we compensated participants with a gift card of $75 USD. Participants were informed of the purpose of the study and were told that they

could refuse to answer any questions and ask for the recording to be paused at any time. All participant data, including interview transcripts and recordings, was stored in a secure location and was not accessible by anyone outside the research team. The research team checked all included quotes to ensure that no identifying information about participants or their organizations was revealed.

Positionality. All authors are, or were at the time of conducting the research, employed at a large technology company. This positioning afforded us access to ongoing conversations within the industry around synthetic data practices, including emerging use cases. At the same time, we recognize that our proximity to industry priorities may shape how we frame risks and recommendations. Throughout this work, we have aimed to remain accountable to perspectives that may not align with institutional interests by foregrounding scholarship from critical data studies and the political economy of the AI supply chain. We offer this reflection to acknowledge how our embeddedness within industry shapes the kinds of questions we ask and the interpretations we bring to our analysis, and to invite ongoing dialogue with communities who approach this topic from different positions or commitments.

Potential adverse impacts. First, our findings focus exclusively on practitioners' and RAI experts' perspectives. Our findings may inadvertently privilege their viewpoints and concerns over those of other crucial stakeholders. This could lead to recommendations that undervalue the needs of affected communities, policymakers, and civil society organizations. Second, our research sheds light on motivations, practices, and challenges for the use of synthetic data in the AI development pipeline. We also expose limitations and ethical considerations. Emphasizing these drawbacks may be interpreted as evidence that the use of synthetic data should be more restricted. However, we caution against this interpretation of our results. In setting policy guidance for synthetic data, policymakers should weigh the benefits as well. Perhaps most notably, synthetic data enables the evaluation of models and systems at a scale that would not be possible through other means; without it, many evaluations simply may not happen. To address this concern, we have aimed throughout this paper to provide a balanced perspective, highlighting both the benefits and drawbacks of synthetic data. We additionally emphasize that norms and practices around the use of synthetic data are rapidly evolving. While we hope that our research inspires future work developing best practices for the responsible use of synthetic data, these practices will need to be constantly evaluated and updated to ensure that they remain practical and relevant.

Acknowledgments

First and foremost, we are grateful to our study participants, without whom this research would not be possible. This work was enriched by many useful conversations with Anubhav Jangra, Nari Johnson, Daniela Massiceti, Cecily Morrison, Sachita Nishal, and the members of the Microsoft Research FATE group, Microsoft's Sociotechnical Alignment Center, and the Tech Solidarity Lab at CMU. Our sincere thanks to Hanna Wallach and Ece Kamar for making connections to the AI practitioner community and helping with recruitment. We are also grateful to Agathe Balayn, Inha Cha, Sarah Fox, and Jordan Taylor for providing feedback on early drafts of this paper.

References

[1] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. 2019. Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer, 510–526.

[2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905* (2024).

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[4] William Agnew, A Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[5] Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. 2020. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*. 1–8.

[6] Pierre Azoulay, Joshua L Krieger, and Abhishek Nagaraj. 2024. *Old moats for new models: Openness, control, and competition in generative ai.* Technical Report. National Bureau of Economic Research.

[7] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. 2023. DigiFace-1M: 1 Million Digital Face Images for Face Recognition. In *2023 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE.

[8] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 167–176.

[9] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[10] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[11] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[12] Sarah Burkhardt and Bernhard Rieder. 2024. Foundation models are platform models: Prompting and the political economy of AI. *Big Data & Society* 11, 2 (2024), 20539517241247839.

[13] Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937* (2023).

[14] Kasia Chmielinski, Sarah Newman, Chris N. Kranzinger, Michael Hind, Jennifer Wortman Vaughan, Margaret Mitchell, Julia Stoyanovich, Angelina McMillan-Major, Emily McReynolds, Kathleen Esfahany, Mary L. Gray, Audrey Chang, , and Maui Hudson. 2024. The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers. Harvard Kennedy School Shorenstein Center discussion paper.

[15] Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory. 2020. Passing the data baton: A retrospective analysis on data science work and workers. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1860–1870.

[16] Jessamyn Dahmen and Diane Cook. 2019. SynSys: A synthetic data generation system for healthcare applications. *Sensors* 19, 5 (2019), 1181.

[17] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–23.

[18] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 473–484.

[19] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 705–716.

[20] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems* 36 (2024).

[21] Madeleine Clare Elish and Danah Boyd. 2018. Situating methods in the magic of Big Data and AI. *Communication monographs* 85, 1 (2018), 57–80.

[22] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* 9 (2021), 391–409.

[23] Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. *arXiv preprint arXiv:2305.08281* (2023).

[24] Andrew Fitzgerald. 2024. Why Synthetic Data Can Never Be Ethical: A Lesson from Media Ethics. *Surveillance & Society* 22, 4 (Dec. 2024), 477–482. https://doi.org/10.24908/ss.v22i4.18324

[25] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (December 2021), 86–92.

[26] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.

[27] Lisa Gitelman. 2013. "Raw Data" Is an Oxymoron.

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[29] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge Distillation: A Survey. *International Journal of Computer Vision* 129 (2021), 1789–1819.

[30] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[31] Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. AI and the transformation of social science research. *Science* 380, 6650 (2023), 1108–1109.

[32] Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, Aonan Zhang, Bowen Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, et al. 2024. Apple intelligence foundation language models. *arXiv preprint arXiv:2407.21075* (2024).

[33] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE intelligent systems* 24, 2 (2009), 8–12.

[34] Lei Han, Tianwa Chen, Gianluca Demartini, Marta Indulska, and Shazia Sadiq. 2023. A data-driven analysis of behaviors in data curation processes. *ACM Transactions on Information Systems* 41, 3 (2023), 1–35.

[35] Alex Hanna and Tina M Park. 2020. Against scale: Provocations and resistances to scale thinking. *arXiv preprint arXiv:2010.08850* (2020).

[36] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. http://arxiv.org/abs/2203.09509 arXiv:2203.09509 [cs].

[37] Amy K Heger, Liz B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–29.

[38] Paula Helm, Benjamin Lipp, and Roser Pujadas. 2024. Generating reality and silencing debate: Synthetic data as discursive device. *Big Data & Society* 11, 2 (June 2024), 20539517241249447. https://doi.org/10.1177/20539517241249447 Publisher: SAGE Publications Ltd.

[39] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

[40] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[41] Qisheng Hu, Kaixin Li, Xu Zhao, Yuxi Xie, Tiedong Liu, Hui Chen, Qizhe Xie, and Junxian He. 2023. Instructcoder: Empowering language models for code editing. *arXiv preprint arXiv:2310.20329* (2023).

[42] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608* (2022).

[43] IBM. 2023. What is Synthetic Data? | IBM. https://www.ibm.com/think/topics/synthetic-data [Online; accessed 2025-01-21].

[44] Apple Inc. 2024. Introducing Apple's On-Device and Server Foundation Models - Apple Machine Learning Research. https://machinelearning.apple.com/research/introducing-apple-foundation-models [Online; accessed 2025-01-21].

[45] Benjamin N Jacobsen. 2023. Machine learning and the politics of synthetic data. *Big Data & Society* 10, 1 (Jan. 2023), 20539517221145372. https://doi.org/10.1177/20539517221145372 Publisher: SAGE Publications Ltd.

[46] Nikita Jaipuria, Xianling Zhang, Rohan Bhasin, Mayar Arafa, Punarjay Chakravarty, Shubham Shrivastava, Sagar Manglani, and Vidya N Murali. 2020. Deflating dataset bias using synthetic data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 772–773.

[47] Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. AI Art and its Impact on Artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 363–374.

[48] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. 2022. Synthetic Data–what, why and how? *arXiv preprint arXiv:2205.03257* (2022).

[49] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*.

[50] Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the Snark: Annotator Diversity in Data Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3544548.3580645

[51] Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[52] Laura Koesten, Elena Simperl, Tom Blount, Emilia Kacprzak, and Jeni Tennison. 2020. Everything you always wanted to know about a dataset: Studies in data summarisation. *International Journal of Human-Computer Studies* 135 (March 2020), 102367. https://doi.org/10.1016/j.ijhcs.2019.10.004

[53] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. 2019. Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2261–2268. https://doi.org/10.1109/CVPRW.2019.00279

[54] Francis Lee, Saghi Hajisharif, and Ericka Johnson. 2025. The ontological politics of synthetic data: Normalities, outliers, and intersectional hallucinations. *Big Data & Society* 12, 2 (2025), 20539517251318289.

[55] Peter Lee. 2024. Synthetic Data and the Future of AI. (2024).

[56] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9493–9500.

[57] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[58] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best Practices and Lessons Learned on Synthetic Data for Language Models. https://doi.org/10.48550/arXiv.2404.07503 arXiv:2404.07503 [cs].

[59] Dieuwertje Luitse and Wiebke Denkena. 2021. The great transformer: Examining the role of large language models in the political economy of AI. *Big Data & Society* 8, 2 (2021), 20539517211047734.

[60] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583* (2023).

[61] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 26 pages.

[62] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[63] Daniella Meeker, Crystal Kallem, Yan Heras, Stephanie Garcia, and Casey Thompson. 2022. Case report: evaluation of an open-source synthetic data platform for simulation studies. *JAMIA open* 5, 3 (2022), ooac067.

[64] Cade Metz, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson, and Nico Grant. 2024. How Tech Giants Cut Corners to Harvest Data for A.I. - The New York Times. https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html [Online; accessed 2025-01-21].

[65] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.

[66] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229.

[67] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[68] Laura Nader. 1972. Up the anthropologist: Perspectives gained from studying up. (1972).

[69] Sergey I Nikolenko. 2021. *Synthetic data for deep learning*. Vol. 174. Springer.

[70] Code of Practice Working Groups. 2024. Second Draft of the General-Purpose AI Code of Practice. https://digital-strategy.ec.europa.eu/en/library/second-draft-general-purpose-ai-code-practice-published-written-independent-experts [Online; accessed 2025-01-20].

[71] Office of the Assistant Secretary for Planning and Evaluation (ASPE). 2022. A Synthetic Health Data Generation Engine to Accelerate Patient-Centered Outcomes Research | ASPE. https://aspe.hhs.gov/synthetic-health-data-generation-engine-accelerate-patient-centered-outcomes-research [Online; accessed 2025-01-22].

[72] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[73] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[74] Samir Passi and Steven Jackson. 2017. Data Vision: Learning to See Through Algorithmic Abstraction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2436–2447. https://doi.org/10.1145/2998181.2998331

[75] Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.

[76] Personal Data Protection Commission (PDPC). 2024. Privacy Enhancing Technology (PET): Proposed Guide On Synthetic Data Generation. https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/other-guides/proposed-guide-on-synthetic-data-generation.pdf [Online; accessed 2025-01-20].

[77] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. *arXiv:2202.03286 [cs]* (Feb. 2022). http://arxiv.org/abs/2202.03286 arXiv: 2202.03286.

[78] Crystal Qian, Michael Xieyang Liu, Emily Reif, Grady Simon, Nada Hussein, Nathan Clement, James Wexler, Carrie J Cai, Michael Terry, and Minsuk Kahng. 2024. The Evolution of LLM Adoption in Industry Data Curation Practices. *arXiv preprint arXiv:2412.16089* (2024).

[79] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[80] Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2020. Synthetic data augmentation for zero-shot cross-lingual question answering. *arXiv preprint arXiv:2010.12643* (2020).

[81] Kevin Roose. 2024. Data for A.I. Training Is Disappearing Fast, Study Shows - The New York Times. https://www.nytimes.com/2024/07/19/technology/ai-data-restrictions.html [Online; accessed 2025-01-21].

[82] Roy A Ruddle, James Cheshire, and Sara Johansson Fernstad. 2023. Tasks and visualizations used for data profiling: A survey and interview study. *IEEE Transactions on Visualization and Computer Graphics* (2023).

[83] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[84] Siamak Shakeri, Noah Constant, Mihir Sanjay Kale, and Linting Xue. 2020. Towards zero-shot multilingual synthetic question and answer generation for cross-lingual reading comprehension. *arXiv preprint arXiv:2010.12008* (2020).

[85] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. 2018. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*. Springer, 1–11.

[86] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature* 631 (2024), 755–759.

[87] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[88] Daniel Susser, Daniel S Schiff, Sara Gerke, Laura Y Cabrera, I Glenn Cohen, Megan Doerr, Jordan Harrod, Kristin Kostick-Quenet, Jasmine McNealy, Michelle N Meyer, et al. 2024. Synthetic Health Data: Real Ethical Promise and Peril. *Hastings Center Report* 54, 5 (2024), 8–13.

[89] Daniel Susser, Daniel S. Schiff, Sara Gerke, Laura Y. Cabrera, I. Glenn Cohen, Megan Doerr, Jordan Harrod, Kristin Kostick-Quenet, Jasmine McNealy, Michelle N. Meyer, W. Nicholson Price II, and Jennifer K. Wagner. 2024. Synthetic Health Data: Real Ethical Promise and Peril. *Hastings Center Report* 54, 5 (2024), 8–13. https://doi.org/10.1002/hast.4911 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hast.4911.

[90] Daniel Susser and Jeremy Seeman. [n. d.]. Dialogue Critical Provocations for Synthetic Data. ([n. d.]).

[91] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[92] Risto Uuk, Annemieke Brouwer, Tim Schreier, Noemi Dreksler, Valeria Pulignano, and Rishi Bommasani. 2024. Effective Mitigations for Systemic Risks from General-Purpose AI. *arXiv preprint arXiv:2412.02145* (2024).

[93] Janet Vertesi and Paul Dourish. 2011. The value of data: considering the context of production in data economies. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 533–542.

[94] Angelina Wang, 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 arXiv:2402.01908.

[95] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose AI Dream? In search of the aspiration in data annotation.. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16.

[96] Cedric Deslandes Whitney and Justin Norman. 2024. Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1733–1744. https://doi.org/10.1145/3630106.3659002

[97] David Gray Widder and Dawn Nafus. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society* 10, 1 (2023), 20539517231177620.

[98] David Gray Widder, Sarah West, and Meredith Whittaker. 2023. Open (for business): Big tech, concentrated power, and the political economy of open AI. *Concentrated Power, and the Political Economy of Open AI (August 17, 2023)* (2023).

[99] David Gray Widder and Richmond Wong. 2023. Thinking upstream: Ethics and policy opportunities in AI supply chains. *arXiv preprint arXiv:2303.07529* (2023).

[100] David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. 2023. It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 467–479.

[101] Tanja Wiehn. 2024. Synthetic Data: From Data Scarcity to Data Pollution. *Surveillance & Society* 22, 4 (Dec. 2024), 472–476. https://doi.org/10.24908/ss.v22i4.18327

[102] Matteo Wong. 2024. The GPT Era Is Already Ending - The Atlantic. https://www.theatlantic.com/technology/archive/2024/12/openai-o1-reasoning-models/680906/?utm_source=chatgpt.com [Online; accessed 2025-01-22].

[103] Kanit Wongsuphasawat, Yang Liu, and Jeffrey Heer. 2019. Goals, process, and challenges of exploratory data analysis: An interview study. *arXiv preprint arXiv:1911.00568* (2019).

[104] Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. 2024. Fairness feedback loops: training on synthetic data amplifies bias. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2113–2147.

[105] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. 2018. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739* (2018).

[106] Zhenpei Yang, Yuning Chai, Dragomir Anguelov, Yin Zhou, Pei Sun, Dumitru Erhan, Sean Rafferty, and Henrik Kretzschmar. 2020. Surfelgan: Synthesizing realistic sensor data for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11118–11127.

[107] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People+ AI Guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[108] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284* (2023).

[109] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? Roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[110] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.

A Additional Study Details

In this appendix, we provide the interview protocols used in each phase of our study.

A.1 Interview Protocol for Phase 1

Note: This interview protocol contains questions that are not relevant to some participants. The moderator decided which question to follow up with based on the participant's AI pipeline and project description. For example, participants whose focus was on evaluation were not asked questions related to training, and vice versa.

A.1.1 Introduction.

- Introduction (role, background, years of experience)
- How large is your team?
- What kinds of projects do you focus on?
- What types of data do you deal with?

A.1.2 ML pipeline. Now would be a good time to shift gears and learn a little more about a specific project where you were involved in data generation, curation, or model evaluation.

- Could you give me a brief overview of your ML pipeline step-by-step, from data development to model training, evaluation, and deployment? It would be helpful to get a full picture of the various stages where you deal with different kinds of data.
- Follow-up clarification questions:
  - Are there any labels involved in this data?
  - What are your data sources? Where does this data come from?
  - Do you augment the dataset in any way or make perturbations?
  - Would you say [data they described] is synthetic data? Why or why not?

[If they do not already talk about model-generated or simulated data as part of the earlier question] Have you ever or do you currently use synthetic, simulated, model-generated, or augmented data in your work? For example, data generated by LLMs that you use to train or evaluate other LLMs, or simulating users with LLMs by automatically generating prompts.

- What do you usually call this kind of data?
- Would you consider this synthetic data?
- Are there other ways you have worked with synthetic, simulated, model-generated, or augmented data?
- What counts as synthetic data in your field? What terms do you use to describe this kind of data?
- (If they say no to all forms of synthetic data) Did you ever consider using these kinds of data but then decide against it? If so, why?
- Do you ever use models to evaluate other models or systems? For example by simulating users or automatically generating prompts? What made you try this approach?
- Do you also incorporate human feedback into their evaluations?
- How do you decide when you need humans in the loop and when you don't?
- Can you walk me through your process of generating synthetic data in some detail?

A.1.3 Model training.

- Do you use synthetic data exclusively for training, or is it combined with real-world data?
- Are there specific things you need to do or look out for before you combine synthetic data with other data sources, such as real world data?
- Have you encountered any performance bottlenecks or computational challenges when working with large-scale synthetic datasets?
- Do you have concerns relating to distribution shifts or 'model collapse'?

A.1.4 Model evaluation.

- What kinds of evaluation did you run for this project?
- What is the goal of the evaluation? Do you have specific objectives for the evaluation?
- Do you incorporate feedback from evaluations with synthetic data into your model development process? If so, how?

A.1.5 Prompt-generated data.

- How did you choose the pre-trained LLM? What were the factors that went into this decision?
- How do you come up with the prompt templates that are used to generate data with the LLM? (Is this a manual or automated process?) What kinds of expertise is needed to select the prompt templates?
- Could you share an example of a prompt that you used? Did you try different variations of prompts before deciding on this approach? When evaluating data for each iteration of the prompt, what were you assessing in the data? Why wasn't it meeting your needs or what led you to changing the prompt?
- What makes this technique effective for your use case?

A.1.6 Understanding synthetic data practices.

- Did you always use synthetic data in your data pipeline, or did you make a shift? If so, why?
- How did you decide to use synthetic data?
- What makes synthetic data appealing for the work you do? [to get at all the benefits]
- What objectives does it serve for you? What goals do you have?
- Did you face technical or organizational challenges when shifting to synthetic data?
- What does good synthetic data look like?
- How do you define success for synthetic data?
- What counts as high-quality synthetic data for your work?
- Do you ever think about the 'realistic-ness' of synthetic data? How 'realistic' does synthetic data need to be for your purposes?
- How do you think about glitches, artifacts and imperfections of real-world data?
- How do you evaluate or validate synthetic data? Fidelity, reliability etc.

- For evaluating the generated data: do you qualitatively look through the samples or use any metrics?
- Where do you face the biggest challenge with data?
- Have you deployed a system that uses synthetic data?
- Have you ever noticed any downstream issues that you attribute to synthetic data?

A.1.7 Responsible AI considerations.

- Are you aware of any responsible AI considerations with the use of synthetic data?
- How do you navigate these concerns?
- Have you come across any guidance or best practices for using synthetic data?
- Are there any resources that you found helpful?
- Do you think synthetic data helps you mitigate responsible AI issues (such as privacy, fairness-related harms, or consent)?
- On the other hand, does it also exacerbate any responsible AI issues?
- Transparency:
  - Do you record any information or create documentation for synthetic data?
  - What kinds of information do you record?
  - Would you say that it is similar to 'real-world' data or are there fields unique to synthetic data that you need to include?
  - Is this data shared publicly or used by other teams? Do you have to process the data to make it usable for other teams?
  - When you are using model-generated data, does the data underlying the model inform your approach? Do you have a good sense of the data that is used to train the model you are using?
  - Does the provenance of synthetic data matter for the work you do?
- Fairness:
  - How do you think about fairness in the context of synthetic data?
  - Is there any form of 'debiasing' that needs to happen with synthetic data?
  - Are there considerations around diversity and representation that are relevant to synthetic data?

Are there things you would like to share that we did not get to discuss today?

Thanks for your time!

A.2 Interview Protocol and Vignettes for Phase 2

A.2.1 Introduction.

- Introduction (role, background, years of experience)
- What does responsible AI mean for the work you do? Are there any specific areas of RAI that you focus on?
- What kinds of projects do you work on?

This is a good time to shift gears and discuss the recent shifts towards using synthetic data or model-generated data or simulated data or augmented data in the ML pipeline (e.g., for data creation or model evaluation), and how that might impact responsible AI considerations.

Before we go into specific scenarios of how this data is being used today, it would be nice to start off hearing your broad perspective on these emerging practices.

- Has your work touched on the use of synthetic, simulated, model-generated, or augmented data across the ML pipeline or have you previously thought about the use of such data?
- Are you aware of any responsible AI considerations with the use of synthetic data?
- Do you think synthetic data helps you mitigate responsible AI issues (such as privacy, fairness-related harms, or consent)?
- On the other hand, does it also exacerbate any responsible AI issues?

Thanks, this is very helpful. Here are the three scenarios. [Show slide with the three scenarios presented in Table 2.]

A.2.2 Reflecting on vignettes. For each vignette:

- What are the major concerns of using models for this purpose?
- What kinds of trade-offs are involved in this scenario and are there good ways to make decisions around this?
- Do you think there is a need to validate the generated data? If yes, which properties might be important to validate?
- What are the implications of not doing this validation?
- What kinds of strategies might help us validate (evaluate) synthetic data or model-generated data?
- Do you anticipate any potential risks or downstream impacts with using model-generated data in this way?

Thanks for covering these scenarios with me.

- What kinds of guidance or best practices should teams follow when creating/using synthetic data?
- Are there contexts where using model-generated data might be better than traditional methods? Why?
- We know of trade-offs around synthetic data (e.g., privacy and utility, realistic data & copying training data). How should practitioners navigate these trade-offs and make decisions on when & how to use synthetic data?

A.2.3 Responsible AI considerations.

- Does synthetic data mitigate some responsible AI issues?
- Does synthetic data also exacerbate some responsible AI issues?
- Does this use of synthetic data introduce new challenges or directions for responsible AI work?
- How can synthetic data help or hinder transparency in AI systems?
- How does the process of creating synthetic data impact ethical considerations?
- Do we need to think about fairness differently in the context of synthetic data?
- Are there considerations around diversity and representation that are relevant to synthetic data?
- How should the provenance of synthetic data be documented to ensure accountability and trustworthiness?

Scenario 1: **Fine-tuning data**	**Scenario 2:** **Evaluation**	**Scenario 3:** **Simulated data**
The SmartTeach team is developing question-answering (QA) systems for tutoring. They would like to offer hints within their system to scaffold learning. Due to limited task-specific data, the team is leveraging a pre-trained Language Model (LLM) to generate hints for an existing QA dataset. The hints dataset will be used for fine-tuning a smaller generative model.	The Responsible AI team works on identifying harms within LLM conversations. People need to interact with an LLM to understand failure modes, but this is often expensive to scale and is also time-consuming. The Responsible AI team is exploring a new technique: they are prompting an LLM with a topic & user identity characteristics. This LLM will have multi-turn conversations with the system-under-test. Following these interactions, another LLM 'scores' the conversation to detect potential harm.	The PrivacyOps team supports internal teams to better understand user feedback while safeguarding data privacy. They are using an LLM to simulate employee interactions within a hypothetical organization. This involves creating personas and generating synthetic documents and emails that resemble real-world situations. This simulated data will be used to test models and train downstream classifiers.

Table 2: Vignettes of Synthetic Data Use

What roles do stakeholders (such as model producers, researchers, regulators, and civil society) play in ensuring responsible use of synthetic data across the ML pipeline?

Are there things you would like to share that we did not get to discuss today?

Thanks for your time!