Institutional Repository - Research Portal
Dépôt Institutionnel - Portail de la Recherche
researchportal.unamur.be
University of Namur

RESEARCH OUTPUTS / RÉSULTATS DE RECHERCHE

Integrating Generative AI into Information Systems Research

Bono Rossello, Nicolas; Simonofski, Anthony; Rossello, Lluc Bono; Castiaux, Annick

Published in:
Proceedings of the 58th Hawaii International Conference on System Sciences, HICSS 2025

DOI:
10.24251/hicss.2025.861

Publication date:
2025

Document Version
Publisher's PDF, also known as Version of record

Link to publication

Citation for pulished version (HARVARD):
Bono Rossello, N, Simonofski, A, Rossello, LB & Castiaux, A 2025, Integrating Generative AI into Information Systems Research: A Framework for Synthetic Data Evaluation. in TX Bui (ed.), *Proceedings of the 58th Hawaii International Conference on System Sciences, HICSS 2025.* Proceedings of the Annual Hawaii International Conference on System Sciences, IEEE Computer Society, pp. 7195-7204, 58th Hawaii International Conference on System Sciences, HICSS 2025, Honolulu, United States, 7/01/25. https://doi.org/10.24251/hicss.2025.861

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 11. Aug. 2026

Integrating Generative AI into Information Systems Research: A Framework for Synthetic Data Evaluation

Nicolas Bono Rossello
University of Namur, Belgium
nicolas.bonorossello@unamur.be

Anthony Simonofski
University of Namur, Belgium
anthony.simonofski@unamur.be

Lluc Bono Rossello
Université Libre de Bruxelles, Belgium
lluc.bono.rossello@ulb.be

Annick Castiaux
University of Namur, Belgium
annick.castiaux@unamur.be

Abstract

Generative AI is paving its way into the research process. Among the plethora of available generative AI solutions, the generation of synthetic data is one of the most controversial. The current division of opinion and the lack of formal approach to AI use in research create a situation of conflicting bad practices and under-used potential. This work aims to add nuance and structure to this research practice by providing a general framework to evaluate the use of synthetic data in different stages of the research process, based on the objective and methods of generation. Relying on a breakout literature review, we explore the fields of Data quality management and Control theory to transfer method theories from these fields to help us build the framework. The resulting conceptual framework provides an iterative scheme where, based on the desired properties of the data and its comparison to the synthetic result, the researcher can improve the outcome of the generation process and, equivalently, formally present the properties that make this data suitable for research.

Keywords: Synthetic data, Data quality, Control Theory, Information Systems Research.

1. Introduction

Machine learning, network analysis or numerical simulations are some examples of computationally intensive techniques that can be found at different stages of the research process (Berente et al., 2019). These techniques enhance the research process by enabling deductive actions or facilitating exploratory purposes. Synthetic data is a potential outcome of these computationally intensive approaches, and has long been used in fields that require numerical applications (e.g., engineering, economics, or physics), including the field of Information Systems (IS) research (Livieris et al., 2024). In this paper, we consider the term synthetic data in a broad sense; that is, any information which is artificially generated rather than produced by real world events (Kowalczyk et al., 2022).

The advantages of synthetic data for research are numerous: the privatization of data, the augmentation of datasets or the in-silico evaluation of physical processes. With the emergence of generative AI, we face the possibility of generating textual or visual data based on pre-existing datasets, expanding the use of synthetic data from purely numerical research to a vast range of fields and applications. These new examples of AI-based synthetic data in research showcase survey responses (Jansen et al., 2023) or synthetic social media posts (Rossi et al., 2023), where Large Language Models (LLMs) or Generative Adversarial Networks (GAN) are used to either augment or recreate real datasets for research purposes.

This change within the research paradigm raises the question of whether this kind of synthetic data could (and should) become a recurrent element in the field of IS research. This inquiry is generating a profound debate around its feasibility, and the actual added value of this type of data (Listgarten, 2024; Noble, 2024). The complexity and variety of elements that compose synthetic data, together with the range of potential sources and methods of generation, make it hard to provide a general answer to this question. Moreover, the polarizing views and the current lack of formalization regarding AI use in research create a situation of conflicting bad practices and under-used potential (Kowalczyk et al., 2023).

As a result, discussions on synthetic data usage tend to overlook peculiarities based on the context and

URI: https://hdl.handle.net/10125/109712
978-0-9981331-8-8
(CC BY-NC-ND 4.0)

objective of the data. *What is the goal of using this data? What are the requirements for that use? Are those requirements fulfilled for that specific dataset?* For example, privacy-focused synthetic data must ensure confidentiality and maintain the meaningful statistical properties of the original dataset (Arnold and Neunhoeffer, 2021), while synthetic data used to evaluate research questionnaires must provide plausible answers and be indistinguishable from those of human participants (Hämäläinen et al., 2023). The use of synthetic data is indeed a multidimensional issue; where ethics, performance, and privacy are just some of the concerns when using artificially generated data (Kowalczyk et al., 2023). In our work, we group all these dimensions under the term data quality (Wang and Strong, 1996).

Overall, the main goal of this paper is to add nuance and structure to the discussion of AI generated (or not) data. To this end, we develop a framework to evaluate the quality of synthetically generated data for research purposes, relying on the context and objectives of its use. By developing this framework, we aim to provide the necessary steps to assess the analytical/critical transparency to the process of generating synthetic data (Dilaver and Gilbert, 2023). That is, assessing the strengths and limitations of data intended for use in IS research, so that researchers can make contributions to the field in line with these assessments.

To build this framework, we explore the potential connections between the generation of synthetic data and the fields of control theory and data quality management. We do so by following a theory adaptation approach in which we implement a "breakout" literature review methodology. As such, we propose a framework structurally based on a closed-loop feedback scheme that uses the main elements of data quality frameworks to assess the data being generated. This approach also facilitates the communication of the properties of the data in a more transparent manner.

The paper is organized as follows. In section 2, we provide a comprehensive theoretical background on the use of synthetic data in research. Section 3 introduces our methodology, while section 4 presents the evaluation framework. Then, section 5 illustrates the use of the framework, and we use section 6 to discuss the results and implications of our work. We conclude the paper with section 7, where we provide the conclusion and elaborate on future work.

2. Synthetic data in research

2.1. Types of Synthetic data

Synthetic data is commonly defined as data that is computationally generated rather than observed and collected from real-world phenomena (Kowalczyk et al., 2022). This definition includes two main sources of generation: simulations and generative model approaches (Steinhoff and Hind, 2024). This type of data can be used to anonymize existing data (Vallevik et al., 2024), to augment datasets (Kowalczyk et al., 2023), or to generate entire samples (Hämäläinen et al., 2023).

Simulations generate data differently based on the modeling approach; either based on a macroscopic representation of the phenomena, commonly ruled by theoretical frameworks, or a microscopic approach modeled by observed individual behavior (Conte et al., 2001). On the other hand, data-driven approaches are more useful in complex scenarios where bias and uncertainty in the modeling of simulation hinder the quality of the results (Eckerli, 2021). In essence, both approaches face the issue of generalization ability in terms of reality gap (Steinhoff and Hind, 2024). That is the transfer and generalization of the findings from the synthetic domain to the real domain. In other words, the challenge in any use of synthetic data is to determine what we can infer about the real world by analyzing it.

2.2. Transparency in synthetic data

New types of synthetic data generation, such as LLMs (Hämäläinen et al., 2023), share similar challenges with more established approaches, such as agent-based or complex systems simulations. These techniques have difficulty assessing the accuracy and reliability of their results, as there is not enough information to compare them to actual systems (Christie et al., 2005), and they can only rely on uncertainty and sensitivity analyses to evaluate the model and its properties (Helton et al., 2007).

Dilaver and Gilbert (2023) describe computational simulations as black-box systems with varying levels of opacity. These levels depend on what is opaque in the simulation process, for whom, and in what sense the simulation is opaque (i.e., which elements are not identifiable or explainable). This description also applies to foundation models and any type of data generation, where the opaqueness of the process and its outcomes varies depending on the phenomenon and the approach.

The generation of data based on partially opaque systems, where only some information of the process and the outcomes is available, resonates with the fields

of control theory and cybernetics (Kuo and Golnaraghi, 1995). Any data generation process can be seen as an opaque system where we apply certain inputs and, based on the observed outcomes, we attempt to modify these inputs to achieve a desired outcome, i.e., useful synthetic data.

Defining and evaluating these outcomes is a wicked problem in itself. Moreover, challenges may vary significantly based on the type of process used to generate the data and its context. Given the level of assessment that is possible from the opaqueness of the system, and drawing from complex systems literature (Christie et al., 2005), we propose to separate synthetic data generation into two types. *Replicative synthetic data* refers to the cases where there is a real dataset that can be compared to in terms of content and statistical properties, e.g., healthcare synthetic data generated for privacy and regulatory concerns (Vallevik et al., 2024). In such cases, the potential for analytical and critical transparency is very high, as the assessment process has a clear reference. *Emergent synthetic data* refers to the cases where, similarly to complex systems, there is not a real dataset or enough knowledge about the real phenomenon to evaluate the data directly, e.g., generation of fake personas (Rossi et al., 2023). In these cases, only statistical inferences or subjective evaluations can be made, reducing the level of critical transparency that can be achieved for that type of data.

2.3. Uses of synthetic data

Different models and types of simulations can help observe various aspects of a given phenomenon, and thus provide different utilities to research (Edmonds, 2017). Again, generative synthetic data can be seen through a similar lens. Rossi et al. (2024) already classified the use of synthetic data from foundation models in research as either data for experiments or data to replicate existing datasets.

These distinctions help to understand the characteristics of the data and identify the important elements accordingly. Large language models fail to provide high-granularity properties, but some macroscopic patterns produced by the analysis of large quantities of data could be useful at certain points in the research process. A similar case occurs in Monte Carlo simulations, which provide a solution to a macroscopic system through simulations of its arbitrarily generated microscopic elements (Lemon and Verhoef, 2016).

Inspired by Rossi et al. (2024) and Edmonds (2017), we propose a new general classification of synthetic data uses in IS research. We define synthetic data based on its research purpose as: speculative, reductionist, and realistic. *Speculative data* is used in the initial steps of the research process, where we do not expect to draw significant conclusions but to validate assumptions, commonly followed by an actual experiment. Hämäläinen et al. (2023) explore the use of LLMs in HCI research to generate open-ended questions to evaluate and pilot potential questions in new experiments. *Reductionist data* aims to provide a limited series of characteristics found in real phenomena, at either a micro or macro level. Data-driven synthetic data for financial markets, such as asset returns (Eckerli, 2021), is an example of generated data that replicates only part of the phenomenon. Lastly, *realistic data* aims to replicate all aspects of real data or real phenomena. The generation of synthetic images of skin lesions and X-rays for medical applications is a example of this type of data (Chen et al., 2021).

Combining these two dimensions—the level of analytical transparency and the research objective—we obtain the classification depicted in Table 1.

3. Methodology

In this conceptual paper, we develop an evaluation framework for the use of synthetically generated data in IS research. We follow a theory adaptation approach (Jaakkola, 2020), in which we examine the literature from various fields and integrate concepts across multiple theoretical perspectives to provide new theoretical lenses and solutions to the source domain.

We apply a breakout literature methodology (Gruner and Minunno, 2023) to evaluate and transfer concepts and methods to the field of synthetic data in research. This approach is grounded in the idea of analogical reasoning (Breslin and Gatrell, 2023) and follows three sequential steps:

Step 1: Identify and review target domain. This step aims to define the target domain, assess its current status, and identify research opportunities within it. As Gruner and Minunno (2023) state, the goal is not to carry out a systematic evaluation of the target domain but to obtain a solid understanding, question common assumptions, and identify potential source domains for a more in-depth literature exploration.

This step is detailed in section 2. In that section, we question the current approach to the use of synthetic data in research, explore various concepts and ideas about synthetic data, and identify potential source domains such as data quality management and technical control theory.

Step 2: Explore source domains and juxtapose knowledge. In this step, we explored and analyzed different elements of data quality management and

	Speculative	Reductionist	Realistic
Replicative	Data that imitates a specific behavior or information of a measurable phenomenon to extract some initial properties. *Example:* Statistical testing data for estimating system reliability (Soltana et al., 2017)	Data that replicates certain characteristics or elements of a known and measurable phenomenon. *Example:* Evolution of asset returns in specific financial markets (Eckerli, 2021)	Data that aims to replicate the exact behavior of a well-known and measurable phenomenon. *Example:* Synthetic images of skin lesions and frontal X-rays (Chen et al., 2021).
Emergent	Data that imitates specific behavior or information of a non-measurable phenomenon to extract initial properties. *Example:* Open-ended questions for preparing HCI experiments (Hämäläinen et al., 2023).	Data that replicates certain characteristics or elements of a non fully measurable phenomenon. *Example:* Evolution of the willingness of citizens to participate in government-led initiatives (Dai et al., 2022).	Data that aims to replicate an exact behavior of a well-known and non fully measurable phenomenon. *Example:* Fake persona social network profile (Rossi et al., 2023).

Table 1. Proposed classification of the different types of synthetic data used in research.

control theory and theorized on how they relate to the generation of synthetic data.

We followed an exploratory literature review to initially evaluate the main concepts of both domains. To do so, we conducted a Scopus search based on title, abstract and keywords using the search terms ("data quality framework") for the literature in data quality and then (("control theory") OR ("closed-loop") OR ("feedback loop") OR ("cybernetics")) AND ("management research" OR "information systems") for control theory. This search yielded 116 and 472 articles, respectively, which were then filtered to 51 articles based on the evaluation of the abstracts. Additional snowballing was performed based on articles relevant to the most fitting concepts. The aim of this step is not to review these fields rigorously (Gruner and Minunno, 2023) but to explore these distant domains to enrich and deepen our understanding of the studied phenomenon.

The second part is to juxtapose knowledge from these source domains in the target domain. For instance, the potential use of synthetic data in research can be associated with the 4Vs of Big Data (Cai and Zhu, 2015). While volume, velocity, and variability are among the advantages of a potential use of synthetic data in research, value represents the main challenge when considering generative AI methods. This idea of value within a specific context can be traced back to the definition of data quality (Cai and Zhu, 2015), usually described as fitness for use (Wang and Strong, 1996).

Traditional methods for assessing the quality of data are applicable beyond traditional contexts. Specifically, traditional data quality approaches are used to assess the quality of training data in machine Learning models (Priestley et al., 2023) with different quality dimensions based on the needs of these models. A similar, more conceptual, adaptation is applied to the evaluation of synthetic data.

Control Theory and Cybernetics focus on the modeling of systems and the potential to control them (Kuo and Golnaraghi, 1995). The generation of synthetic data can be seen as a black-box system where we have a limited number of actions and partial information on the internal states of the system and the final outcome. This opaqueness may complicate the task of assessing the generated data and the determining the steps required to improve its quality.

Step 3: Evaluate and articulate insights. We evaluated the different elements and frameworks from both fields of literature in the context of generating synthetic data for research. The field of data quality management provides concepts such as dimensions, assessment, and measurement of data quality. Cybernetic control schemes offer the structure and sequential actions needed to apply these concepts in the interaction with a given system, in this case, the generator of synthetic data.

These concepts are articulated and presented in the evaluation framework in section 4, together with the challenges associated with their application. The overall research process followed in this paper is illustrated in Figure 1.

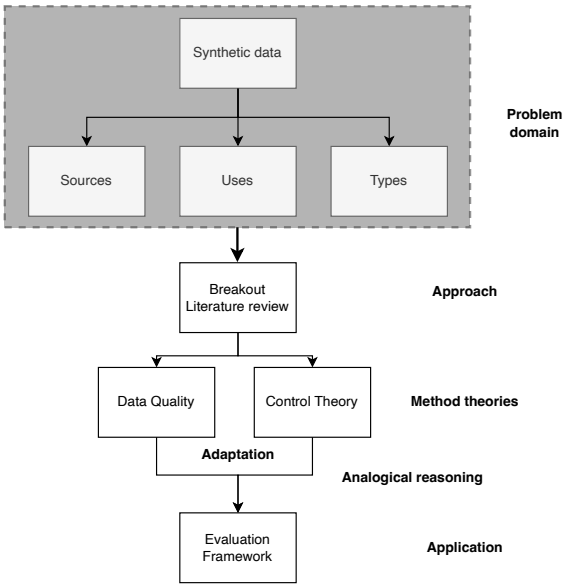

Figure 1. Methodological approach followed in the paper.

4. Evaluation Framework for IS synthetic data

The proposed framework embraces the idea of generating synthetic data as an iterative process, where traditional concepts from data quality management are adapted in a structured and sequential manner. The goal of this work is to provide a framework that is easy to follow and that offers a means to communicate how this process was carried out.

The framework scheme we propose is presented in Figure 2. The process is composed of an input, on the left-hand side of the scheme, and two outputs. The main input represents all information about the research case and the type of data to be generated. The outputs correspond to the synthetic data generated and the misalignment within the generation process, i.e., the difference between the desired properties of the data and the properties actually obtained.

We aim to provide a framework that is general enough to be applied to all types of synthetic data generation (e.g., simulations or generative models). Moreover, we adhere to the data quality literature, which lists the requirements of a comprehensive evaluation framework (Cichy and Rass, 2019) as: i) a definition of relevant quality attributes, ii) data quality assessment steps, and iii) data quality improvement steps.

The block diagram presented in Figure 2 is inspired by the closed-loop control scheme (Kuo and Golnaraghi, 1995). Closed-loop control regulates a system's outcome towards a desired reference by using the difference between the desired and actual outcomes as input to compute feedback loop actions. Following the literature in control theory, the links represent data and information, while the blocks indicate the different processes/agents within the framework.

4.1. Elements of the framework

4.1.1. Research case (Link 1) The first link, and input to the evaluation framework, contains the information associated with the data being generated, such as type of data, objective, etc. This could include survey data that appears realistic and helps prepare experiments (Jansen et al., 2023), fake Twitter profiles with images and text indistinguishable from real profiles (Rossi et al., 2023) or healthcare data that is anonymized yet retains the statistical properties of the original data (Vallevik et al., 2024).

4.1.2. Case evaluation This process involves identifying, based on all the elements of the research case, the dimensions, indicators, and reference values for data quality.

Data quality is a multidimensional concept affected by content and context-related factors. Data quality dimensions are attributes that represent a single aspect or construct of the data quality (Cai and Zhu, 2015) and which will allow to define indicators employed to assess quality. Thus, identifying these dimensions is a fundamental step in establishing the metrics to be evaluated within the framework.

In the case of synthetic data, traditional dimensions of data quality, such as accessibility or temporality (Cichy and Rass, 2019), are not necessarily transferable. Quality attributes can be linked to the classification in Table 1, as the selected type of synthetic data can help researchers identify which dimensions are important for that type of data generation.

In this step, the researcher must justify why these indicators and dimensions are sufficient to validate specific claims about the data being generated (Kowalczyk et al., 2023). Vallevik et al. (2024) identified several aspects such as representativeness, privacy, and fairness, as quality dimensions for healthcare synthetic data for privacy concerns. Hämäläinen et al. (2023) focused on the plausibility of the answers and their distinguishability from real human responses.

4.1.3. Desired data properties (Link 2) This link provides the desired data properties that will assist in evaluating the data being generated. It defines

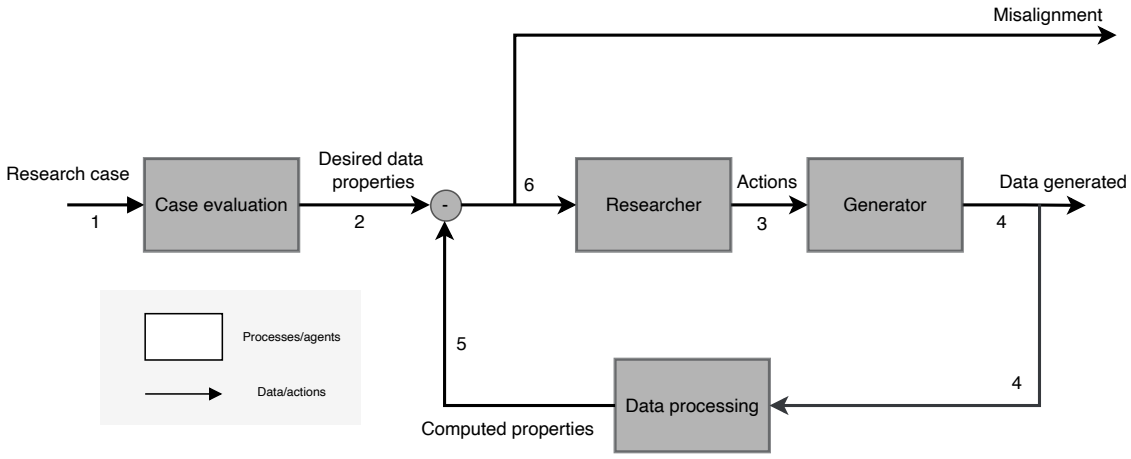

Figure 2. Main components of the evaluation framework.

the elements to be evaluated (similarly to the definition of dimensions and elements in data quality frameworks (Cai and Zhu, 2015)) and the desired values associated with these indicators (equivalent to the reference in control loop frameworks (Kuo and Golnaraghi, 1995)). This link enables the iterative comparison between the outcome and the desired values. Therefore, it is crucial that these indicators can be computed from the generated data.

Often, this type of information is not explicitly provided by researchers. Based on information from some studies, these references might take the following forms: in Hämäläinen et al. (2023) a 75% validity rate in the evaluation of the errors produced by ChatGPT; in Rossi et al. (2023) a similar rating for the similarity of synthetic profiles to the real ones; and in Eckerli (2021) a similar probability density function distribution of the synthetic and real returns.

4.1.4. Researcher Unlike traditional data quality management, where data is collected rather than generated, the active generation of data introduces additional control and complexity for researchers. This complexity requires more structured guidance within the evaluation process. Traditional data quality improvement actions differentiate between data-driven or process-driven actions (Batini et al., 2009). In our framework, data-driven actions are related to parameters and inputs within the interaction with the data generator, e.g training data provided to the GAN (Eckerli, 2021) or prompt information given to an LLM (Rossi et al., 2023). Process-driven actions imply modifications to elements of the framework itself, changing the data generator (e.g., replacing LLM processes for agent-based modeling) or varying the claims associated with the data, thus affecting the case evaluation (e.g., shifting from realistic data to speculative data used to pilot experiments).

Note that while Loshin (2010) differentiates between proactive and reactive actions in data quality improvement, our framework leverages the ability to generate data. Consequently, we conceptualize the evaluation process in an iterative manner, i.e., acting based on the comparison between desired and obtained properties. This contrasts with approaches where actions are performed either a priori or a posteriori.

4.1.5. Actions (Link 3) Based on the type of generator, the number of actions and their impact on the generation process will vary.

For LLMs, these actions might include: the prompts provided to the LLM; the parameters, such as temperature, that control the probability distribution of the output; or the LLM model being used, which could vary in size or be specifically trained/tuned (Hämäläinen et al., 2023). In agent-based modeling, the topology of the network, the behavioral rules associated with the agents or the value parameters of these rules are also elements that will vary the outcome of the simulations (Dai et al., 2022).

4.1.6. Generator This block represents the computational process that generates the synthetic data. That is any model, algorithm, or AI system that will produce synthetic data based on the actions and inputs from the researcher.

In the literature we find different types of synthetic data generators: LLMs to generate survey responses (Jansen et al., 2023), GANs to generate fake persona (Rossi et al., 2023), machine learning to

extend datasets (Kowalczyk et al., 2023) or agent-based modeling simulating the behavior of citizens within a community (Dai et al., 2022).

4.1.7. Data generated (Link 4) Synthetic data is one of the outcomes of the evaluation framework, and it also connects with the data processing step initiating the iterative feedback loop.

The generated data can be textual as in the case of open-ended answers (Hämäläinen et al., 2023), images for fake personas (Rossi et al., 2023) or numerical values representing asset returns (Eckerli, 2021).

4.1.8. Data processing This step involves extracting the computed properties from the generated data. Data quality frameworks differentiate between a measurement—the action of measuring values of data quality dimensions—and assessment, which is the comparison between measurements and reference values (Batini et al., 2009). Data processing represents the action of obtaining these measurements so they can be compared to the desired properties established earlier (assessment). This process can be performed quantitatively, by extracting properties with data analysis techniques (Eckerli, 2021), qualitatively, by using humans to evaluate realism (Hämäläinen et al., 2023) or via quantitative surveys (Rossi et al., 2023).

4.1.9. Computed properties (Link 5) This link represents the properties extracted from the generated data, which should have the same dimensions as those from the desired data properties. This allows for comparison with the previously established reference values.

These properties can be derived using different approaches: from human evaluations, as in Hämäläinen et al. (2023), where the authors compute the percentage of distinguishability between fake and real answers, to similarity scores based on multivariate and univariate distributions (Vallevik et al., 2024).

4.1.10. Misalignment (Link 6) This link results from the comparison between the desired characteristics of the data and the computed properties (assessment). It represents the misalignment between each of the selected indicators. This information is used to guide the researchers' actions and can be used to evaluate the quality of the synthetic dataset.

This link indicates the state of the process after each iteration. If the values are acceptable, the process ends; otherwise, the researchers may take actions based on this misalignment (e.g., changing model parameters, training data, or the model itself).

In Rossi et al. (2023), the misalignment would be 0, as the values obtained associated to the suspiciousness of the fake profiles are within the desired range (equivalent to real profiles). In Eckerli (2021), testing different GANs, the pdf distribution show more or less similarity, providing information regarding the best dataset generated.

5. Applying the framework

To illustrate how the framework could be implemented retrospectively, we use the article by Hämäläinen et al. (2023) as an illustrative example. In their work, Hämäläinen et al. (2023) evaluate the potential use of LLMs to generate synthetic data to ideate and pilot HCI questionnaires. We chose this article thanks to the detailed information provided by the authors. We briefly map their research onto our framework in Figure 3.

The main actions during and between experiments, i.e., changes in the input to the generator, can be seen as iterations of our framework. These changes are based on attempts to improve the quality of responses from ChatGPT, moving from a mean of 64% valid responses to 79% in *Experiment 2*. This can be mapped to our framework by considering 75% of valid responses as reference, which results in a first iteration with a negative misalignment. After modifying the prompt, the data improved to reach the target of 75% valid answers. Similarly, in *Experiment 3*, several LLM models are compared. This can also be seen as iterative actions aimed at improving the values of different metrics such as Frechet distance or precision.

When mapped to our framework, the main elements missing are the explicit definition of desired properties and some additional iterations to improve the outcome. For instance, we observe that the final experiment in Hämäläinen et al. (2023) provides low values in diversity. If applying the framework, additional iterations could be performed to adjust other parameters. Since ChatGPT is a next-word predictor/generator, to achieve diversity one can adjust the parameters that drive this prediction. These actions include: increasing the temperature, which controls the randomness of predictions; lowering the top-P value, which limits token selection to a subset with cumulative probability above a threshold P, allowing the model to explore a broader set of possible next words; setting a higher top-K value, which restricts prediction to the top K most likely tokens (e.g., from 50 to 100); or increasing frequency and presence penalties, which discourages repetition of the same tokens.

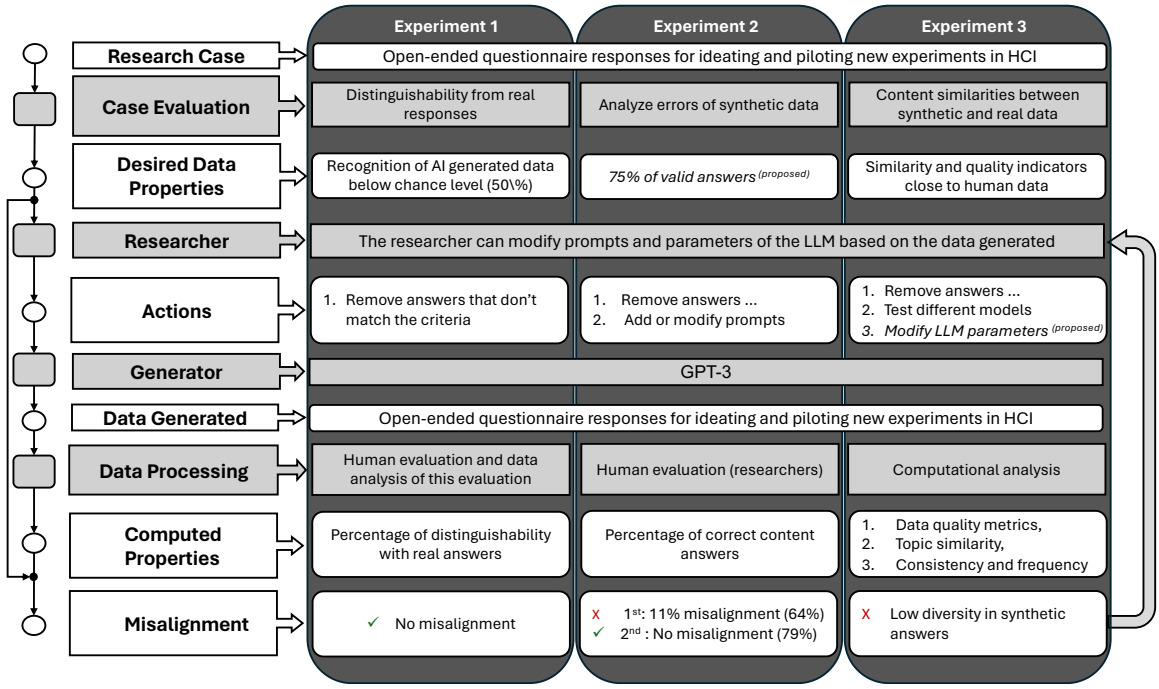

Figure 3. Transparency Table for Synthetic Data Reporting. Data from Hämäläinen et al. (2023).

6. Discussion

The presented framework is based on the idea that synthetic data can enhance the research process. This assumption does not necessarily advocate for all uses of generative AI and synthetic data in research. Instead, it emphasizes the proper evaluation of each use, based on the conditions, objectives, and type of generative process. As such, the framework aims to assist researchers in validating these aspects. The first part—case evaluation and desired data properties—focuses on formalizing the research objectives and potential ethical issues of using synthetic data in a given research context. The classification in Table 1 summarizes the different uses and can help researchers better frame their case study. The second part of the framework—the iterative loop—aims to validate the properties of the synthetic data, including potential biases, inaccuracies, or defects in the generated data that would render it unsuitable for research.

Our work not only provides a means to evaluate the synthetic data being generated but also offers a means to communicate this process and its outcomes to IS journal outlets. As such, the framework ensures transparency regarding the process and its limitations, i.e., analytical transparency (Dilaver and Gilbert, 2023), and transparency regarding the data and its properties. Thanks to this, both aspects can be evaluated by the research community, aligning our work with the ethical guidelines proposed by Rossi et al. (2024). We believe that the main risk of generative tools does not lie in their stochastic nature but in their misuse and lack of formalization within the research process. That said, the framework can be used as a pre-evaluation tool for researchers to potentially discard the use of synthetic data in the early stages of their research.

Nonetheless, this work is subject to two main limitations: the methodological approach used to develop the framework and the capabilities of the framework in its current state.

Based on the work of Gruner and Minunno (2023), we followed a breakout literature review approach to expand the current literature and understanding of the use of synthetic data in research. However, given the non-exhaustive nature of this approach, other domains besides data quality and control theory could be explored to further expand and enrich our contributions.

It is also important to note that the proposed framework, or any framework, cannot overcome the problems inherent to synthetic data or the use of generative AI in research; it helps in their characterization and provides a structured and sequential approach for researchers to interact with these systems. The efficacy of this tool will ultimately depend on the researcher; who must decide, motivate and demonstrate the validity of a certain type of synthetic data in a given research scenario. To reduce this dependency, future iterations of the framework will

focus on guiding researchers to better navigate its various elements. The first step involves clarifying the types of uses of synthetic data to guide researchers on which use could fit their project. This can be achieved by extending the mapping from Table 1 by adding a technological dimension to link specific uses with suitable generation tools.

Another aspect worth mentioning is the current level of detail in the framework. Our aim is to provide a general framework that can be adapted to different data generation tools (whether generative AI or not) and that remains useful as new methods emerge. At this stage, we have presented the main elements of the framework and their roles, providing a low level of specificity constrained by text space and the initial state of our research. This is particularly noticeable in the improvement actions to be carried out based on the misalignment observed after each iteration. Given the variety of generators and potential research uses, such guidelines will be explored in future work, narrowing the scope to more specific use cases. A more detailed characterization of actions available, based on the technology and how they can affect the outcome, will further enhance the applicability and the guidance for researchers using the framework. Additionally, other frameworks focused on specific characteristics of synthetic data, such as the choice of appropriate metrics and dimensions (Dankar et al., 2022), could be integrated into our framework to enrich certain implementations.

To improve the framework and address the stated limitations, future work will focus on assessing its added value. There are several aspects of the framework that could be tested: its applicability, its effectiveness in improving the generation of synthetic data, and its effectiveness in communicating the methodology followed. These elements will be evaluated by applying our framework more extensively to a large variety of real-case scenarios, as shown in Figure 3. This will provide researchers with more practical guidelines on how to use the framework and will help us improve its effectiveness while validating it.

7. Conclusion

In this work, we present a novel framework for the generation and evaluation of synthetic research data across different contexts and purposes. The goal is to present a methodological tool that helps researchers integrate synthetic data in a formal and transparent manner.

We use a breakout literature approach, exploring the domains of data quality and control theory, to establish the foundations of the framework. The framework takes the form of a closed-loop system, leveraging the properties of most generators to compare desired properties with the output, and modify the process accordingly. This framework can be used with any type of synthetic data and provides a high level of transparency in its use within the research process.

8. Acknowledgement

The research pertaining to these results received financial aid from the Belgian Science Policy Office according to the agreement of subsidy no. [B2/223/P3/BeCoDigital].

References

Arnold, C., & Neunhoeffer, M. (2021, October). Really Useful Synthetic Data – A Framework to Evaluate the Quality of Differentially Private Synthetic Data [arXiv:2004.07740 [cs, stat]].

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52.

Berente, N., Seidel, S., & Safadi, H. (2019). Research Commentary—Data-Driven Computationally Intensive Theory Development. *Information Systems Research*, *30*(1), 50–64.

Breslin, D., & Gatrell, C. (2023). Theorizing Through Literature Reviews: The Miner-Prospector Continuum. *Organizational Research Methods*, *26*(1), 139–167.

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*(0), 2.

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*(6), 493–497.

Christie, M. A., Glimm, J., Grove, J. W., Higdon, D. M., Sharp, D. H., & Wood-Schultz, M. M. (2005). Error Analysis and Simulations of Complex Phenomena. (29).

Cichy, C., & Rass, S. (2019). An Overview of Data Quality Frameworks. *IEEE Access*, *7*, 24634–24648.

Conte, R., Edmonds, B., Moss, S., & Sawyer, R. K. (2001). Sociology and Social Theory in Agent Based Social Simulation: A Symposium. *Computational & Mathematical Organization Theory*, *7*(3), 183–205.

Dai, L., Han, Q., & de Vries, B. (2022). Agent-Based Simulation of Citizen Participation in Nature-Based Projects. *SSRN Electronic Journal*.

Dankar, F. K., Ibrahim, M. K., & Ismail, L. (2022). A Multi-Dimensional Evaluation of Synthetic Data Generators. *IEEE Access*, *10*, 11147–11158.

Dilaver, O., & Gilbert, N. (2023). Unpacking a Black Box: A Conceptual Anatomy Framework for Agent-Based Social Simulation Models. *Journal of Artificial Societies and Social Simulation*, *26*(1), 4.

Eckerli, F. (2021). Generative Adversarial Networks in finance: An overview. *SSRN Electronic Journal*.

Edmonds, B. (2017). Different modelling purposes. *Simulating social complexity: A handbook*, 39–58.

Gruner, R. L., & Minunno, R. (2023). Theorizing across boundaries: How to conduct a 'breakout' literature review. *International Journal of Management Reviews*, ijmr.12356.

Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19.

Helton, J., Johnson, J., Oberkampf, W., & Storlie, C. (2007). A sampling-based computational strategy for the representation of epistemic uncertainty in model predictions with evidence theory. *Computer Methods in Applied Mechanics and Engineering*, *196*(37-40), 3980–3998.

Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, *10*(1-2), 18–26.

Jansen, B. J., Jung, S.-g., & Salminen, J. (2023). Employing large language models in survey research. *Natural Language Processing Journal*, *4*, 100020.

Kowalczyk, P., Röder, M., Rottmann, J., & Thiesse, F. (2023). Designing a method to nudge analytics with artificially generated data. *ICIS 2023 Proceedings. 4.*

Kowalczyk, P., Welsch, G., & Thiesse, F. (2022). Towards a Taxonomy for the Use of Synthetic Data in Advanced Analytics [Publisher: arXiv Version Number: 1].

Kuo, B. C., & Golnaraghi, M. F. (1995). *Automatic control systems* (Vol. 8). Prentice hall Englewood Cliffs, NJ.

Lemon, K. N., & Verhoef, P. C. (2016). Understanding Customer Experience Throughout the Customer Journey. *Journal of Marketing*, *80*(6), 69–96.

Listgarten, J. (2024). The perpetual motion machine of AI-generated data and the distraction of ChatGPT as a 'scientist'. *Nature Biotechnology*, *42*(3), 371–373.

Livieris, I. E., Alimpertis, N., Domalis, G., & Tsakalidis, D. (2024, April). An evaluation framework for synthetic data generation models [arXiv:2404.08866 [cs]].

Loshin, D. (2010). *The practitioner's guide to data quality improvement*. Elsevier.

Noble, W. S. (2024). Response to "The perpetual motion machine of AI-generated data and the distraction of ChatGPT as a 'scientist'". *Nature Biotechnology*.

Priestley, M., O'donnell, F., & Simperl, E. (2023). A Survey of Data Quality Requirements That Matter in ML Development Pipelines. *Journal of Data and Information Quality*, *15*(2), 1–39.

Rossi, S., Kwon, Y., Auglend, O., Mukkamala, R., Rossi, M., & Thatcher, J. (2023). Are deep learning-generated social media profiles indistinguishable from real profiles? *Proceedings of the 56th Hawaii International Conference on System Sciences*, 134–143.

Rossi, S., Rossi, M., Mukkamala, R. R., Thatcher, J. B., & Dwivedi, Y. K. (2024). Augmenting research methods with foundation models and generative AI. *International Journal of Information Management*, 102749.

Soltana, G., Sabetzadeh, M., & Briand, L. C. (2017). Synthetic data generation for statistical testing. *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 872–882.

Steinhoff, J., & Hind, S. (2024). Simulation and the Reality Gap: Moments in a prehistory of synthetic data.

Vallevik, V. B., Babic, A., Marshall, S. E., Elvatun, S., Brøgger, H. M., Alagaratnam, S., Edwin, B., Veeraragavan, N. R., Befring, A. K., & Nygård, J. F. (2024). Can I trust my fake data – A comprehensive quality assessment framework for synthetic tabular data in healthcare. *International Journal of Medical Informatics*, *185*, 105413.

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33.