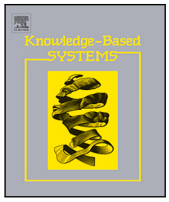
Knowledge-Based Systems 286 (2024) 111440

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Generation of probabilistic synthetic data for serious games: A case study on cyberbullying

Jaime Pérez [a,*], Mario Castro [a,b], Edmond Awad [c,d,e], Gregorio López [a]

[a] Institute for Research in Technology (IIT), ICAI Engineering School, Universidad Pontificia Comillas, Madrid, 28015, Spain
[b] Grupo Interdisciplinar de Sistemas Complejos (GISC), Madrid, 28015, Spain
[c] Department of Economics, University of Exeter, Exeter, EX4 4PU, United Kingdom
[d] The Oxford Uehiro Centre for Practical Ethics, University of Oxford, Oxford, OX1 1PT, United Kingdom
[e] Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, 14195, Germany

ARTICLE INFO

Keywords:
Synthetic data
Serious games
Cyberbullying
Item response theory
Bayesian network
Hierarchical Bayesian model
Computational social science

ABSTRACT

Synthetic data generation has been a growing area of research in recent years. However, its potential applications in serious games have yet to be thoroughly explored. Advances in this field could anticipate data modeling and analysis, as well as speed up the development process. To fill this gap in the literature, we propose a simulator architecture for generating probabilistic synthetic data for decision-based serious games. This architecture is designed to be versatile and modular so that it can be used by other researchers on similar problems (e.g., multiple choice exams, political surveys, any type of questionnaire). To simulate the interaction of synthetic players with the game, we use a cognitive testing model based on the Item Response Theory framework. We also show how probabilistic graphical models (in particular, Bayesian networks) can introduce expert knowledge and external data into the simulation. Finally, we apply the proposed architecture and methods in the case of a serious game focused on cyberbullying. We perform Bayesian inference experiments using a hierarchical model to demonstrate the identifiability and robustness of the generated data.

1. Introduction

The Internet has become an integral part of young people's lives. Minors under 18 years old already accounted for nearly one-third of Internet users worldwide in 2017, according to UNICEF's "Children in a Digital World" report [1]. However, uncontrolled access to the Internet also opens the door to new threats targeted toward minors, making them more accessible to bullies, harassers, and sex offenders. Around 10% of European children are already victims of cyberbullying (CB) every month [2], and 49% have experienced a CB-related situation at least once [3].

Traditionally, law enforcement agencies and policymakers have focused their efforts on addressing the CB issue from the criminal component. This research is part of the European research project H2020 RAYUELA[1] [4], which aims to leverage the natural appeal of a *Serious Game* (SG) to use it as an educational and research tool to study CB, thus fostering a preventive approach. Specifically, the project aims to understand better which factors influence risky online behavior in a friendly, safe and non-invasive way. Players are immersed in a SG where they must make decisions involving potentially hazardous cybercrime-related situations.

SG are tools designed for purposes beyond pure entertainment (e.g., educational, training, awareness, marketing) [5]. They have gained prominence in recent years in research, industry, and education [6–8], offering immersive and interactive experiences to users. The idea of using games as a research tool to investigate humans is not new and has been gaining popularity in recent years. For example, an experiment embedded in a video game showed that the complexity of the city where a child lives influences his or her future navigation skills [9]. A video grammar game called "*Which English?*" probed the existence of a "critical period" for learning a second language that extends into adolescence [10]. *"The Moral Machine"* experiment, a dilemma-based game involving millions of people, explored the moral values of our societies and how they vary between countries [11].

As the demand for SG increases, so does the need for diverse and realistic datasets to improve their development and evaluation. Synthetic data is a good candidate to address some of these challenges. For example, it can help with data privacy, fairness and augmentation, compensate for data deficiencies such as category imbalance or even produce data before the real one is available [12]. Although synthetic

* Corresponding author.
E-mail address: jperezs@comillas.edu (J. Pérez).
[1] https://www.rayuela-h2020.eu/.

https://doi.org/10.1016/j.knosys.2024.111440
Received 25 May 2023; Received in revised form 10 January 2024; Accepted 22 January 2024
Available online 23 January 2024

data is not a replacement for real data, it can accelerate the SG development process and facilitate advanced data modeling and analysis [13]. In recent years, interest in using synthetic data in social or behavioral science research has also notably increased [14–16].

This paper's primary goal is to present a methodology for generating synthetic data for decision-based SGs, such as the one used in the RAYUELA project. Although the methodology presented can be used to generate synthetic data in any decision-making scenario (e.g., multiple choice exams, political surveys, any type of questionnaire, or games/simulations). To this end, we propose a simulator architecture and bring two innovations to the state of the art. First, we present a generic methodology to introduce external data to the simulator through probabilistic graphical models, particularly Bayesian Networks (BN) [17]. BN modeling has become a popular tool in recent years [18], including some examples in CB research [19,20]. Second, the model that mimics player behavior is based on the Item Response Theory (IRT) cognitive modeling framework. This paradigm has been extensively studied in the literature and proven far superior to classical test theory [21–23].

Through Bayesian inference experiments employing a hierarchical model, we demonstrate the identifiability and robustness of synthetic data, showing the potential of our approach in generating high-quality datasets for serious game development. This paper provides a practical solution to the challenge of synthetic data generation for SG, and lays the groundwork for future exploration and refinement of methodologies at this emerging intersection of research domains.

2. State of the art

Interest in synthetic data has been increasing over the last few years, offering a solution to the challenges associated with limited or inaccessible real data sets. Moreover, this data may be difficult, expensive, or unethical in many domains. Conceptually, synthetic data have similar statistical properties to real data. If an analyst works with a synthetic dataset, the expectation is that the analysis outcomes should closely resemble those derived from real data. This section provides an overview of the current state of the art in synthetic data generation, encompassing main approaches and methods regardless of the application domain. There are three types of synthetic data depending on their generation process. The first type is generated from actual data, the second type does not use real data, and the third type is a hybrid of these two [24].

2.1. Synthesis from real data

The methods included in this subsection are also known as data augmentation. The intuition is that synthetic data can act as a regularizer, thus reducing variance in the final model. The goal of data augmentation may include addressing data imbalance, improving the generalization and robustness of data-driven models, reducing overfitting, or preserving user privacy [25].

Classical statistical imputation methods (e.g., SMOTE, ADASYN) are widely used in unbalanced datasets. However, their capabilities are very limited in replicating complex relationships between variables. Also widely used are those approximations known as heuristics, such as linear or geometric transformations to the data [26,27].

In recent years, sophisticated machine and deep learning techniques have begun to be used to capture particularly complex relationships between variables. Within this category, techniques such as Variational Autoencoder (VAE) [28,29], Generative Adversarial Networks (GAN) [30,31], or diffusion models [32,33] are achieving the greatest success. Recent advances in generative AI promise significant advances, although special care must be taken to ensure that models do not collapse due to self-consuming loops [34,35]. Numerous updated surveys address the usefulness of data augmentation techniques depending on the data type, whether time series [36,37], images [38], or text [39].

2.2. Synthesis without real data

This type of synthetic data covers generation methods that do not use real data. Instead, it uses computational models describing known behaviors or expert knowledge to generate the synthetic samples. Simulators are used in the most complex cases. They can be, for instance, gaming engines creating synthetic scenes that obey a set of specific rules (e.g., physics laws, production line processes, financial market behavior, board game rules).

Over the last few years, it has been proven the great potential of using simulators to train highly advanced AI models based on Reinforcement Learning such as AlphaZero [40]. Furthermore, it has been used in developing robots, since it enables the algorithms to train for thousands of hours in realistic simulations, subsequently improving their performance in the real world [41]. The concept of Digital Twin is applied when the aim is to computationally mimic specific facilities, operational processes, or physical products [42].

2.3. Hybrid synthesis

This type of synthetic data combines methods from the other two groups to generate data that not only replicates the statistical characteristics of real-world data but also incorporates domain-specific insights and expertise. The generation process usually starts with an existing real dataset, and then domain experts contribute their insights to the generation process. This may involve incorporating known patterns, relationships, or nuances that purely data-driven approaches might not fully capture [24].

Simulations play also a crucial role in this approach by generating scenarios that may not be well represented in the existing data [43,44]. The synergy between data-driven augmentation and expert-guided simulations results in a hybrid synthetic dataset with a more complete and nuanced representation of the underlying domain [45].

Hybrid synthetic data generation is most commonly used in specialized fields where expert knowledge is essential, but we also have some external data from which we can learn. For example, in specialized industrial processes [46,47] or medical systems [48,49]. Some proposals in the literature already propose using BNs to generate synthetic data, as they are a convenient approach to merging expert knowledge and data [50–52].

The work developed in this paper fits into the hybrid synthesis category, since we will use existing external data and expert knowledge to enrich the simulation. We contribute to this field by proposing a modular architecture to generate synthetic data for an iterative decision-making serious game. Unlike other agent-based simulations, the goal is not to "win" the game but to replicate realistic human behavior while playing.

3. Simulator

3.1. Design considerations

Before detailing the proposed simulator architecture, we will review the design considerations and project constraints that led to the decisions made. Firstly, although the proposed architecture can be applied to other environments where participants must make a series of decisions or answer categorical questions, this work focuses on the specific problem of an interactive narrative serious game. Besides, it is noteworthy that in our work, the synthetic players do not aim to "win" the game but to approximate realistic human behaviors, in an approach more similar to [53,54].

To ensure that the synthetic data better reflects reality, it is desirable to be able to introduce external information into the generative process (e.g. expert knowledge, surveys, prevalence data, etc.). It is also desirable to do this in a generic way, so that it is easy to experiment and introduce additional data at any time, and so that the proposed

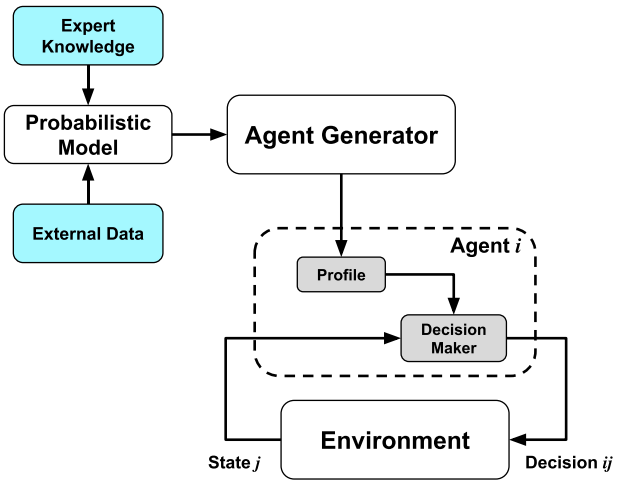


Fig. 1. Simulator's overall architecture and components. The generated *agents* respond to the *environment* and make a decision according to its *profile* in a non-deterministic manner. The blue boxes represent external information that is fed into the simulation. The gray boxes represent the internal states and models of the synthetic agents.

architecture can be used to address issues other than CB. To meet this design need, we propose using probabilistic graphical models, particularly BN [17]. This model is a powerful visual and quantitative tool for expressing probabilistic relationships among variables. BNs consist of a direct acyclic graph (DAG) structure that encodes which variables are causally related to others, and each node of the network contains a conditional probability table (CPT). The structure and parameters of a BN can be learned from data, manually constructed (usually with the help of experts in the specific problem being addressed), or a combination of both. In addition, a trained BN can be used to generate synthetic data by sampling from the learned probability distributions.

The ultimate goal of RAYUELA's serious game is to identify different groups/clusters of players through the answers collected. In other words, to investigate whether the answers given in the serious game provide information about the players' behaviors in the real world. We can model this environment as a sequence of multi-choice questions, where each player's latent state changes the probability of choosing each option. This paper aims to generate synthetic data reflecting the internal states of the players and their cognitive decision-making process. Therefore, the proposed simulator must have a "decision maker" module that obtains the probabilities of choosing each answer from a given player profile and a question. To meet this design need, we propose using the IRT framework [55], a testing theory based on the idea that the probability of a correct response to an item is a mathematical function of the respondent and item parameters. IRT is often regarded as superior to classical test theory [56], primarily because in addition to inferring the "ability" of the participant, it also takes into account the "difficulty" of each question when assessing (and other possible parameters in complex models). Our work will use IRT to generate synthetic data rather than for statistical inference. With this approach, we achieve to model the interactions between players and the in-game decisions they have to confront based on widely used psychological theories.

In summary, our proposed simulator models players' decision processes probabilistically using a widely used test theory (IRT), while incorporating expert knowledge and external data (e.g. surveys, prevalence data, etc.) through the use of BN.

3.2. Architecture

Considering the technical and design considerations outlined in the previous subsection, we summarize the proposed overall architecture of the simulator in Fig. 1. This modular architecture allows tweaking specific features of the simulator, thus avoiding future significant redesigns (e.g., if we want to create a new agent model for other case studies, we would only have to modify that module). In particular, we have designed the simulator architecture to contain four components: the probabilistic model, the agent generator, the agent, and the environment (see Fig. 1), which we describe in the following subsections.

3.2.1. Probabilistic model

This module is responsible for incorporating expert knowledge and other external information (e.g., surveys or prevalence data) into the simulator, thus aligning the synthetic data with reality and making it as helpful as possible. This is achieved using a probabilistic model, such as BNs, where the expert knowledge is encoded into the network's DAG structure, and network parameters (CPTs) are learned from external information. External information can also be incorporated to define the prior belief probability distributions. Moreover, BNs allow us to *interrogate* the model using "What if...?" questions to obtain quantifiable responses for events for which we have little or no data.

We propose using a trained BN to generate synthetic data that the Agent Generator module will use to produce synthetic players with an individualized profile (in a probabilistic way). In addition, the synthetic data generated by the BN will be incorporated into the final synthetic dataset to make it more informative and helpful. If we desire more control over the generation, we can condition chosen variables (e.g., Age = *18*, Gender = *Male*) to produce *stratified* synthetic data.

3.2.2. Agent generator

This module generates synthetic agents with distinct parameters representing varied profiles (e.g., psychological or sociological profiles). The output of the Probabilistic Model (i.e., probabilities of the variable of interest) drives this generation process. Although the exact transformation process to obtain each synthetic player's profile is a design decision that will change drastically depending on the issue addressed and the number of profiles desired. It, therefore, allows for controlled generation at the service of researchers (e.g., generating an intentional imbalance in the synthetic data that more closely captures reality). Section 4 will explain in detail the implementation we have done for our case study on CB.

3.2.3. Agent

This module aims to re-create the interaction of the synthetic agents with the simulator questions/dilemmas, obtaining as output the answers/decisions taken according to their profile (in a non-deterministic way). Two main components constitute the Agent module:

(i) *Profile* (α_i): This is a fixed internal parameter, unique for each agent, representing the agent's profile. This numerical value is inherited from the Agent Generator module. For instance, in our case study on cyberbullying, the profile parameter will represent the risk propensity of each agent. Positive values of α_i would represent more risk-prone agents, and negative values represent agents with lower risk propensity. Values of α_i around zero represent a random player.

(ii) *Decision maker*: This submodule will simulate the decisions made by the agents in the game, according to the profile and question parameters, trying to align them probabilistically (thus capturing the uncertainty in human decision-making). The implementation is common to all agents.

The approach implemented in the Decision Maker module borrows ideas from the IRT paradigm. However, some adjustments must be performed to make the model properly fit our particular case. As we explained before, the ultimate goal of our project's serious game is to identify different groups/clusters of players through the answers

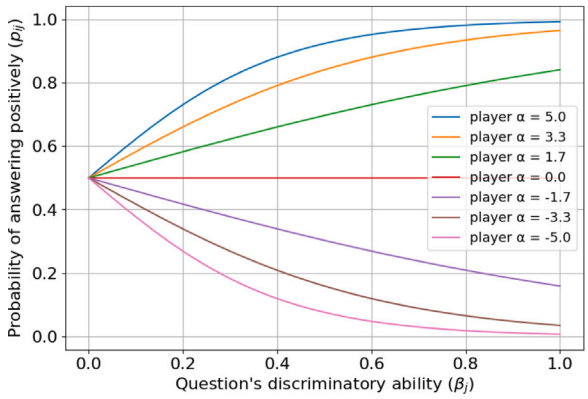


Fig. 2. Visualization of the probability p_{ij} equation values as a function of α and β for the dichotomous/binary case. The values of the equation are shown for 7 values of α, represented in different colors.

collected. Therefore, there will not be correct or incorrect answers, but answers representing greater alignment with certain profiles.

In the simplest case, where agents will make dichotomous choices (i.e., two possible answers) can be formally expressed as Eq. (1). The answers of each player i to each question j are random samples from a Bernoulli probability distribution, with a probability p_{ij} that depends on the agent's profile $\alpha_i \in \mathbb{R}$ and the question parameter $\beta_j \in [0, 1]$, for $i \in [0, N]$ players and $j \in [0, Q]$ questions. Eq. (1) is valid for dichotomous/binary questions, but it can be generalized to multiple choices questions by replacing the Bernoulli with a Categorical probability distribution and using a polytomous IRT-based model in the probability computation [57].

$$\text{Answer}_{ij} \sim \text{Bernoulli}(p_{ij}), \text{ with } p_{ij} = \frac{1}{1 + e^{-\alpha_i \beta_j}} \tag{1}$$

The question parameter β_j is a numerical value unique for each question and represents its discriminatory ability to extract valuable information related to the agent's profile. It is inherited from the Environment module. A value of $\beta_j = 0$ represents null information given by the question, and a value of $\beta_j = 1$ represents perfect information. Namely, answering positively to a question with $\beta_j = 1$ provides a good measurement of the agent's profile (α_i). In actual games, a question with a value close to $\beta_j = 0$ will represent those decisions whose answer is unrelated to the variable of interest (for instance, in our case study, it would be unrelated to CB).

Fig. 2 illustrates how the probability p_{ij} of answering a question positively varies depending on the values of α and β. When $\beta \to 1$ (i.e., high discriminatory ability), agents have a high probability of choosing the response that matches their profile. However, when $\beta \to 0$ (i.e., low discriminatory ability), each agent has a probability that tends to 0.5, regardless of the value of their individual α_i. A random player ($\alpha_i = 0$) will always answer randomly, regardless of the question or β_j value.

3.2.4. Environment

This module simulates the game's narrative structure and is the component with which the synthetic agents interact. It provides the beta values of the questions to the agent module. In interactive narrative games, the internal structure of the scenarios and questions accessed by the player is in the form of a tree. Each node of the tree provides the possible choices that the agent can make in each question/situation of the game. As explained in the previous section, the β_j parameter of the questions indicates its discriminatory ability to extract valuable information related to the agent's profile. During the simulation, these parameters are sampled from a probability distribution with values between 0 and 1 (e.g. Beta distribution).

3.3. Generation process

The proposed method to generate informed synthetic data using a BN can be summarized in the following steps. A graphical representation of these steps is shown in Fig. 3.

1. Build the BN structure (i.e., DAG) from expert knowledge.
2. Train the BN with external data to learn the parameters (i.e., CPTs) using a learning algorithm such as Maximum Likelihood Estimator or Expected Maximization [58].
3. Sample synthetic data from the BN using a sampling algorithm such as Bayesian Model Sampling or Gibbs sampling [59], yielding the characteristics that define each agent.
4. Check the value of the variable of interest (in our case study, having experienced a CB-related situation) to determine the profile of each synthetic player/agent.
5. Sample the profile value (α_i) of each agent according to whether they belong to the group of risky or safe players.
6. Sample the environment values (β_j).
7. Obtain the answers of the agent i using the IRT model (Eq. (1)).

4. Case study: Serious game on cyberbullying

In this section we will explain how we applied the proposed simulator architecture to the serious game of the RAYUELA project [4]. The game is an interactive narrative focused on CB, aiming to identify different groups/clusters of players through the collected responses. Specifically, to differentiate between risky and safe players regarding their online behavior.

Following the proposed architecture (Fig. 1) and in order to generate synthetic data more faithful to reality, we will use a BN to introduce expert knowledge and external data into the simulation. The external data consists of a survey of minors in schools in Spain (Madrid and Valencia) during the year 2022. We collected 665 responses from students between 13 and 17 years old (Mean = 14.5, SD = 0.9), where 50.8% identified themselves as males, 47.4% as females, and 1.8% as non-binary. In this survey, we collected a series of demographic data, some questions about the participants' relationship with new technologies (e.g., IoT devices) and the Internet, and finally, some inquiries about situations related to CB or cyber-harassment. Table 1 shows a random sample of 5 survey participants.

The H2020 RAYUELA project,[2] in which this research is framed, consisted of an interdisciplinary team including psychologists and anthropologists with expertise in CB. The development of the BN structure employed in this case study (Fig. 4) results from iterative discussion and meticulous research conducted in the RAYUELA project through a collaborative effort. The network structure encodes the causal relationships between the variables collected through the survey and how they affect the likelihood of experiencing CB-related events.

The BN is trained using the Expected Maximization algorithm [60], a de-facto standard due to its ability to deal with missing data, being this a pervasive problem in serious game or social science research. GeNIe Modeler[3] software was used to construct and train the BN. Uniform prior probability distributions were set in all the BN nodes to maintain a neutral stance and minimize possible biases. This deliberate choice was intended to ensure that the subjective beliefs of the researchers did not unduly influence the training procedure.

Once the BN has been trained, we begin to generate synthetic data using the Bayesian Model Sampling algorithm to finally obtain a binary probability distribution on the variable of interest (i.e., *having experienced CB related situations*) for each synthetic agent. This probability on

[2] https://www.rayuela-h2020.eu/.
[3] https://www.bayesfusion.com/genie/.

Table 1
External data: Sample of 5 randomly selected survey participants. The last column, "Having experienced situations related to cyberbullying in the last year" is an aggregation of 3 questions in the survey about specific cyberbullying-related situations.

Gender	Age	Sexual orientation	Immigrant background	Daily hours on the Internet for leisure	Cyberbullying concern (1–5)	Family communication on cyber-threats (1–4)	Having experienced situations related to cyberbullying in the last year (Aggregated)
Female	13	Bisexual	No	Between 3 h and 4 h	5/5 (very concerned)	2/4 (rarely)	No
Male	14	Heterosexual	No	Between 1 h and 2 h	2/5 (unconcerned)	3/4 (often)	No
Male	16	Bisexual	No	Between 3 h and 4 h	4/5 (concerned)	3/4 (often)	No
Female	14	Heterosexual	No	More than 4 h	5/5 (very concerned)	4/4 (very often)	Yes
Female	16	Heterosexual	Yes	Between 3 h and 4 h	4/5 (concerned)	3/4 (often)	No

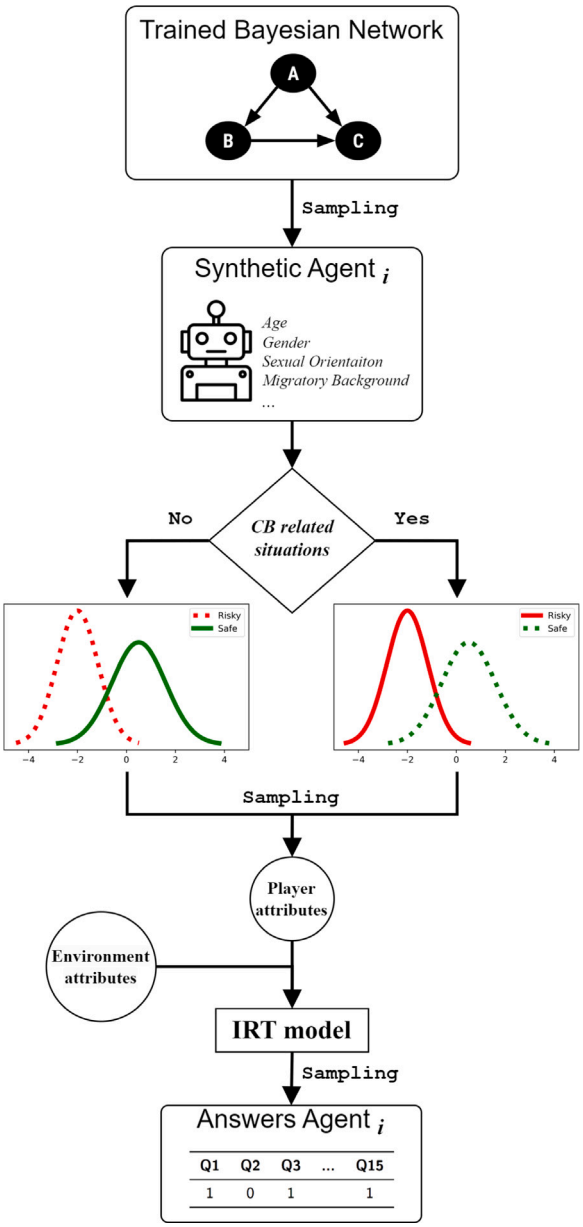


Fig. 3. Conceptual graphical representation of the steps involved in the synthetic data generation. The elements of the generator architecture interact to produce a synthetic dataset representing decisions (0 or 1 in this case) representative of the modeled problem.

the variable of interest will condition whether the agent belongs to the risky or safe group/cluster and, therefore, the numerical value of its risk profile (α_i).

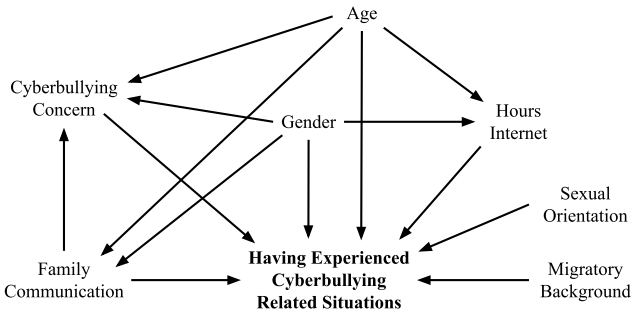


Fig. 4. Probabilistic Model: Bayesian Network structure which encodes the experts' hypotheses of causal relationships among the variables collected in the survey to minors.

For these two groups/clusters of players (risky and safe), we have made the assumption that their risk profiles (α_i) are samples of two different Gaussian distributions. Allowing for some overlapping to account for the intrinsic uncertainty underlying human decision-making processes.

Once the BN has been trained with the survey data (i.e., external information) and we have defined the procedure for using the obtained probabilities to obtain the agents' risk profiles, we can start generating synthetic data. We have created a dataset of 500 synthetic players participating in a game simulation of 15 dichotomous/binary questions. For this case study, we have defined the hyperparameters for the Gaussian distributions as described in Eq. (2). Note that the specific values we gave the hyperparameters are just an example that we believe is somewhat realistic. However, real players may behave differently, or the risky/safe groups may not even exist in the real world.

$$\begin{aligned} \alpha_i|\text{safe} &\sim \text{Normal}(\mu = -2, \sigma = 0.7) \\ \alpha_i|\text{risky} &\sim \text{Normal}(\mu = 0.5, \sigma = 1.2) \end{aligned} \tag{2}$$

Table 2 shows 5 samples of the generated synthetic dataset, where each agent is stored in a row, and also includes synthetic personal information obtained from the BN (age, gender, sexual orientation, immigrant, hours of internet use, CB concern and family communication about cyber-threats). In the columns of the dataset where each agent's answers are stored (Q1 to Q15), the 1s mean that the agent chose the option implying the highest risk propensity. And the opposite with the 0s, the agent has chosen the option implying the lowest risk propensity. Fig. 5 shows a histogram of the generated agents' risk-profile (α_i) parameters. The bimodality reflects the fact that α_i comes from two different Normal distributions and the asymmetry because the incidence of risky profiles in the survey data is lower than 50%.

4.1. Identifiability analysis

Once we have generated the synthetic data, we will perform an empirical identifiability analysis to ensure it is possible to estimate the parameters' values used in the simulator just from the generated data ($N = 500$ players, $Q = 15$ questions). Specifically, we will use a Bayesian hierarchical model with the same structure used to produce

Table 2
Synthetic data: Sample of 5 agents randomly selected from the dataset generated ($N = 500$ agents). Columns Q1 to Q15 indicate the questions of the simulation that has been created for this example. In those, the 1s means that the agent has chosen the option implying the highest risk propensity, and the 0s mean that it has chosen the option implying the lowest risk propensity.

Risk profile (α_i)	Q1	Q2	Q3	...	Q13	Q14	Q15	Gender	Age	Sexual orientation	Immigrant background	Daily hours on the Internet for leisure	Awareness cyberbullying (1–5)	Family communication on cyber-threats (1–4)
−2.16	0	1	0	...	0	0	0	Male	13	Heterosexual	No	Between 1 h and 2 h	4/5 (concerned)	3/4 (often)
1.69	1	0	1	...	0	1	0	Female	16	Heterosexual	No	More than 4 h	2/5 (unconcerned)	1/4 (never)
0.42	1	0	1	...	1	1	0	Female	14	Heterosexual	No	Between 2 h and 3 h	4/5 (concerned)	3/4 (often)
−1.4	1	0	0	...	1	0	1	Female	14	Heterosexual	Yes	Between 2 h and 3 h	5/5 (very concerned)	4/4 (very often)
1.03	0	0	0	...	0	1	1	Non-binary	17	Non-heterosexual	No	Between 2 h and 3 h	5/5 (very concerned)	1/4 (never)

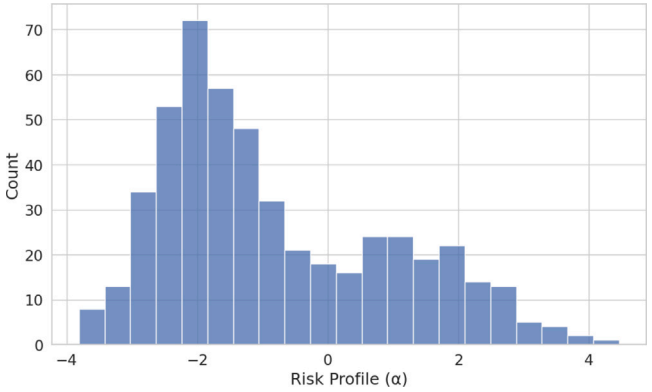


Fig. 5. Histogram of the risk profiles (α_i) parameters of the synthetic generated dataset ($N = 500$ agents). Lower α values encode agents with lower risk propensity and vice versa.

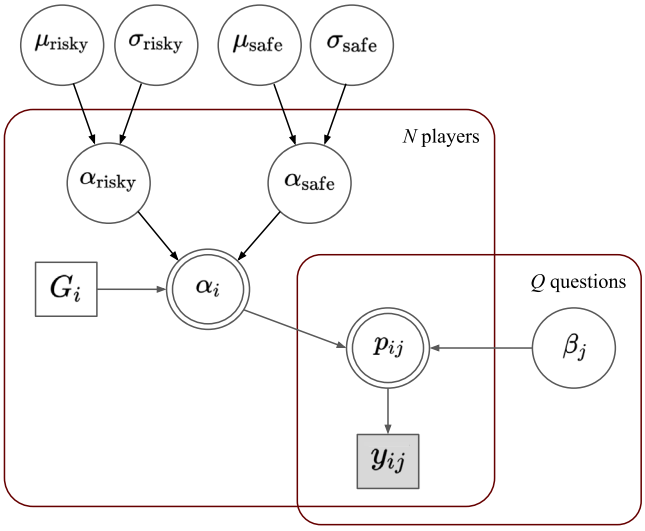


Fig. 6. Probabilistic graphical model of the hierarchical Bayesian model. Circular nodes represent continuous variables and square nodes discrete ones. Double-bordered nodes represent deterministic variables. Shaded nodes represent observed variables.

the data (described in the previous subsection) to estimate the hyperparameters defining the agents and the questions' β_j parameters. We will use only the synthetic responses ($Q_1 \dots Q_{15}$) to train the Bayesian hierarchical model, as these are the data generated with the parameters we want to estimate. So, we will not use the synthetic data generated through the BN (e.g., gender, age, sexual orientation) to reconstruct these parameters. Eq. (3) describes the prior distributions introduced in the hierarchical Bayesian model, and Fig. 6 shows the graphical representation. The parameter p_{ij} is described in Eq. (1).

As depicted in Fig. 6, the risk value was specifically modeled under the paradigm of *Latent Mixture Models* [61,62]. This type of modeling assumes that the observed data are generated by two distinct processes that combine, and specific crucial properties of this combination remain unobservable or latent. In our context, these processes correspond to "risky" and "safe" player behaviors, with the latent variable being the group membership of each player. In essence, we assume that players can solely originate from two hypergroups, and their group membership is treated as a latent variable. This modeling strategy enables the inference of the probability of each player belonging to either the "risky" or "safe" hypergroup.

The validation of this Latent Mixture Model in our case study consists of inferring the value of the risk variable in each player using only the simulated responses in the synthetic dataset. If these estimates are sufficiently accurate, we can say that the synthetic dataset is identifiable and that the synthetic data generation process is successful.

In Bayesian inference, unlike in the Machine Learning or Deep Learning fields, we do not get a singular value due to the prediction; we obtain posterior probability distributions as a result. These distributions represent the epistemic uncertainty about the inferred statistical parameter conditional on the collection of observed data. We have made the implementation using the open-source library *PyMC* [63], a state-of-the-art software tool for probabilistic programming and statistical computation.

$$
\begin{aligned}
\mu_{\text{safe}} &\sim \text{Normal}(-1, 2) \\
\sigma_{\text{safe}} &\sim \text{Exponential}(1) \\
\mu_{\text{risky}} &\sim \text{Normal}(1, 2) \\
\sigma_{\text{risky}} &\sim \text{Exponential}(1) \\
G_i &\sim \text{Bernoulli}(0.5) \\
\alpha_i &\leftarrow \begin{cases} \text{Normal}(\mu_{\text{safe}}, \sigma_{\text{safe}}) & \text{if } G_i = 0 \\ \text{Normal}(\mu_{\text{risky}}, \sigma_{\text{risky}}) & \text{if } G_i = 1 \end{cases} \\
\beta_j &\sim \text{Beta}(1, 1) \\
y_{ij} &\sim \text{Bernoulli}(p_{ij})
\end{aligned}
\tag{3}
$$

In Fig. 7, we find the posterior probability distributions of the hyperparameters of the Gaussian distributions that generate the agents' profiles. Fig. 8 shows the posteriors of the beta parameters of the questions. In both figures, the true value used in the generation process is shown in orange. Both figures also show the High Density Interval (HDI) of the posterior distributions [64]. The HDI is an interval within which the value of an unobserved parameter falls with a certain probability. In Bayesian inference, if the true parameter is within the 94% HDI of the posterior distribution, it is usually considered a "correct guess" [65].

As can be seen, the parameters were reconstructed quite accurately for the setting presented in this analysis. Therefore, in this sense, the synthetic data generation was successful, as the generated data encapsulate (probabilistic) information about the agent groups/clusters and the discriminative ability of the questions.

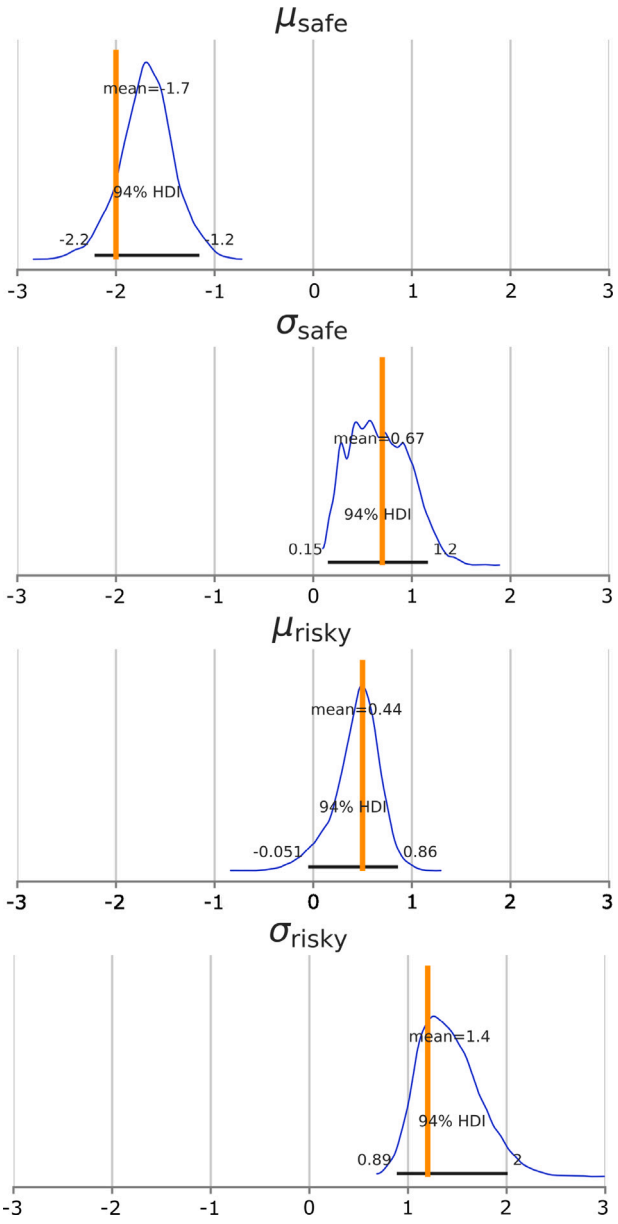


Fig. 7. Posterior probability distributions of the hyperparameters of the Gaussian distributions generating the agents' profiles in the case study. The true values used in the generation process are shown in orange. The black line at the bottom of each plot represents the HDI (94%).

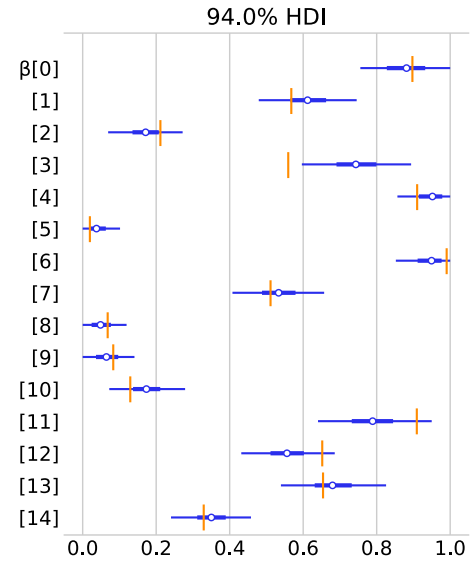


Fig. 8. HDIs (94%) of the posterior probability distributions of the questions' parameters (β_j). The true values used in the generation process are shown in orange. Note that all the HDIs of the inferred parameters, except β_3, include the true value.

4.2. Robustness analysis

Following the analysis of the synthetic data and the proposed model, in this subsection we will analyze the robustness in the reconstruction of the parameters using the synthetic data, as a function of the number of agents and questions. To do so, we have used the hierarchical Bayesian model shown in Fig. 6.

The motivation for this analysis is that, as we have seen in the previous subsection, the parameters used to generate the synthetic data are reconstructable (i.e. the true value is in the 94% HDI of the posteriors). However, if these posteriors are too wide (i.e., low confidence in the prediction), they will not be helpful, even if the true value is still within the HDI range. Therefore, in this analysis, we will systematically examine the "width" of the estimated posteriors while varying the number of agents and questions generated. In other words, we will analyze the confidence with which the Bayesian model has inferred the generation parameters.

To quantify the "width" of the posterior distributions with a single metric, we will use the entropy [66] of the distributions. This metric measures the average amount of *information* or *uncertainty* in a random variable. Given a discrete random variable X, which takes values in the range of $\mathcal{X}$ and is distributed according to the probabilities p, Eq. (4) defines its entropy.

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) = \mathbb{E}[-\log_2 p(X)] \tag{4}$$

In the experiments, we varied the number of agents from 5 to 1000 and the number of questions from 1 to 50. Then, we trained the hierarchical Bayesian model on each combination. Subsequently, we calculated the average entropy of $P(\alpha_i|\text{Data})$ and $P(\beta_j|\text{Data})$. Low entropy values represent that the model has high confidence in its prediction (i.e., the distribution is narrow) and vice versa. As we treat $p(x)$ as discrete (sample) probabilities, and to be able to compare among sets of parameters, we make a histogram of each distribution with the same number of bins in the same range of the parameter. To reduce sampling variability, we performed each experiment (with a fixed number of players and questions) 5 times, then normalized and averaged the obtained entropies. The final results are shown in Figs. 9 and 10.

The entropy value 1 represents complete uncertainty (i.e., the data do not contain any information about the parameter), and 0 represents perfect parameter information. To further facilitate the interpretation of the results obtained in the heatmaps, in Fig. 11 we show two examples of posterior distributions of the α parameters, also indicating the corresponding normalized entropy value.

The results of this robustness analysis give us an indication of how many players or questions we will need, depending on the precision with which we want to estimate the latent parameters. If we assume that the proposed model reflects the behavior of real players sufficiently well, this will help us to design the serious game of the RAYUELA project (our case study) and give us an idea of the precision we can expect depending on the number of participants, thus speeding up the development process.

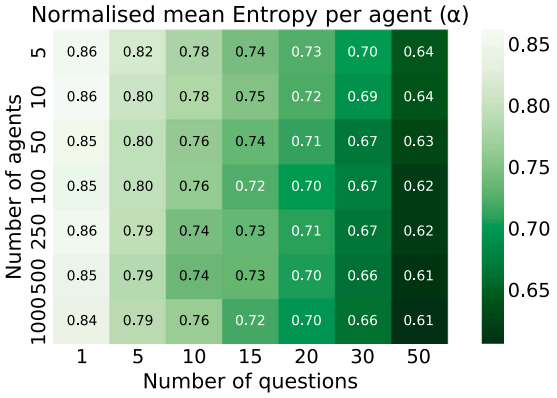


Fig. 9. Robustness experiments using the hierarchical Bayesian model on the agents' parameters (α_i), varying the number of agents and questions. The results show the normalized entropy of the posterior distribution of the inferred parameters, being 1 maximum entropy (i.e., no valuable information about the parameters) and 0 minimum entropy (i.e., complete information about the true value of the parameters).

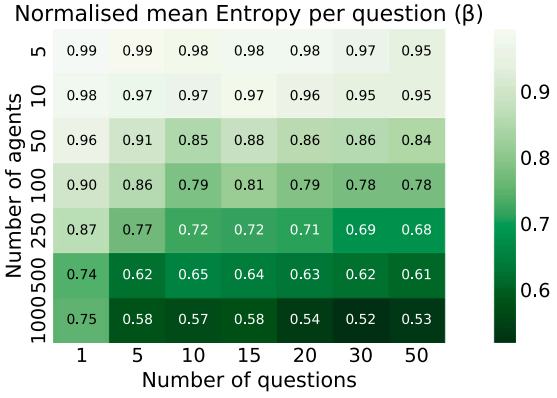


Fig. 10. Robustness experiments using the hierarchical Bayesian model on the questions' parameters (β_j), varying the number of agents and questions. The results show the normalized entropy of the posterior distribution of the inferred parameters, being 1 maximum entropy (i.e., no valuable information about the parameters) and 0 minimum entropy (i.e., complete information about the true value of the parameters).

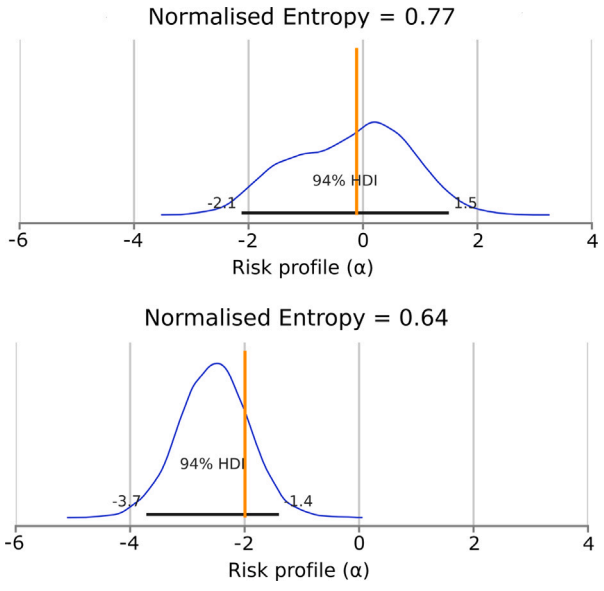


Fig. 11. Examples of posterior probability distributions of the α parameters. The corresponding normalized entropy value is also indicated to facilitate the interpretation of the results obtained in the robustness experiments.

5. Conclusions and limitations

This paper introduces a novel approach for generating probabilistic synthetic data explicitly tailored for decision-based SG. Although the methodology presented can be extended to create synthetic data for any decision-making scenario, such as multiple-choice exams, political polls, psychological questionnaires or other games/simulations. We demonstrated the validity of our proposal through a case study focused on cyberbullying.

The heart of our contribution lies in designing and implementing a simulator architecture explicitly conceived for generating synthetic probabilistic data (Fig. 1). This architecture provides flexibility for recreating decision-making scenarios within SG, capturing the complexity and uncertainty inherent in real-world situations. This work falls into the category of hybrid synthetic data since, by construction, causal BNs combine expert knowledge (through the design of the graph and the causal arrows between variables) and external data to inform the type and characteristics that feed the synthetic data simulator. Furthermore, the model mimicking player behavior is based on the IRT, a cognitive modeling framework for test scoring.

Our approach to synthetically generated data has proven a strategic advantage in SG development. In particular, in the RAYUELA project, this methodology has made significant contributions in two areas. Firstly, it has allowed us to refine the number of questions necessary for the serious game to achieve the desired results through robustness analysis (Figs. 9 and 10). Secondly, as in many large-scale social science projects, the actual data was not available until the last part of the project. The first half of the project was aimed primarily at the design, programming, and testing of the game itself. In this context, our generator enabled us to define the desired data structure in RAYUELA and prepare the pipeline software ahead of actual data collection, management, and analysis. In conclusion, this work has accelerated project development times and facilitated design, analysis, and data management. Including a synthetic data generation stage can become a customary methodological step to improve outcomes and reduce risks in large-scale projects.

Acknowledging the limitations of our methodology, we require expert knowledge to design the BN structure. Besides, our current focus is on decision-based SG, which narrows the scope of application as such games represent a fraction of the broader gaming landscape.

There are additional limitations due to the use of BN. Firstly, the results obtained rely on the truthfulness of the structure designed by the experts (Fig. 4) and the data quality. However, these assumptions are explicit, improving transparency and promoting discussion, which mitigate the impact of this issue in the research process and results. Secondly, as the number of variables increases, the complexity of BN grows exponentially. Learning the structure and parameters of large BN becomes computationally intensive, and the resulting models may become difficult to interpret.

As future lines of research, reducing expert bias (and producing more robust DAGs) requires the definition of novel methodologies to integrate expert knowledge with data. Also, explore the potential impact of introducing memory on the agents, which could lead to a more realistic model, allowing past responses to influence future responses. It would also be interesting to perform tests on discriminative tasks to verify the validity of the results. Finally, we encourage researchers to employ our methodology to generate synthetic data for other decision-making problems.

Our work contributes to the growing body of research at the intersection of SG and synthetic data generation, offering a valuable tool for decision-based game developers and researchers. This emerging field holds immense potential to advance the development of SG and decision-based simulations and analysis.

CRediT authorship contribution statement

Jaime Pérez: Writing – review & editing, Writing – original draft, Software, Investigation, Data curation. **Mario Castro:** Writing – review & editing, Validation, Software, Investigation, Formal analysis. **Edmond Awad:** Writing – review & editing, Validation, Supervision, Investigation. **Gregorio López:** Writing – review & editing, Validation, Supervision, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 882828. The authors would like to thank all the partners within the consortium for the fruitful collaboration and discussion. The sole responsibility for the content of this document lies with the authors and in no way reflects the views of the European Union. This work has been partially supported by Grant PID2022-140217NB-I00 funded by MCIN/AEI/ 10.13039/501100011033 and, by "ERDF A way of making Europe".

References

[1] UNICEF, The State of the World's Children 2017: Children in a Digital World, tech. rep., UNICEF Division of Communication, 2017, ISBN: 978-92-806-4930-7.
[2] D. Smahel, H. Machackova, G. Mascheroni, L. Dedkova, E. Staksrud, K. Ólafsson, S. Livingstone, U. Hasebrink, EU Kids Online 2020: Survey Results from 19 Countries, tech. rep., EU Kids Online, 2020, ISSN: 2045-256X.
[3] European Commission, J.R. Centre, B. Lobe, A. Velicu, E. Staksrud, S. Chaudron, R. Di Gioia, How Children (10-18) Experienced Online Risks During the COVID-19 Lockdown : Spring 2020 : Key Findings from Surveying Families in 11 European Countries, Publications Office of the European Union, 2021.
[4] G. López, N. Bueno, M. Castro, M. Reneses, J. Pérez, M. Riberas, M. Álvarez-Campana, M. Vega-Barbas, S. Solera-Cotanilla, L. Bastida, et al., The H2020 project RAYUELA: A fun way to fight cybercrime, Jornadas Nac. Investig. Ciberseguridad - JNIC (2021).
[5] C.C. Abt, Serious Games, University press of America, 1987.
[6] S. Çiftci, Trends of serious games research from 2007 to 2017: A bibliometric analysis, J. Educ. Train. Stud. 6 (2) (2018) 18–27, ISSN-2324-805X.
[7] Y. Zhonggen, A meta-analysis of use of serious games in education over a decade, Int. J. Comput. Games Technol. 2019 (2019) 4797032.
[8] K. Larson, Serious games and gamification in the corporate training environment: a literature review, TechTrends 64 (2) (2020) 319–328.
[9] A. Coutrot, E. Manley, S. Goodroe, C. Gahnstrom, G. Filomena, D. Yesiltepe, R.C. Dalton, J.M. Wiener, C. Hölscher, M. Hornberger, H.J. Spiers, Entropy of city street networks linked to future spatial navigation ability, Nature 604 (7904) (2022) 104–110.
[10] J.K. Hartshorne, J.B. Tenenbaum, S. Pinker, A critical period for second language acquisition: Evidence from 2/3 million English speakers, Cognition 177 (2018) 263–277.
[11] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan, The Moral Machine experiment, Nature 563 (7729) (2018) 59–64.
[12] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S.N. Cohen, A. Weller, Synthetic Data – what, why and how? 2022.
[13] J. Pérez, M. Castro, G. López, Serious games and AI: Challenges and opportunities for computational social science, IEEE Access 11 (2023) 62051–62061.
[14] S. Grund, O. Lüdtke, A. Robitzsch, Using synthetic data to improve the reproducibility of statistical results in psychological research, Psychol. Methods (2022).
[15] D.S. Quintana, A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation, eLife 9 (2020).
[16] B. Howe, J. Stoyanovich, H. Ping, B. Herman, M. Gee, Synthetic data for social good, 2017.
[17] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan kaufmann, 1988.
[18] B.G. Marcot, T.D. Penman, Advances in Bayesian network modelling: Integration of modelling technologies, Environ. Model. Softw. 111 (2019) 386–393.
[19] K. Sitnik-Warchulska, Z. Wajda, B. Wojciechowski, B. Izydorczyk, The risk of bullying and probability of help-seeking behaviors in school children: A Bayesian network analysis, Front. Psychiatry 12 (2021).
[20] J. Li, Y. Jin, S. Xu, A. Wilson, C. Chen, X. Luo, Y. Liu, X. Ling, X. Sun, Y. Wang, Effects of bullying on anxiety, depression, and posttraumatic stress disorder among sexual minority youths: Network analysis, JMIR Public Health Surveill. 9 (2023) e47233.
[21] J. Petrillo, S.J. Cano, L.D. McLeod, C.D. Coon, Using classical test theory, item response theory, and rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples, Value Health 18 (1) (2015) 25–34.
[22] J.C. Cappelleri, J.J. Lundy, R.D. Hays, Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures, Clin. Ther. 36 (5) (2014) 648–662.
[23] S.L. Vincenzi, E. Possan, D.F.d. Andrade, M.M. Pituco, T.d.O. Santos, E.P. Jasse, Assessment of environmental sustainability perception through item response theory: A case study in Brazil, J. Clean. Prod. 170 (2018) 1369–1386.
[24] K. El Emam, L. Mosquera, R. Hoptroff, Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data, O'Reilly Media, 2020.
[25] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S.N. Cohen, A. Weller, Synthetic Data – what, why and how? 2022.
[26] T.T. Um, F.M.J. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, D. Kulić, Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 216–220.
[27] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019).
[28] Z. Wan, Y. Zhang, H. He, Variational autoencoder based synthetic data generation for imbalanced learning, in: 2017 IEEE Symposium Series on Computational Intelligence, SSCI, 2017, pp. 1–7.
[29] Z. Islam, M. Abdel-Aty, Q. Cai, J. Yuan, Crash data augmentation using variational autoencoder, Accid. Anal. Prev. 151 (2021) 105950.
[30] A. Antoniou, A. Storkey, H. Edwards, Augmenting image classifiers using data augmentation generative adversarial networks, in: Lecture Notes in Computer Science, Springer International Publishing, 2018, pp. 594–603.
[31] J. Pérez, P. Arroba, J.M. Moya, Data augmentation through multivariate scenario forecasting in Data Centers using Generative Adversarial Networks, Appl. Intell. 53 (2) (2023) 1469–1486.
[32] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, M.-H. Yang, Diffusion models: A comprehensive survey of methods and applications, ACM Comput. Surv. 56 (4) (2023).
[33] P. Dhariwal, A. Nichol, Diffusion models beat GANs on image synthesis, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J.W. Vaughan (Eds.), in: Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 8780–8794.
[34] S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A.I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, R.G. Baraniuk, Self-consuming generative models go MAD, 2023, arXiv preprint arXiv:2307.01850.
[35] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, R. Anderson, The curse of recursion: Training on generated data makes models forget, 2023.
[36] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, H. Xu, Time series data augmentation for deep learning: A survey, 2020, arXiv preprint arXiv:2002.12478.
[37] B.K. Iwana, S. Uchida, An empirical survey of data augmentation for time series classification with neural networks, Plos One 16 (7) (2021) e0254841.
[38] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 60.
[39] C. Shorten, T.M. Khoshgoftaar, B. Furht, Text data augmentation for deep learning, J. Big Data 8 (1) (2021) 101.
[40] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, D. Hassabis, A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play, Science 362 (6419) (2018) 1140–1144.
[41] W. Zhao, J.P. Queralta, T. Westerlund, Sim-to-real transfer in deep reinforcement learning for robotics: a survey, in: 2020 IEEE Symposium Series on Computational Intelligence, SSCI, 2020, pp. 737–744.
[42] F. Tao, B. Xiao, Q. Qi, J. Cheng, P. Ji, Digital twin modeling, J. Manuf. Syst. 64 (2022) 372–389.
[43] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, S. Birchfield, Training deep networks with synthetic data: Bridging the reality gap by domain randomization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018.

[44] B. Osiński, A. Jakubowski, P. Zięcina, P. Miłoś, C. Galias, S. Homoceanu, H. Michalewski, Simulation-based reinforcement learning for real-world autonomous driving, in: 2020 IEEE International Conference on Robotics and Automation, ICRA, 2020, pp. 6411–6418.

[45] C.N. Vasconcelos, B.N. Vasconcelos, Increasing deep learning melanoma classification by classical and expert knowledge based image transforms, 2017, ArXiv, abs/1702.07025.

[46] O. Petrovic, D.L.D. Duarte, S. Storms, W. Herfs, Towards knowledge-based generation of synthetic data by taxonomizing expert knowledge in production, in: Intelligent Human Systems Integration (IHSI 2023): Integrating People and Intelligent Systems, Vol. 69, AHFE Open Acces, 2023, no. 69.

[47] B. Yang, Z. Liu, G. Duan, J. Tan, Mask2Defect: A prior knowledge-based data augmentation method for metal surface defect inspection, IEEE Trans. Ind. Inform. 18 (10) (2022) 6743–6755.

[48] G. Shi, J. Wang, Y. Qiang, X. Yang, J. Zhao, R. Hao, W. Yang, Q. Du, N.G.-F. Kazihise, Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification, Comput. Methods Programs Biomed. 196 (2020) 105611.

[49] C.-E. Kuo, G.-T. Chen, P.-Y. Liao, An EEG spectrogram-based automatic sleep stage scoring method via data augmentation, ensemble convolution neural network, and expert knowledge, Biomed. Signal Process. Control 70 (2021) 102981.

[50] G. Lederrey, T. Hillel, M. Bierlaire, DATGAN: Integrating expert knowledge into deep learning for synthetic tabular data, 2022.

[51] D. Kaur, M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta, N. Markuzon, Application of Bayesian networks to generate synthetic health data, J. Am. Med. Inform. Assoc. 28 (4) (2020) 801–811.

[52] G. Gogoshin, S. Branciamore, A.S. Rodin, Synthetic data generation with probabilistic Bayesian Networks, Math. Biosci. Eng. 18 (6) (2021) 8603–8621.

[53] B. Wang, T. Sun, X.S. Zheng, Beyond winning and losing: Modeling human motivations and behaviors with vector-valued inverse reinforcement learning, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, Vol. 15, 2019, pp. 195–201.

[54] B. Lin, G. Cecchi, D. Bouneffouf, J. Reinen, I. Rish, A story of two streams: Reinforcement learning models from human behavior and neuropsychiatry, 2019, arXiv preprint arXiv:1906.11286.

[55] S.E. Embretson, S.P. Reise, Item Response Theory, Psychology Press, 2013.

[56] S.E. Embretson, S.P. Reise, Item Response Theory for Psychologists, Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 2000, p. xi, 371.

[57] R. Ostini, M.L. Nering, Polytomous Item Response Theory Models, in: Quantitative Applications in the Social Sciences, No. 144, SAGE Publications, Inc., 2006.

[58] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1) (1977) 1–22.

[59] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6 (6) (1984) 721–741.

[60] S.L. Lauritzen, The EM algorithm for graphical association models with missing data, Comput. Statist. Data Anal. 19 (2) (1995) 191–201.

[61] A.G.C. Wright, M.N. Hallquist, J.Q. Morse, L.N. Scott, S.D. Stepp, K.A. Nolf, P.A. Pilkonis, Clarifying interpersonal heterogeneity in borderline personality disorder using latent mixture modeling, J. Pers. Disord. 27 (2) (2013) 125–143.

[62] S. Sugasawa, Grouped heterogeneous mixture modeling for clustered data, J. Amer. Statist. Assoc. 116 (534) (2021) 999–1010.

[63] J. Salvatier, T.V. Wiecki, C. Fonnesbeck, Probabilistic programming in Python using PyMC3, PeerJ Comput. Sci. 2 (2016) e55.

[64] R. McElreath, Statistical Rethinking: A Bayesian Course with Examples in R and Stan, Chapman and Hall/CRC, 2020.

[65] D. Makowski, M. Ben-Shachar, D. Lüdecke, bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework, J. Open Source Softw. 4 (40) (2019) 1541.

[66] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.