Impact of Serious Games and Causal Artificial Intelligence on Social Science Research:

A Case Study on Cyberbullying

by
Jaime Pérez Sánchez

supervised by
Dr. Gregorio López López
Dr. Mario Castro Ponce

A document submitted for the degree of
Doctor of Philosophy
at
ICAI School of Engineering
Universidad Pontificia Comillas

Madrid, 2024

«La aspiración a la verdad es más valiosa que su posesión segura.»
- Gotthold Ephraim Lessing

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Jaime Pérez Sánchez
Madrid, 2024

Acknowledgement

I am deeply grateful to my supervisors, Dr. Mario Castro and Dr. Gregorio López, for their mentorship and kindness throughout these years. I feel very fortunate to have had the opportunity to grow up under your tutelage. Thank you for the trust you have placed in me.

Thank you to my colleagues at IIT, especially Juan Luis Gómez, for their support and the beautiful friendship we have built.

Finally, I would like to express my deepest gratitude to my friends, family, and beloved partner, Celia, for their unwavering support, encouragement, and love. This work is dedicated to you.

Abstract

This thesis explores the transformative potential of integrating Serious Games (SG) and causal Artificial Intelligence (AI) into social science research. This integration holds the key to addressing persistent challenges in social science research, such as limited sample sizes or the difficulty of reaching certain populations, like minors. Designed with a purpose beyond entertainment, SGs are emerging as a promising tool to engage diverse audiences and challenge traditional social research paradigms. The interactive and immersive nature of video games opens up opportunities to increase participant involvement, simulate real-life scenarios with fewer ethical constraints, and reach populations with greater demographic variability.

The main contribution of this thesis is to advocate the use of Probabilistic Graphical Causal Models (PGCM) to analyze data derived from the use of SGs in research contexts. These models facilitate the understanding of complex causal relationships and provide an analytical framework that overcomes some of the limitations of traditional statistical and machine learning approaches. In particular, we propose using these models to analyze sensitive social issues, such as cyberbullying (CB).

Firstly, we present a methodology for constructing Directed Acyclic Graphs (DAG) that harmoniously combine expert knowledge with data-driven algorithms. This hybrid approach aims to produce robust and realistic causal models that accurately reflect the complexity of the phenomena under study. Secondly, we introduce a novel synthetic data generation technique for SGs, based on a psychometric theoretical framework. This technique allows us to simulate player behavior in a parameterized and realistic manner. The incorporation of synthetic data into this area of research enhances the efficiency of development processes by overcoming obstacles such as data scarcity, and facilitates the design, analysis and management of data.

The empirical validation of these methods was carried out within the framework of the European H2020 project RAYUELA. This project is a pioneering multidisciplinary initiative that studies cybercrime involving minors and seeks to propose methodologies to better understand the factors that influence it. The results demonstrate the potential of SGs to observe and understand the complex dynamics of CB. This thesis posits that the integration of SGs and causal AI in social science research can significantly advance our capacity to address complex social issues by enabling large-scale ethical experiments to explore into the causal relationships that define such issues.

In conclusion, this thesis makes a significant contribution to demonstrating the potential of SGs as a valuable research tool in the social sciences and deepens existing methodologies in this field by integrating causal AI techniques. In doing so, it contributes to the multidisciplinary scientific community by offering new ways to explore and address relevant social challenges. The implications of this research extend beyond academia, providing practical tools and knowledge to develop effective prevention and intervention strategies against CB, thus improving the online well-being of minors.

Resumen

Esta tesis explora la integración de los juegos serios y la Inteligencia Artificial causal en la investigación en ciencias sociales. Esta integración tiene el potencial de hacer frente a retos que han estado presentes asiduamente en el estudio de las ciencias sociales, tales como tamaños de las muestras insuficientes o la dificultad para llegar a algunos sectores de la población, como los menores de edad. Los juegos serios, definidos como juegos diseñados con fines que van más allá del entretenimiento, se postulan como una prometedora herramienta para atraer a públicos diversos y superar algunos paradigmas tradicionales de la investigación social. La naturaleza interactiva e inmersiva de los videojuegos permite aumentar la implicación de los participantes, simular escenarios de la vida real con menos limitaciones éticas y alcanzar poblaciones con una mayor variabilidad demográfica.

La principal contribución de esta tesis es abogar por el uso de modelos gráficos causales probabilísticos para analizar los datos derivados del uso de juegos serios en contextos de investigación. Estos modelos facilitan la comprensión de relaciones causales complejas y proporcionan un marco analítico que permite superar algunas de las limitaciones de los enfoques estadísticos y de aprendizaje automático tradicionales. En particular, se propone aquí utilizar estos modelos para el análisis de cuestiones sociales delicadas, como el ciberacoso.

En primer lugar, proponemos una metodología para construir grafos acíclicos dirigidos que combinan el conocimiento experto con algoritmos basados en datos. Este enfoque híbrido pretende producir modelos causales robustos y realistas que reflejan con precisión la complejidad de los fenómenos estudiados. En segundo lugar, presentamos una técnica de generación de datos sintéticos para juegos serios basada en un marco teórico psicométrico que permite simular el comportamiento de los jugadores de forma parametrizada y realista. La introducción de datos sintéticos en esta área de investigación permite mejorar la eficacia de procesos de desarrollo al permitir superar obstáculos como la escasez de datos, así como facilitar el diseño, análisis y gestión de los datos.

La validación empírica de estos métodos se ha llevado a cabo en el marco del proyecto europeo H2020 RAYUELA. Este proyecto es una iniciativa multidisciplinar pionera que estudia la ciberdelincuencia en la que se ven involucrados menores de edad y busca proponer metodologías para comprender mejor los factores que influyen en esta. Los resultados ponen de manifiesto el potencial de los juegos serios para observar y comprender las complejas dinámicas vigentes en el ciberbullying. Esta tesis defiende que la integración de los juegos serios y la Inteligencia Artificial causal en la investigación en ciencias sociales puede constituir un avance significativo en nuestra capacidad para abordar cuestiones sociales complejas, al permitir la realización de experimentos éticos a gran escala, que permitan ahondar en las relaciones causales que definen dichas cuestiones.

En definitiva, esta tesis contribuye significativamente a demostrar el potencial de los juegos serios como una herramienta valiosa de investigación en ciencias sociales y profundiza en las metodologías existentes en este campo, integrando técnicas de Inteligencia Artificial causal. De este modo, contribuye a la comunidad científica multidisciplinar ofreciendo nuevas formas de explorar y abordar retos sociales relevantes. Las implicaciones de esta investigación se extienden más allá del ámbito académico, proporcionando herramientas y conocimientos prácticos para desarrollar estrategias eficaces de prevención e intervención contra el ciberbullying, mejorando así el bienestar online de los menores de edad.

Acronyms

ACE	Average Causal Effect
ATE	Average Treatment Effect
BF	Bayes Factor
BIC	Bayesian Information Criterion
CB	Cyberbullying
CPD	Conditional Probability Distribution
CPT	Conditional Probability Table
CSS	Computational Social Science
DAG	Directed Acyclic Graph
LL	Log-Likelihood
OR	Odds Ratio
PGCM	Probabilistic Graphical Causal Model
SG	Serious Game

Contents

1 Introduction

This chapter discusses the motivation, scope and objectives of this thesis. The research framework project is also described, which helps to understand the justification for the work accomplished. Finally, the overall structure of the thesis is presented.

1.1 Motivation

Social Science research has encountered, traditionally, numerous challenges, such as the constraints posed by small sample sizes, ethical dilemmas inherent when conducting experiments, and the difficulties associated with cultural biases [41, 71, 206]. Consequently, these issues have significantly shaped the scope, research methodologies, and outcomes within the field, making pursuing generalizable insights exceptionally challenging. As the discipline seeks innovative solutions to these persistent issues, exploring serious games as a research tool emerges as a promising avenue [189].

Serious Games (SG) —games designed with purposes beyond mere entertainment— have shown significant potential in various domains, including education, training, awareness initiatives, and healthcare [189]. Their purpose-driven design aims to achieve specific outcomes such as skill development, knowledge enhancement, or behavioral change [3]. The versatility and broad appeal of SGs have established them as a valuable tool in engaging diverse audiences, effectively transcending traditional educational and training paradigms [271, 40]. However, their application as a research tool within the social sciences has been relatively underexplored, an oversight that neglects the substantial benefits they offer, including enhanced participant engagement, a reduction in inhibitions compared to traditional research settings, and the potential to scale experiments to larger and more culturally diverse sample sizes while maintaining manageable costs.

There are several advantages of using SGs in research. First, their interactive and engaging nature can increase participation and motivation among subjects, potentially yielding richer and more authentic data. Also, the immersive environments of SGs can facilitate observing behaviors and decisions in contexts that closely simulate real-life situations, without the ethical and practical constraints associated with real-world experimentation. The scalability of SGs also permits data collection from a broad range of demographic and cultural backgrounds, thus addressing the challenge of small sample sizes and enhancing the generalizability of research findings. Such benefits position SGs as an attractive solution for overcoming traditional research obstacles.

However, leveraging SGs as a methodological tool in scientific inquiry demands the application of rigorous analytical frameworks capable of extracting reliable data from game-based interactions. Traditional statistical models and machine learning techniques, while useful, often suffer from limitations due to their susceptibility to spurious correlations and biases inherent in the data [159, 269]. This issue becomes especially significant when addressing sensitive topics, such as cyberbullying (CB), where the accuracy and reliability of research findings can have profound real-world consequences.

To overcome these challenges, this thesis proposes the adoption of Probabilistic Graphical Causal Models (PGCMs) as a viable analytical solution for SG-derived data. PGCMs offer an intuitive and robust framework for elucidating complex causal relationships, enabling researchers to not only identify but also understand the underlying dynamics of the phenomena under investigation [238]. Additionally, PGCMs enable a straightforward inclusion of expert knowledge, intrinsically quantify the uncertainty associated with estimations, and can simulate interventions. Analyzing data derived from SGs with PGCMs could advance our approach to social science research, enabling us to overcome longstanding obstacles, and offering a nuanced and reliable understanding of causal mechanisms. This integration is not only intended to improve the validity and comprehensiveness of research results but also to foster a more ethically and culturally sound research practice.

PGCMs have two main ingredients: a network of interaction between variables (causes and effects) and a mathematical description of those interactions, typically in the form of conditional probabilities of the effects given the causes. A pivotal challenge in employing PGCMs, especially within social science research, is constructing the network structure. This network, known as a causal Directed Acyclic Graph (DAG), defines and models the causal relationships among variables. The efficacy and accuracy of PGCMs hinge on the robustness of the DAGs, which should accurately reflect the complexities of the phenomena under study. The construction of these DAGs can be approached through expert knowledge, data-driven automatic structure learning algorithms, or a mixture of both approaches. The reliability of the data-driven structure learning algorithms remains a concern, particularly in situations with limited available data and under causal assumptions [128, 218].

To address the intricacies of DAG construction, this thesis introduces a methodology aimed at aligning DAGs proposed by subject matter experts with those generated from data using algorithms. This method attempts to leverage the detailed knowledge of domain experts and empirical patterns discovered through data analysis, producing DAGs that are not only robust but also reflect the real-world complexity of the problems being addressed. The proposed consensus-based approach combines theoretical knowledge with empirical evidence, resulting in a more comprehensive and reliable causal model.

Another significant challenge in social science research, particularly in the context of behavioral science, is the scarcity of available data. Artificial Intelligence (AI) offers a promising solution to this issue by generating synthetic data, a technique that has witnessed increasing interest in recent years. Synthetic data, though artificially generated, share statistical properties with real-world data, rendering them invaluable in situations where acquiring actual data is impractical, costly, or ethically questionable [114]. This thesis proposes a specific methodology for generating synthetic data within the domain of SGs. By building a parameterized model of player behavior informed by cognitive testing frameworks, this methodology aims to simulate player responses to various in-game scenarios accurately. Including probabilistic elements in the data generation process further aims to mimic the variability observed in real-life closely.

The initiatives detailed in this thesis are part of a larger effort framed by the European research project H2020 RAYUELA[a]. The proposed methodologies for constructing consensus-based casual DAGs, generating synthetic data, and analyzing SG data from a causal standpoint were developed within the context of this project. However, these techniques were also designed with a broad perspective so they can be applied to other domains and scenarios. This project provides a practical framework for applying and validating the novel procedures through case studies leveraging real experimental data derived from the project's endeavors. The next section will explore the H2020 RAYUELA project in greater depth, outlining its goals and the specific challenges it seeks to overcome. This discussion aims to position the thesis within a broader research context, underscoring its contributions to the field and its potential to enhance our understanding and exploration of SGs as a research tool.

1.2 Framework Project: H2020 RAYUELA

The advent of the digital era has precipitated a significant escalation in Internet usage among children. According to the UNICEF "Children in a Digital World" report from 2017, minors under the age of 18 constituted nearly one-third of the global Internet user base [243]. The onset of the COVID-19 pandemic has fur-

[a] https://www.rayuela-h2020.eu/

ther amplified this trend, resulting in increased digital engagement among children at progressively younger ages. Young people exhibit a profound capacity to leverage the multifaceted benefits and opportunities of the Internet and related technologies. Nonetheless, uncontrolled access to the Internet exposes children to additional risks, exacerbating the vulnerabilities of those already at risk. Despite their familiarity with digital environments —often referred to as digital natives— many young users remain inadequately informed about the potential perils and prospects that accompany digital technology and Internet utilization.

Law enforcement and security agencies and researchers devote considerable efforts to examining the technical and mechanical dimensions of cybercrime, including analysis of malware, exploit tools, and forensic examination of coding and techniques. However, relatively little attention is given to exploring the social and psychological factors of cybercriminal behavior, including identifying perpetrators, their motivations, and the genesis of their criminal activities. Recognizing the importance of these elements is crucial to moving from a punitive approach to one focused on deterrence and prevention. It is imperative to understand the trajectories that lead certain adolescents into cybercriminal activities to formulate effective preventive and intervention strategies. Investigating how young people are initiated into cybercrime will enhance educational initiatives in family and academic settings, especially for the most vulnerable, and contribute to developing comprehensive policing strategies and best practices in various cybercrime domains.

The European Commission, recognizing the profound implications of these issues, launched in 2012 the European Strategy for a Better Internet for Children[b] (BIK strategy). This initiative, designed as an extension of the previous Safer Internet Program, is based on four key principles: 1) promoting high-quality online content for young users, 2) improving awareness and empowerment, 3) establishing a safe online environment for children, and 4) vigorously combating child sexual abuse and exploitation. This strategic envisioning underscores the critical importance of addressing both the opportunities and challenges of children's engagement with the digital world. It is a significant step towards creating a safer and more beneficial Internet for the younger generation.

RAYUELA (empowe**R**ing and educ**A**ting **Y**o**U**ng p**E**ople for the internet by p**LA**ying) [155] represents a pioneering multidisciplinary research initiative that brings together law enforcement agents, sociologists, anthropologists, psychologists, lawyers, philosophers, computer scientists, and engineers. This collaboration is underpinned by a central objective: to develop innovative methodologies to understand better the multiple factors that influence certain forms of cybercrime among minors (namely, CB, online grooming, human trafficking for sexual exploitation, misinformation and deception, and technological threats related to IoT devices). The project proposes to achieve these goals by playing, through a SG, representing a novel approach to investigate and mitigate cybercriminal behavior in a friendly and non-invasive way.

In addition, RAYUELA aims to harness the potential of young people in the field of cybersecurity, as well as encourage adherence to exemplary online practices. This endeavor directly aligns with the foundational pillars of the European Strategy for a Better Internet for Kids (BIK strategy), as outlined above. RAYUELA's focus on educating and training young talent in cybersecurity and promoting sound online practices embodies a proactive approach. Through educational initiatives and the development of best practices, the project aims to equip young people with the knowledge and skills necessary to navigate the digital world safely and responsibly.

M. J. Salganik [206] defines two distinct approaches to computational social science research: *readymade* and *custommade*. The *readymade* methodology uses existing big data sources and repurposes them ingeniously. In the *custommade* methodology, on the other hand, researchers start with a concrete question and create an ad-hoc digital tool to study that question. The RAYUELA project would fit into the *custommade* category, as it designs and creates a specific SG to answer the research questions posed.

The methodology and lines of action defined in the RAYUELA project are shown in Figure 1.1:

[b] https://www.betterinternetforkids.eu/policy/newbikstrategy

- *Breaking the ice*: First, it aims to create a knowledge base on the drivers of cybercrime and to assess the technological landscape and other interdisciplinary factors from psychology, criminology, and anthropology.
- *Weighting in*: Then, create preliminary profiles and unify knowledge about cybercrime in minors, from the information coming from social and technological research.
- *It's child's play!*: Based on the knowledge acquired, design and development of the SG that will be used in the following phases.
- *Getting the full picture*: Experimental phase of pilots across the EU using the developed SG.
- *The ins and outs*: Analysis and interpretation of data for prevention and awareness of cybercrime among minors. This phase is expected to provide refined insights to inform future evidence-based policies and strategies.

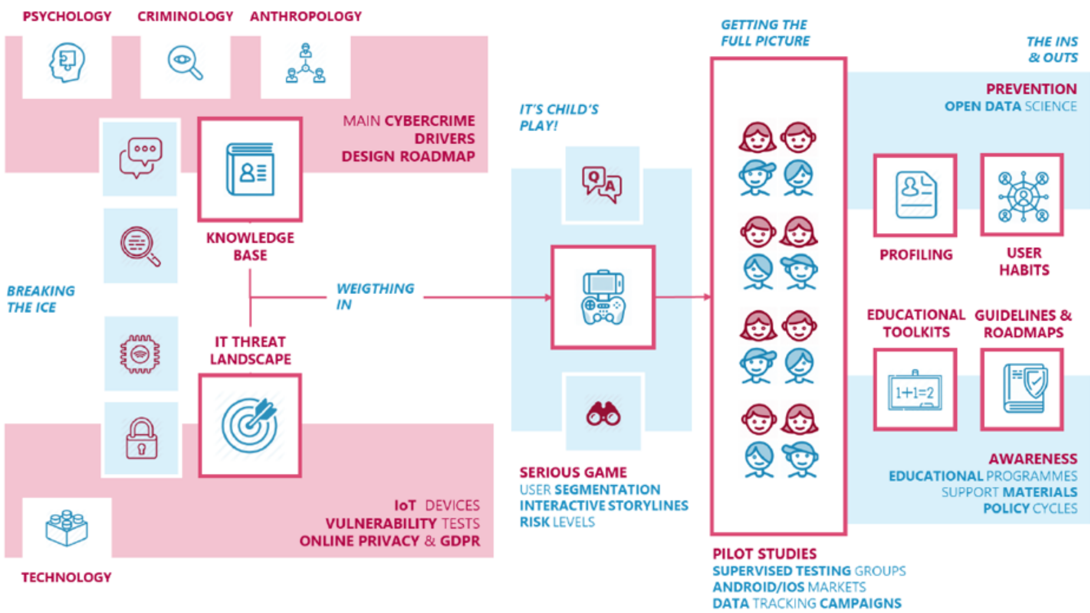


Figure 1.1: Diagram describing the methodology of the RAYUELA project. The strength of the approach relies upon the multidisciplinary of the participants and the combination of technology and state-of-the-art data analysis methods. Source: RAYUELA Grant Agreement No 882828.

This doctoral thesis and the methodologies presented have been developed in the context of the RAYUELA project, mainly within the line of action *The ins and outs*. Besides, the empirical basis of this thesis is grounded in data gathered through RAYUELA, marking a significant contribution given the inherent challenges associated with collecting data on such a delicate subject involving minors. Several teams within the RAYUELA consortium have dealt with the ethical, legal, and technological aspects regarding the creation of the game and the creation of databases, as well as the collection of data in the pilots. Therefore, these aspects are not part of the development of this doctoral thesis.

Our analysis is built on the use cases derived from this data collection effort to validate the theoretical propositions, underscoring the practical implications of our research findings. Specifically, this thesis' case studies focus on CB, as it represents one of the most widespread cybercrimes. Approximately 10% of European children are cyberbullied every month [229], and nearly 50% have experienced a CB-related incident at least once [65]. Moreover, CB is the only cybercrime among those considered for which we have obtained a validated questionnaire to be filled in by the participants, serving as the "ground truth" for the analyses. This way, we can test whether the developed SG is a rigorous tool and whether the results are valid.

In essence, this thesis represents an academic effort to advance the understanding of using SG as a novel tool in social science research, validating the propositions practically through CB-related use cases. Additionally, aiding in a better understanding of the complex dynamics of CB, and contributing to the development of informed strategies to safeguard digital environments inhabited by young individuals.

1.2.1 Data Collection

In this section, we describe the two datasets collected through the RAYUELA project and used in this thesis's case studies to demonstrate the validity of the proposed methodologies.

All data collection and experimental procedures were approved by the ethics committees of the RAYUELA consortium in each of the countries involved in the research. In addition, RAYUELA's legal experts also took the necessary measures to ensure that the collection, storage, and disclosure of data is in compliance with the European GDPR. In each session, the researchers and teachers explained the project and its main objective, as well as the data to be collected. Depending on the age, the participants or their parents signed an informed consent to participate.

Survey of Spanish minors

This dataset was collected through a representative survey of children in schools in Madrid (Spain). The survey collected responses from 665 students aged 13 to 17, where 50.6% identified themselves as males, 48% as females, and 1.4% as non-binary. We gather demographic and environmental information (*age*, *gender*, *sexual orientation*, *migratory background*, and *family communication*), participants' relationship with technology and the Internet, and inquiries about CB-related situations. Table 1.1 provides the variable values and their percentage of occurrences (i.e., marginal probability). Appendix B provides a more exhaustive exploratory data analysis.

Experimental pilots using the serious game

This dataset was created with input from the pilots using the developed SG in European schools and institutes. Specifically, data were collected mainly in Spain, Greece, Belgium, Estonia, Portugal, the United Kingdom, and the Netherlands. We gathered 1055 responses from students between 12 and 16 years old (Mean=13.9, SD=1.34), with 57.7% identifying as males, 38.1% as females, and 1.3% as non-binary. Table 1.2 shows summarized statistics of the data collected, the possible values of each variable, and their percentage of occurrence (i.e., the marginal probability). Appendix C provides a more exhaustive exploratory data analysis. These data have been published openly [187], addressing the scarcity of shared data in this research domain.

Pre-gameplay data

Before starting the game session, all participants registered and filled out demographic and psychological questionnaires. RAYUELA's CB experts meticulously chose each demographic and psychological variable to collect, aligning with prior and conducted research suggesting these factors may substantially impact the susceptibility and response to CB [197, 198, 199]. The following variables were obtained during the first phase of each pilot, before playing the SG:

- Demographic variables: *Age*, *gender*, *sexual orientation*, *migratory background*, and *daily hours spent on the Internet*. These variables were considered to understand the diverse demographic background of the participating minors, along with a measure of their relationship with technology.

- Psychological and environmental variables: *Social support* (isolation), *family support*, and *self-esteem*. These variables were chosen to provide a baseline for understanding the minors' emotional, social, and psychological state before playing the game. They were extracted from the following validated questionnaires: the *Rosenberg Self-Esteem Scale* [204] and *The Multidimensional Scale of Perceived Social Support* [272].

Gameplay data

The data collected exclusively through the SG encompasses two essential elements: (i) the player's decisions at each game dilemma and (ii) a post-game question ("*Did you play as you would behave in the real world?*"). Although minors were asked to play as they would act in real life at the beginning of the sessions, this post-game question acts as a calibration and control on the "honesty" while playing, since the game format may encourage some players to make more adventurous choices to explore narrative options.

Post-gameplay data

After each game session, minors completed a questionnaire about their past CB experiences, serving as the "ground truth" for the data analysis. The validated questionnaire was the *European Cyberbullying Intervention Project Questionnaire* [28]. The questionnaire quantifies whether the person has suffered or committed (or both) CB-related acts.

Table 1.1: Survey of Spanish minors dataset ($N = 665$): The table shows the possible values of each variable and its marginal probability (i.e., the percentage of observation).

Variable	**Response Values**	**Marginal Probability**
Gender	Male	50.6%
	Female	48%
	Non Binary	1.4%
Age	13	8.2%
	14	46.5%
	15	31.4%
	16	11.7%
	17	2.2%
Sexual Orientation	Heterosexual	84.5%
	Non Heterosexual	15.5%
Migratory Background	No	82.9%
	Yes	17.1%
Family Communication	1 (Never)	5.3%
	2	20%
	3	42.3%
	4 (Very frequently)	32.4%
Daily Hours of Internet	Less than 1 h	5.1%
	1-2 h	35.9%
	2-4 h	36.5%
	More than 4 h	22.5%
CB Awareness	1 (Not worried)	7.2%
	2	5.4%
	3	11%
	4	23.1%
	5 (Very worried)	53.3%
Suffered CB-related Situations	No	86.5%
	Yes	13.5%

Table 1.2: Dataset of experimental pilots using the serious game ($N = 1055$): The table shows the possible values of each variable and its marginal probability (i.e., the percentage of observation). The «missing» label aggregates incorrect values and items that respondents have chosen not to answer.

Variable	**Response Values**	**Marginal Probability**
Gender	Male	57.7%
	Female	38.1%
	Non Binary	1.4%
	«missing»	2.8%
Age	12	18.3%
	13	20%
	14	19.8%
	15	22.7%
	16	13.8%
	«missing»	5.4%
Sexual Orientation	Heterosexual	66.4%
	Non Heterosexual	14.8%
	«missing»	18.8%
Migratory Background	No	72.9%
	My parents were born in another country	13.9%
	I was born in another country	13.2%
Self-Esteem	Low	20.8%
	Medium	41.6%
	High	37.6%
Social Support	Low	6.6%
	Medium	34.6%
	High	58.8%
Family Support	Low	6.8%
	Medium	27%
	High	66.2%
Daily Hours of Internet	Less than 2 h	8.2%
	2-3 h	19.7%
	3-4 h	23.2%
	4-5 h	18.7%
	More than 5 h	6.5%
	«missing»	23.7%
Previous CB Offending	Yes	26.2%
	No	73.8%
Previous CB Victimization	Yes	40.7%
	No	59.3%

1.3 Scope and Objectives of the Thesis

The main objective of this thesis is to apply causal-based computational methods for analyzing SG data for social science research purposes. This will serve as a practical demonstration of the enormous potential and advantages of conducting social or behavioral research through games. Moreover, it will allow us to study the complex phenomenon of CB and the most influential driving factors, through the presented case studies from the RAYUELA project. The central research question we seek to answer is whether SGs can be a valuable research tool in computational social science research.

In Chapter 2 of this thesis, we conducted a state-of-the-art analysis of the field of serious games and their intersections with AI. At the end of this chapter, we identify several research gaps, from which we define the **objectives** of this thesis:

1. **Develop a methodology to generate robust causal model structures (DAG) that unify expert knowledge and results of automatic structure learning algorithms.**

   Defining a causal model structure (DAG) is one of the first steps to work with PGCM. This objective will provide a tool that can be applied systematically to any case study where we have data and access to expert knowledge. We will validate the proposed methodology with the survey data from the RAYUELA project.

2. **Develop a general probabilistic framework to produce synthetic data that models human behavior in decision-based serious games.**

   Synthetic data generation is currently a very active area of research with many potential applications. It can anticipate data modeling and analysis, speed up the development process, or increase the available data (potentially improving analysis and enabling more advanced techniques such as deep learning). With this goal in mind, we will create a parametric tool to generate synthetic data for serious decision-based games, such as the one in the RAYUELA project. The tool should mimic human behavior in the game as realistically as possible.

3. **Build a causality-based computational methodology for analyzing Serious Game data.** To employ SGs as a social science research tool, it is crucial to develop a reliable and robust methodology to analyze the data obtained through the game. Throughout this thesis, we will also discuss why causal analysis is the most appropriate for this research. This is the core objective of the thesis and will be validated through the case study of the SG from the RAYUELA project.

4. **Identify relevant risk factors for cyberbullying using causality-based techniques.**

   This objective is based on the case studies from the RAYUELA project, which were used in the thesis to verify the proposed methodologies. The aim is to demonstrate, through empirical evidence, the positive outcomes achieved, how they can be utilized to research CB, and the significant factors involved. This makes the results of this thesis both methodological and empirical.

1.4 Thesis Outline

The structure and content of each chapter are briefly described below. Figure 1.2 shows a diagram of the general structure of the thesis and how the chapters are interconnected.

- Chapter 2 aims to identify gaps in research intersecting SGs with AI. To provide further context, a classification of the main fields of application of serious games is also proposed. The content of this chapter has already been published in a peer-reviewed scientific journal [189].
- Chapter 3 presents a methodology for unifying proposed causal model structures (DAG) from expert knowledge and data-driven algorithms. This methodology will serve in the following chapters as a first step to start working with PGCM.
- Chapter 4 presents and discusses the structure of a probabilistic simulator capable of generating synthetic data for decision-based serious games. To validate the proposal, we use the dataset of the survey of minors carried out in the RAYUELA project. The content of this chapter has already been published in a peer-reviewed scientific journal [188].
- Chapter 5 proposes and applies a causality-based methodology for analyzing data from serious games. This chapter draws on the results of Chapter 4, as this helped to accelerate the project and allowed us to devise data structures and pipelines prior to obtaining real data. To validate the proposal, we use the dataset of the experimental pilots carried out in the RAYUELA project.
- Chapter 6 is the last chapter of the thesis and concludes the work presented. First, the conclusions are discussed. Secondly, the contributions of this thesis are presented. Finally, promising lines of future work in the field of serious games and their use as a research tool are presented.

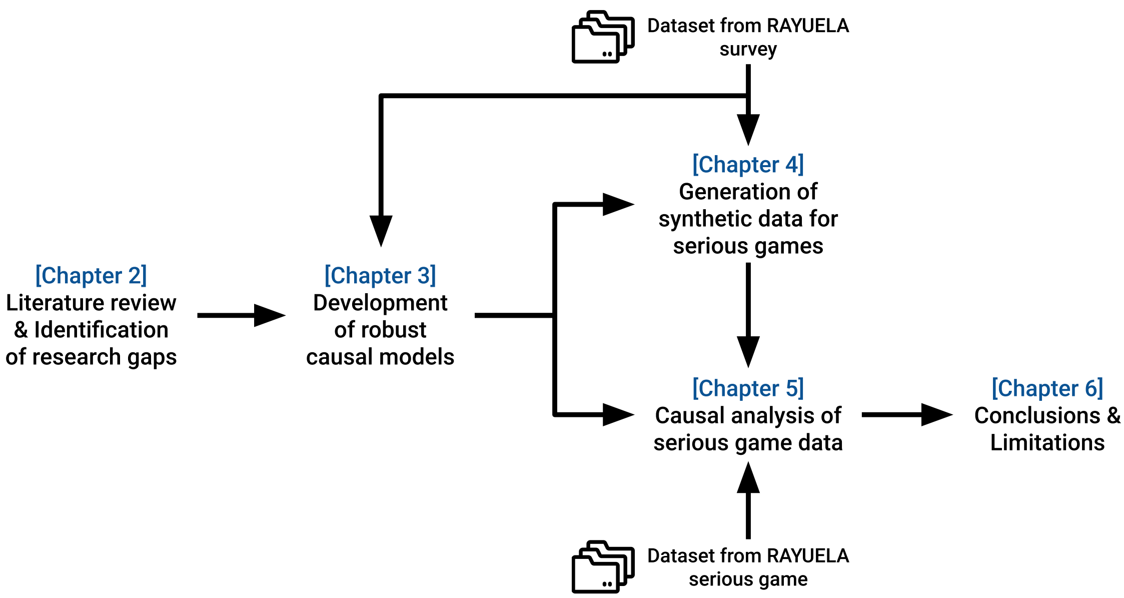


Figure 1.2: Diagram of the general structure of this doctoral thesis.

This dissertation covers material from the following publications:

1. J. Pérez, M. Castro and G. López, "Serious Games and AI: Challenges and Opportunities for Computational Social Science," in *IEEE Access*, vol. 11, 2023, pp. 62051-62061. DOI: 10.1109/ACCESS.2023.3286695. JCR: 3,900 Q2 (2022) - SJR: 0,926 Q1 (2022).

2. J. Pérez, M. Castro, E. Awad and G. López, "Generation of probabilistic synthetic data for serious games: A case study on cyberbullying," in *Knowledge-Based Systems*, Volume 286, 2024, pp. 111440, ISSN 0950-7051. DOI: 10.1016/j.knosys.2024.111440. JCR: 8,800 Q1 (2022) - SJR: 2,065 Q1 (2022).

3. J. Pérez, E. Awad, M. Castro, G. López, N. Bueno-Guerra, M. Reneses, M. Riberas-Gutiérrez and A. Gómez-Dorado, "A computational framework for understanding risk factors in cybercrime," *8th International Conference on Computational Social Science - IC2S2 2022*, Chicago (USA). 19-22 Julio 2022.

4. J. Pérez, V. Balmaseda, A.L. Urbistondo, E. Awad, M. Castro, G. López , "A child's play: an agent-based simulator to protect minors online," *8th International Conference on Computational Social Science - IC2S2 2022*, Chicago (USA). 19-22 july 2022.

5. J. Pérez, E. Awad, M. Castro and G. López, "Causality guiding survey analysis: a use case on cyberbullying," *9th International Conference on Computational Social Science - IC2S2 2023*, Copenhague (Denmark). 17-20 july 2023. URL: `https://openreview.net/forum?id=99EHn8T0r3L`.

6. S. Solera-Cotanilla, M. Vega-Barbas, J. Pérez, G. López, J. Matanza, and M. Álvarez-Campana, "Security and Privacy Analysis of Youth-Oriented Connected Devices," in *Sensors* 22, no. 11: 3967. DOI: 10.3390/s22113967. JCR: 3,900 Q2 (2022); - SJR: 0,764 Q1 (2022).

7. J. Fúster, S. Solera-Cotanilla, J. Pérez, M. Vega-Barbas, R. Palacios, M. Álvarez-Campana and G. López, "Analysis of security and privacy issues in wearables for minors," in *Wireless Networks*, 2023. DOI: 10.1007/s11276-022-03211-6 . JCR: 3,000 Q2 (2022) - SJR: 0,706 Q2 (2022).

8. C. Valero, J. Pérez, S. Solera-Cotanilla, M. Vega-Barbas, G. Suarez-Tangil, M. Alvarez-Campana and G. López, "Analysis of security and data control in smart personal assistants from the user's perspective", in *Future Generation Computer Systems*, Volume 144, 2023, Pages 12-23, ISSN 0167-739X, 10.1016/j.future.2023.02.009. JCR: 7,500 Q1 (2022) - SJR: 2,043 Q1 (2022).

9. G. López, N. Bueno-Guerra, M. Castro, M. Reneses, J. Pérez, M. Riberas-Gutiérrez, M. Álvarez-Campana, M. Vega-Barbas, S. Solera-Cotanilla, L. Bastida, A. Moya, R. Fernández, V. Vázquez, G. Zango and P. Vicente, "The H2020 project RAYUELA: a fun way to fight cybercrime," *Jornadas Nacionales de Investigación en Ciberseguridad - JNIC 2021*. 09-10 junio 2021. DOI: 10.18239/jornadas_2021.34.27

10. S. Solera-Cotanilla, J. Fúster de la Fuente, J. Pérez, R. Palacios, M. Vega-Barbas, M. Álvarez-Campana and G. López , "Análisis de problemas de seguridad y privacidad en wearables usados por menores," *VII Jornadas Nacionales de Investigación en Ciberseguridad - JNIC 2022*. 27-29 Junio 2022. ISBN 978-84-88734-13-6, pp. 209-215. URL: `https://2022.jnic.es/Actas_JNIC_2022_v11.pdf`.

11. C. Valero, J. Pérez, S. Solera-Cotanilla, M. Vega-Barbas, G. Suárez, G. López and M. Álvarez-Campana "Evaluando la seguridad y privacidad de los asistentes personales inteligentes: ¡Ojo con el juguete!," *VII Jornadas Nacionales de Investigación en Ciberseguridad - JNIC 2022*. 27-29 junio 2022. ISBN 978-84-88734-13-6, pp. 220-227. URL: `https://2022.jnic.es/Actas_JNIC_2022_v11.pdf`.

2 Serious Games and AI: Challenges and Opportunities for Computational Social Science

In this chapter, we will explore the latest developments in Serious Games (SG) and their interactions with Artificial Intelligence (AI), particularly for their application to Social Science research. Firstly, we will discuss the motivation behind this work. Secondly, we will describe the research methodology and the questions we aim to answer. Then, we will analyze the primary applications of SGs and the role played by AI in this field. Lastly, we will address the challenges and new research opportunities, and share our conclusions. The content of this chapter has already been published in a peer-reviewed scientific journal [189].

2.1 Motivation

Games have existed in all human societies and many other animal species. While some of the oldest board games, such as Go, Backgammon, or Checkers, are still played today, video games have become one of the most relevant forms of entertainment in our society. However, since the origin of games, they have had intentions and benefits beyond entertainment, such as teaching social norms, strengthening social bonds, or developing imagination and planning skills.

The rise of video games has had a remarkable social impact, helping to establish new social interaction and entertainment patterns [201, 85]. A prominent example of this trend is the gamification that our lives have experienced [130], from the workplace (e.g., *Habitica*, *LifeUp*) to romantic relationships (e.g., *Tinder*, *Grindr*) or education (e.g., *Kahoot!*, *Duolingo*). Video games, though high levels of interactivity, can raise motivation, engagement, and fun in almost any activity. While the video game industry is proliferating, the board game industry continues growing [25]. We can draw a clear conclusion: our society loves games, and they permeate many of our activities and interactions.

Some games —referred to as *serious games* (SG) [3]— are explicitly designed for a primary purpose beyond pure entertainment (e.g., training or learning new skills, conveying values, raising awareness). Nevertheless, being entertaining is part of their attractiveness. The first SGs were released in a wide range of formats, from sports to board games (e.g., *Monopoly*, *Suffragetto*), so this concept precedes the digital era.

The current re-emergence of SGs in industry and research [44] has coincided with the eruption of Artificial Intelligence (AI). Nowadays, and increasingly so, almost every entertainment element and digital product are at the service of data analysis and AI algorithms. Games are no exception [170], and in fact, the amount of data available via video games far exceeds other media.

AI has demonstrated its potential to analyze and better understand the functioning of our societies, interactions, and behaviors. The synergy between SGs and AI offers an exceptional window of opportunity for large-scale, non-invasive, and inexpensive social studies, leveraging their disinhibition and entertainment effects to collect large amounts of meaningful data. Moreover, games' casual and playful nature can help break conventional communication boundaries, encouraging participants to interact openly and discuss topics that might otherwise be complicated or too sensitive.

The primary objective of this chapter is to identify gaps in research and potential areas of inquiry at the intersection of SGs, AI, and Social Sciences. We also provide a comprehensive overview of SG research, identifying relevant application areas that differ from previous research focusing solely on a single application or sector.

Furthermore, while previous research has examined the potential applications of AI in SGs broadly [258], this work adopts a more focused approach by identifying specific ways in which AI can enhance SGs within Computational Social Science (CSS) research [140], specifically as novel research tools for understanding human behavior and society. Figure 2.1 provides a visual overview of the chapter's content to facilitate readers' comprehension.

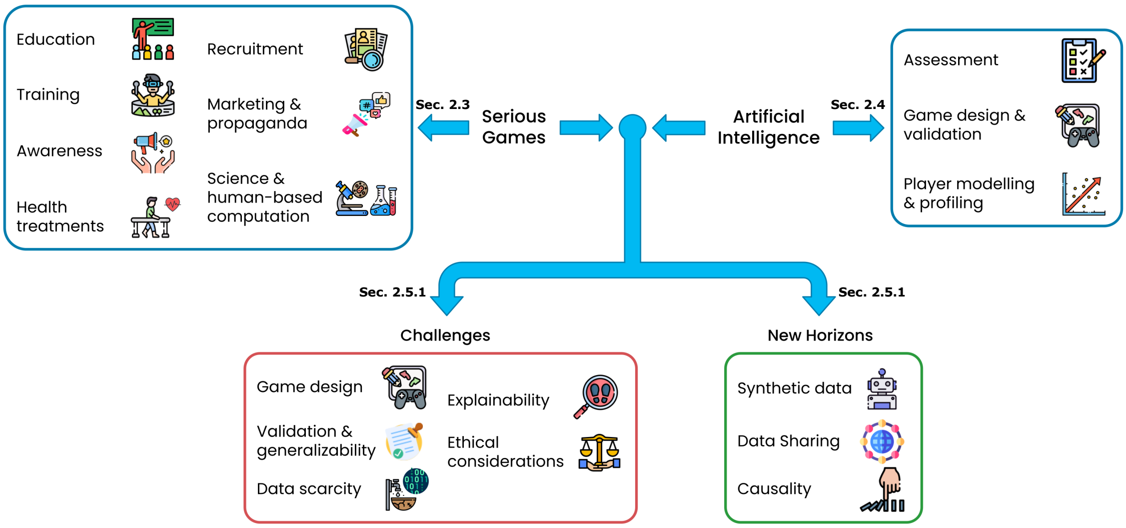


Figure 2.1: Graphical overview of the chapter. The blue boxes are the applications found for SGs and AI applied to them. The red box indicates the challenges faced by this union for its use as a research tool. And the green box indicates promising lines of work in this direction.

2.2 Methodology

This chapter aims to explore the potential of SGs in combination with AI, specifically focusing on the potential applications within CSS. To achieve this goal, we pose the following research questions (RQ):

RQ1: What are the main application fields of serious games?

RQ2: What are the main usages of AI for serious games with potential applications for computational social science?

RQ3: What are the main challenges and new horizons for the interaction of AI and serious games for computational social science research applications?

We have analyzed the available scientific literature to answer these research questions. We conducted our search using the Google Scholar search engine. The inclusion and exclusion criteria of the large number of research pieces identified were based on factors such as the number of citations, the popularity of the game, or the innovation of the approach (*i.e.*, pioneers in applying a novel technique or idea). First, to get an overview of the field, terms such as "serious games," "serious games applications", "serious games AI" and "serious games analytics" were used. Next, specific search terms were used for each of the identified serious games application areas. For example, "educational games", "game-based training", "social impact games", "game-based therapies", etc.

Finally, we screened studies based on their title, abstract, and full text to determine whether they met our inclusion criteria. We excluded studies that were not related to SGs or AI, were not published in English, or did not meet appropriate scientific standards.

2.3 Applications of Serious Games

This section aims to answer RQ1. The upsurge that SGs have been experiencing in recent years [44] may lead us to think this is a new phenomenon. However, the origin of SGs dates back to the 1970s. Clark C. Abt is credited for coining the term *Serious Games*, defining them as "games with an explicit and carefully thought-out educational purpose that are not intended to be played primarily for amusement". Clark C. Abt studied the potential of games as a vehicle for political, educational, or marketing ideas. Another of the leading figures in the history of SGs is Ian Bogost, the author of groundbreaking books on the theory behind them, such as "*Persuasive Games: The expressive power of video games*" [26].

Even though both concepts mirror the same social phenomenon, it is relevant to highlight the distinction between gamification and SGs. Gamification consists of using and integrating game elements into non-game concepts, while SGs refer to the design of entire games for non-playful primary purposes. Although both are concepts from the last century, they have recently resurfaced in the academic and commercial arenas.

Among the first serious video games, we find examples of how they are employed to convey particular values (e.g., *Captain Bible in the Dome of Darkness*, *The Oregon Trail* —Fig. 2.2a—, *Mobility*), disease awareness (e.g., *Captain Novolin* —Fig. 2.2b—), or military training (e.g., *Bradley Trainer* —Fig. 2.2c—). Nevertheless, the line between *regular* and *serious games* is quite blurred regarding games that convey specific beliefs or ideologies. Like any artistic or intellectual creation, video games always carry an implicit political and philosophical perspective. For example, popular video games such as *The Sims*, *Papers Please*, or *This War of Mine* convey strong political messages. However, they were not developed under the idea of being SGs. Focusing on those SGs that consider themselves as such and are designed for that purpose, we will now outline the main areas in which they have proven remarkably valuable.

(a) *The Oregon Trail* (b) *Captain Novolin* (c) *Bradley Trainer*

Figure 2.2: Screenshots of some of the first Serious Games in history.

Education

This category refers to SGs designed for the player to learn a series of concepts of a specific subject. To do so, the players must demonstrate their knowledge during the game and score their performances. Education has been one of the main focuses of action for SGs, based on the principle that learning while having fun is possible and efficient. This field has been extensively explored, and success and failure factors have been analyzed in depth [271, 196]. Prominent examples of building STEM skills might include *Garfield's Count Me In* [78], *Minecraft: Education Edition* (Fig. 2.3a) [169], the *Kahoot DragonBox* maths apps [57], and the *LightBot* coding apps [45]. SGs for educational purposes have also become popular in higher medical education [217], although some authors question their usefulness at such high educational levels, possibly complementing to more traditional learning methods [83, 92].

Training

Closely related to education, this category refers to games designed for players to learn and practice specific skills that will enable them to perform those actions in the real world with improved safety, confidence, and knowledge. This approach is widely used in companies where human failure is critical or costly. One of the best-known examples is flight simulators, such as *Microsoft Flight Simulator* (Fig. 2.3b) [168], where aspiring pilots must spend hours practicing before flying an actual commercial aircraft. There are notable examples of training healthcare professionals [255], mine site inductions and safety [235, 13], industrial engine maintenance training [174], cybersecurity trainees [99, 245], law enforcement agencies or military forces [214, 210, 274]. Another widespread use is training to manage complex business situations or administering teams and resources [138, 1, 273] or simulating trading operations [175]. This approach has also been used outside the workplace, for example, to teach people with intellectual disabilities how to use public transport [35] or to teach minors strategies for dealing with cyberbullying [33].

Awareness

Thanks to their high levels of interactivity, games evoke deep levels of empathy, making them an ideal vehicle to convey an awareness of relevant social issues. A classic example is *Darfur is Dying* (Fig. 2.3c) [205], which sought to tell the story of the humanitarian crisis in the Darfur region of South Sudan. We can find examples on a wide range of topics, such as cyberbullying [33, 69], drug consumption and trafficking [233, 70], gender equality [24], misinformation [119, 193], climate change [73], and environmental sustainability [53, 18, 232].

(a) *Minecraft: Education Edition* (b) *Microsoft Flight Simulator* (c) *Darfur is Dying*

Figure 2.3: Screenshots of some examples of Serious Games.

Health Treatments

This category is framed in healthcare but focuses more on patients than professionals. Well-known examples might be the *Wii Fit* and *Brain Training* games, which aim to have fun and stay fit (physically and mentally) simultaneously. Other notable examples can be found in the field of mental health therapy [68, 72], increasing self-efficacy and physical activity in people with chronic diseases [67, 27], helping the learning process and support of children with autism [77, 123], palliative care and memory training for older people or people with dementia [173, 172], and guidance and motivation in rehabilitation processes [154, 117, 20, 167]. Notably, in 2020 the US Food and Drug Administration approved the first video game-based treatment, *EndeavorRx*, targeting children between the ages of eight and twelve with certain types of Attention Deficit Hyperactivity Disorder (ADHD) [244].

Recruitment

If we combine games' interactivity with players' ability to make decisions in a well-designed environment, we can infer some behaviors or aspects of the players' abilities with reasonable confidence. For this reason, SGs have also been used to optimize the recruitment process in private companies [31, 134] and even military forces [11]. In these games, players face complex situations where they must make decisions and act under certain constraints or pressures. A recent notable example is the *CodinGame*[a] platform, where users practice their programming skills while playing, and many tech companies recruit profiles they find interesting. Another great example is the *GT Academy*[b], a competition in which the best players of a car racing video game have the opportunity to become professional drivers.

Marketing & Propaganda

When the game is developed primarily for marketing purposes, it is often known as an *advergame*. This category of games aims to convey ideas and create desires in a less intrusive and easily customizable way. It should not be confused with games that introduce advertising during gameplay for economic profit. The principal medium for these advergames is smartphones due to their proliferation, ease of development, and everyday use among young people.

Major brands such as *Volkswagen*, *Magnum*, *Chupa Chups*, and *M&Ms* have developed advergames. Related to the Recruitment category, in some cases, companies use these games to present and profile themselves to attract new employees and trainees, or to discover talent. Likewise, there have also been attempts to use video games as a tool to disseminate electoral campaigns, such as the video game *Corbyn Run* [47], or to encourage citizen participation in public decisions [163, 213].

Science & Human-based Computation

This category encompasses games to advance scientific knowledge in some way. One of the most common approaches is employing human players to perform seemingly trivial tasks, either too costly, too complex, or unfeasible for computers. These tasks may include labelling data, transcribing text, using common sense, or activities based on the human experience.

One of the first examples of this category was "*The ESP Game*" [251], in which players in pairs had to guess the photo labels created by their partner, to address the problem of creating complex metadata. Google's reCAPTCHA[c] used human players to label images while identifying legitimate users for accessing online resources. In "*EteRNA*" [64] players had to design RNA sequences that fold into a particular form. The solutions were evaluated to improve computer-based RNA folding prediction models. Other prominent examples might be "*Foldit*" [74] to predict protein structures, "*Eyewire*" [66] to map retinal neurons, "*MalariaSpot*" [160] to help diagnose malaria cases, "*Phylo*" [121] to optimize alignments of nucleotide sequences, or "*Quantum Moves*" [194] to improve how atoms move in a quantum computer.

2.4 Role of AI in Serious Games

This section aims to answer RQ2. Games have long been the test bed for AI as they provide a controlled environment with simple rules for algorithms to learn sophisticated strategies. AI is a rapidly evolving field that encompasses a vast array of techniques and methods, each with its own unique potential applications within the realm of SGs. Therefore, to narrow the contextualization to the point advocated by this chapter,

[a] CodinGame `https://www.codingame.com`

[b] GT Academy `https://www.gran-turismo.com/es/academy/`

[c] reCAPTCHA `https://www.google.com/recaptcha/about/`

our research has focused on exploring the specific areas of SGs where AI has played a prominent role. It has the potential to generate new insights and knowledge in the realm of CSS research.

In recent years, video games have been utilized as sources of vast amounts of player data, providing relevant information about human players, which is helpful inside and outside the game. However, SGs have particular objectives that we want to satisfy, and as such, the techniques and purposes of AI differ notably from those of *regular* games. Additionally, there is significant heterogeneity in these purposes of SGs, resulting in technical differences among them. Despite this heterogeneity, it is possible to discern the main branches encompassing all primary AI applications in SGs with significant potential for CSS research.

Assessment

Game-based assessment is a fruitful field in SGs [126], primarily used in education, training, and recruitment. Players are scored based on their knowledge or skills in a particular subject. [185] stipulate three primary purposes of assessment: (i) to assist learning (formative assessment), (ii) to evaluate the player's capabilities, and (iii) to evaluate programs. In general, collecting, analyzing, and extracting information through educational SGs is known as Game Learning Analytics [75].

The main difference with traditional evaluation methods or test gamification is that game-based assessment also uses in-game and interaction data (e.g., response times) to evaluate the player. Numerous authors have demonstrated the utility of using additional in-game data to evaluate students [150, 124, 8] or to predict learning results [9, 100]. It has also been used successfully to evaluate recruitment processes [136]. Although nowadays, they are more of a complement to the traditional exam-based assessment.

The techniques used are diverse, from simple descriptive statistics and correlations to supervised machine learning algorithms (e.g., linear regression, decision trees, Naive Bayes, Neural Networks) [7, 82]. Some papers use knowledge inference with Bayesian networks [147, 224], which explicitly allows the application of causality and latent state models, but flawed assumptions will negatively influence the results significantly.

This branch of AI applications in SGs is one of the most researched and developed, thanks to the technological push that is changing the way education is delivered. However, much work remains, particularly in demonstrating that they can outperform and generalize traditional approaches [118].

Game Design & Validation

Game design is planning the content, rules, and mechanics of a game to create valuable interactive experiences. The many artistic and technical factors in this process make any analytical information about the players extremely valuable. On the other hand, game validation employs data and evidence to verify and calibrate the game tasks and their difficulty. In the case of SGs, in addition to maintaining engagement, we also want to ensure that the game meets its primary objective (e.g., to train players in a particular skill, increase awareness of an issue, etc.).

Data-driven SG design and validation have flourished in academia in recent years, where we can find successful examples of using analytics to design, improve, personalize, and test these games [101, 237, 34, 76, 184]. This category is closely related to the previous one (Assessment), as it is almost essential to use data-driven validation during the development stage to calibrate how the game evaluates the players [246, 132]. Such analytics can go further to adapt the game's difficulty in real-time [98, 102] and detect player frustration [50].

In this category, the most commonly used techniques are descriptive statistics and visualizations [116, 127, 36], Randomized Control Trials (to test the usefulness of an intervention) [32, 34] and unsupervised machine learning algorithms (to find similar types of players and common patterns in the game) [34]. Using these analytical techniques enables creators and researchers to ensure that their games are entertaining, engaging, and well-designed to fulfill their objectives. Designing games to adapt to players dynamically is inseparable from player modeling so we will discuss this in the following subsection.

Player modeling & Profiling

Player modeling is the creation of computational models to detect, predict, and characterize the human player attributes that manifest while playing a game [265]. These models can be any mathematical representation, rule set, or probability set that maps parameters to observable variables and are built on dynamic information obtained during game-player interaction. On the other hand, player profiling usually refers to categorizing players based on static information that does not alter during gameplay (e.g., personality, cultural background, gender, age). Despite their dissimilarities, these concepts can complement each other, contributing to more reliable player models.

Recent advances in AI, specifically in large language models such as *GPT-4* and *LLaMa*, hold significant potential for advancing the modeling of player behaviors in SGs. These language models have demonstrated an ability to learn from vast amounts of natural language data and generate coherent and meaningful responses, enabling them to comprehend and simulate human language and behavior [107, 52]. Incorporating this technology into SGs can lead to a better understanding of human behavior and social dynamics, gaining insights into how individuals interact and make decisions. Moreover, incorporating them into the game design can provide more immersive and personalized experiences by tailoring the gameplay to the player's behavior, patterns, and preferences. However, there are challenges, such as the potential biases and ethical implications of using this technology.

W. Westera *et al.* [258] provides a comprehensive review of the use of AI in SGs. In particular, it analyzes the literature on player experience modeling using facial emotion recognition and text analysis using natural language processing techniques. Y. Y. Dyulicheva *et al.* [59] explores AI and immersive technologies in serious educational games, focusing on using AI to customize and personalize the player experience. It also examines AI-based SGs for teachers and students that do not require programming skills. D. Hooshyar *et al.* [104] conducted a systematic literature review that profoundly analyzes the computational and data-driven techniques used for player modeling between 2008 to 2016. As this is such a broad and promising field, the variety of algorithms used is immense: descriptive statistics and correlations, path/network analysis, supervised learning (e.g., Neural Networks, Linear Regression, Hidden Markov Models, Decision Trees), unsupervised learning (e.g., k-means, Linear Discriminant Analysis, Self-Organizing Maps), probabilistic algorithms (e.g., Bayesian / Markov Networks), evolutionary methods (e.g., Genetic algorithms), reinforcement learning methods (e.g., Multi-armed bandits), etc. Some proposals based on latent mixture models (e.g., Bayesian hierarchical models) [236, 125] yield more interpretable results, but flawed model assumptions will negatively influence the results.

Player modeling can be helpful both inside and outside the game itself. The most straightforward goal is to improve the game design, tailoring the content to increase engagement and enhance the gaming or learning experience [49]. We find some prominent examples in the *regular* video game industry, such as *Left 4 Dead* [145], where an AI tracks player behavior and adapts future waves of enemies to maintain rhythm and tension. Perhaps the most famous example is the video game *Silent Hill Shattered Memories* [226], which uses a psychological approach where an AI system tries to manipulate players' emotions using the *Five Factor Model* of personality [56]. Outside the game, the most common use of player modeling in the gaming industry is for personalized marketing campaigns, since the commercial sector is very interested in understanding customer behaviors and preferences. In these cases, the games are often presented as free to play in exchange for an intrusion into personal privacy [58]. Besides the "advergames" discussed in the section Marketing & Propaganda, a famous example outside SGs is *Farmville* [260], which monitored the players' behavior to adapt *Amazon* marketing campaigns to them.

In academia, especially in psychology, experiments have been conducted using (*regular* and *serious*) games for research, but primarily focusing on analyzing how the player's personality is projected in the gameplay patterns [137, 267, 91, 54, 165]. Nevertheless, studying psychological characteristics or phenomenology using SGs seems to be an up-and-coming field, especially when introducing AI techniques into the equation.

2.5 Challenges and New Horizons

This section aims to answer RQ3. In the previous sections, we have discussed the main applications of SGs and the current trends in their synergies with AI. In this section, we emphasize the great potential of SGs partnering with AI to serve as research tools, particularly in CSS research, examining the most critical challenges and promising new lines of work.

As discussed in the Motivation, games allow research to be entertaining, provide high levels of empathy, and can lead to a disinhibition effect that is highly sought after in Social Science investigations. Games can evoke dynamic and complex emotions in players [265]. Besides, these complex reactions are difficult to capture with the traditional approaches. Therefore, using SGs as research tools, supported by novel AI techniques, can contribute to advancing many Social Sciences fields in an entertaining and non-invasive way [153].

The latest advances in AI allow us to analyze vast amounts of data and find patterns or behaviors that would be very difficult to observe with traditional methods. So far, the main application given to large AI models that study our interactions through social networks and personal data is for marketing purposes [157]. This practice has been done almost since the beginning of social networks without considering the negative social consequences it could have, particularly for children and adolescents [122, 2]. With this work, we also aim to help achieve more laudable goals using AI.

2.5.1 Challenges

Through our research, we have identified the following challenges:

- ***Game design***: Game design is a creative process that involves a wide range of artistic and technical skills. Game designers must consider story, character development, game mechanics, user interface, graphics, sound, and more. The harmonious integration of these elements is necessary for creating a compelling game experience that engages players. However, game design is not a purely scientific process but often relies on the intuition and creativity of the designer. It is challenging to quantify and standardize design elements, making it difficult to predict the success of a game or to identify what makes a game genuinely engaging. Therefore, the success of a SG as a research tool depends largely on its design and playability.

- ***Validation and generalizability***: The use of SGs as a research tool is becoming increasingly popular in a variety of fields, including education, health care, and Social Sciences. However, one of the biggest challenges researchers face when using SGs is demonstrating the validity and reliability of their findings. Unlike traditional research methods, SGs involve complex interactions between players and the game environment, making it challenging to ensure the consistency and accuracy of data collection. While some areas of SG research have shown promising results, such as game-based Assessment, it is still necessary to establish standard validation procedures for each game and its intended purpose. Each game is unique, and the research questions and hypotheses behind it must be carefully designed and tested. This necessitates a personalized validation approach, which can be time-consuming and resource-intensive.

- ***Data scarcity***: Training and enhancing AI models require a large amount of high-quality data. However, academic experiments in SGs often rely on small and biased data sets, which limits the effectiveness of the models. Consequently, research outcomes and findings may be inaccurate, prejudiced, or biased, hindering the research process. To overcome this challenge, researchers must find ways to increase the number of participants, collect diverse and representative data, and optimally utilize the available data. This may involve partnering with the industry to access larger data sets, using synthetic data generation, or developing standardized data collection and management methods. Moreover, researchers must establish appropriate methods for sharing sensitive data while ensuring the anonymity and privacy of participants. This practice can be immensely beneficial for Social Science research [30].

- ***Explainability***: The increasing complexity and opacity of advanced AI tools pose significant challenges to their use in studying human and social behavior. These advanced AI models are often regarded as "black boxes" since their inner workings are not completely transparent or understandable. Computer science has historically emphasized prediction over explanation, further complicating the use of AI tools in social research. Understanding the results produced by AI models is crucial for studying human behavior. Although progress has been made in developing explainable AI techniques such as LIME or SHAP, many challenges remain. One significant obstacle is the trade-off between model complexity and explainability. More complex models may provide better predictions, but they may also be more difficult to interpret. Furthermore, explainable AI techniques' suitability must be evaluated for specific research contexts, especially for small or highly heterogeneous data sets. Addressing these challenges requires interdisciplinary collaboration between AI researchers and social scientists. This includes developing best practices for incorporating explainable AI into SG experiments, evaluating the appropriateness and accuracy of different techniques for specific research contexts, and balancing the trade-off between model complexity and explainability.

- ***Ethical considerations***: Using personal data in SG experiments raises significant ethical considerations that cannot be overlooked. Clear and unambiguous ethical guidelines are necessary to ensure that the potential benefits outweigh the risks, and to prioritize the safety and well-being of participants. This is especially crucial when dealing with sensitive data from marginalized populations or children. To establish these ethical standards, it is necessary to develop interdisciplinary collaboration between computer scientists and social scientists, who often have different approaches to research ethics [207]. While computer scientists often prioritize technical aspects of privacy and security, social scientists are more concerned with the impact of research on human subjects and communities. Developing ethical guidelines for SGs experiments involving AI requires a deep understanding of the specific ethical concerns associated with this area of research. For instance, researchers must thoughtfully consider how to obtain informed consent from participants, ensure the privacy and security of personal data, and mitigate potential harm resulting from using AI models in research.

2.5.2 New Horizons

Despite the challenges mentioned above, we can also find promising new horizons and future lines of work regarding the interaction between SGs and AI:

- ***Synthetic data***: The AI field has extensive experience developing agents that can play games at a high level and even outperform human players [94]. However, in recent years we are experiencing a wave of generative AI techniques that can create new and diverse content such as images, music, or text. These techniques have shown remarkable potential in creating data miming human behavior and preferences [177, 262]. Combined with data augmentation techniques (e.g., agent-based *imitation learning* [108]), they can enable us to produce large volumes of diverse data that closely mimics real-world human behavior to train other AI models and improve their performance. This opens up the possibility of exploring new research questions and hypotheses that were previously difficult or impossible to address due to limitations in data availability. Within a SG project development, synthetic data enables refining game design, prepare data structures and pipelines, and accelerate development times even ahead of actual data collection.

- ***Data sharing***: The emergence of SGs and their potential to address social issues and promote learning has opened a new avenue of research for Social Science. SGs offer a unique opportunity to collect data from many participants in a more controlled and interactive setting. One of the challenges that CSS has faced is finding and sharing open data, especially from private companies [141]. Well-designed SGs can tackle the issue without collecting sensitive data. In addition, using anonymization and privacy-preserving algorithms has proven very useful in recent years, allowing researchers to extract meaningful insights from data while protecting participants' privacy. This approach has the potential to help researchers overcome one of the significant obstacles in Social Science research: the need to balance ethical considerations of participant privacy with the scientific need for large, diverse data sets.

- ***Causality***: The Social Sciences have traditionally aimed to explain human behavior through interpretative means, with randomized controlled trials being the gold standard for establishing causal relationships. However, computer scientists have typically prioritized the development of accurate predictive models that may not necessarily represent the underlying causal mechanisms of a phenomenon. In recent years, there has been a growing interest in computational causality techniques [166, 182] that can help us estimate how systems work with greater robustness and generate more plausible alternatives, even for observational data [151]. This approach can help us better understand human behavior by combining predictive and explanatory approaches and building more comprehensive models that account for the complexity and nuance of human decision-making. In the field of SGs, explanatory models can help us identify the mechanisms by which behavior is reflected in the game and account for biases in the data. Conversely, predictive models can make accurate predictions when explainability is not a critical constraint. By developing more comprehensive models that account for causal mechanisms and predictive accuracy, we can generate insights to inform the design of more effective and impactful research in CSS [103].

2.6 Conclusions

Gaming has been a part of human and animal leisure and development for thousands of years. Incorporating AI into SGs can enhance the user experience by offering personalized gameplay and improving the analysis and understanding of collected data. Furthermore, player disinhibition and engagement effects enable the collection of highly relevant data on player behavior. This combination provides opportunities for Social Science research, creating an interactive and appealing environment to study complex social phenomena through observation, experimentation, and analysis.

In this chapter, we have presented a systematic overview of the main application domains of SGs. Secondly, we have analyzed the successful interactions of SGs with the field of AI, providing a comprehensive overview of the overlap between these two fields. Thirdly, we have identified critical challenges and research gaps in this overlapping area. Finally, we have proposed three promising new research areas concerning synthetic data, data sharing, and causality that will be explored in the following chapters of this thesis.

Further research is essential to develop Serious Games (SGs) for social science research that ensure data privacy and ethical compliance while utilizing advanced AI models capable of comprehending complex human behaviors in gaming environments. To achieve this, computer scientists must collaborate with professionals in psychology, sociology, anthropology, and others. This interdisciplinary approach can lead to innovative ways of comprehending and using SGs. The continued use of games to facilitate scientific discoveries can yield significant results. In conclusion, the convergence of AI and SGs offers a distinctive chance to examine social phenomena in a more intricate and sophisticated manner.

3 Human-Machine Consensus for Robust Causal DAG Generation

This chapter outlines a methodology for unifying expert knowledge and data for creating robust and reliable causal Directed Acyclic Graphs (DAG). First, we discuss the motivation behind the search for consensus between experts and data in this field. Second, we describe in detail the proposed methodology and success metrics. Third, a case study is conducted using real data collected in Spanish schools from minors regarding cyberbullying (CB). Finally, the results are discussed, and the conclusions reached are presented.

3.1 Motivation

Conventional statistical techniques based on the search for correlations are ubiquitous in social science research. However, they can sometimes be problematic because they rely on spurious correlations and fail to account for data biases [202]. This is a common issue in research where it is impossible (or unethical) to perform a randomized control trial, and only observational studies are feasible. If data processing and analysis are not done carefully, there is a high risk of exaggeration, minimization, or unintentional reversal of causal effects.

In recent years, we have witnessed a flourishing of data-driven causal inference research and techniques [89]. This scientific and modeling philosophy promises to understand complex issues better, unraveling the intricate webs of relationships between variables and shedding light on the underlying mechanisms governing these phenomena. It also motivates researchers to keep a skeptical attitude about data acquisition and manipulation and to explicitly present assumptions and hypotheses, fostering fruitful discussion [86].

Establishing causality is essential in research and policy development, where traditional methods may have limitations and fail to provide solid evidence consistently. CB represents an excellent example of the kind of research where interventional studies are impossible (unethical) and where the interpretability and explainability of potential conclusions are especially important. This stresses the need for a causal approach to develop effective prevention and intervention strategies for CB. Some recent examples in the literature have already attempted to address CB from a causal inference framework [43, 10] and identify causal structures in cyberstalking behaviors [156].

Recent interest is in applying Machine Learning techniques to detect CB [240, 266]. While these methods seem to be a promising alternative to traditional statistical methods, they lack three key ingredients that we consider mandatory for such a sensitive subject: (i) the ability of a systematic quantification of the uncertainty in the analysis, (ii) the lack of a causal-oriented perspective aimed at deploying interventions or informing policymakers; and (iii) the absence of a systematic procedure to integrate expert knowledge with available data. These limitations hinder the effectiveness of Machine Learning approaches in providing comprehensive insights into complex causal relationships.

To address these challenges, we advocate using Probabilistic Graphical Causal Models (PGCM) [238], also known as causal Bayesian Networks (BN). This intuitive framework incorporates probabilistic, structural, and graphical aspects of causal inference. These models provide a solid basis for analyzing intricate causal structures nuancedly, improving our ability to understand complex issues such as CB more effectively.

However, the initial step in modeling causal relationships involves determining a structure for the PGCM, commonly known as causal DAG. This process poses a significant challenge, as identifying an optimal DAG that accurately represents the causal relationships within the data is non-trivial. Although many data-driven algorithms for automatic structure discovery have been proposed [252, 268], their reliability remains a concern, particularly in situations with limited available data and under causal assumptions [128, 218].

Without substantial data, relying solely on expert knowledge is a common practice. However, this approach poses a risk as it may introduce biases inherent to the experts' field or current paradigm. Striking a balance between leveraging expert knowledge and incorporating data-driven algorithms is essential for obtaining robust causal DAGs that accurately reflect the complexity of real-world systems and exploit the benefits of both approaches.

In light of these considerations, we propose a novel methodology for reaching consensus among expert knowledge and data-driven algorithms. This hybrid strategy aims to combine the strengths of both sources of information, mitigating the limitations associated with each and yielding robust causal DAGs upon which to work on real problems. Numerous proposals in the literature exist for constructing DAGs, combining expert knowledge and data-driven algorithms [115, 164, 37, 12]. In this chapter, we have attempted to simplify and systematize a bidirectional information flow between these two actors.

3.2 Methodology

This section describes the method proposed to combine expert knowledge with the data to obtain the causal DAG structure. We will also detail the analyses performed to interrogate the resulting model in the case study and the data collection process.

3.2.1 Causal DAG Construction

The generalized form (without causal implications) of the PGCM is the BN, a graphical model representing the joint probability distribution of random variables. A BN comprises a DAG and conditional probability tables (CPT) [131, 178]. Given a DAG, namely G, and a joint probability distribution P over a set of discrete variables $X = \{X_1, \ldots, X_n\}$, we can say that G is modeling or representing P correctly if there is a one-to-one correspondence between the variables in X and G such that Eq. (3.1) is satisfied. Where pa_i are the direct *parent* nodes of x_i in G, and $P(x_i \mid pa_i)$ is the conditional probability distribution (CPD).

$$P(x_1, \ldots, x_n) = \prod_i P(x_i \mid pa_i) \tag{3.1}$$

Conditional probabilities play a crucial role in establishing causality, as they allow us to compute the probability that one event will occur, given that another has already occurred. However, interpreting a BN as carrying conditional dependence and Independence assumptions does not necessarily imply causality; a valid graph set can be constructed from independent variables with any ordering, not necessarily causal or chronological [179]. Determining causal relationships requires additional information or assumptions beyond the data, such as experimental manipulation or adjustment for confounding variables in observational studies [183]. In this work, we assume that the expert-defined DAG is known and its connections have a causal interpretation since it has been constructed based on insights gained through practical experience and literature study [199].

To introduce causality into BNs (i.e., work with PGCM) we must make the following **assumptions** [215]:

i) There is a DAG, namely G, representing the causal relationships among the variables used.

ii) The causal Markov condition: The joint probability distribution of the variables used obeys the Markov property in G.

iii) Faithfulness: The joint probability distribution satisfies *exclusively* the conditional independence relations implied by the causal Markov property.

Once we have exposed the formal semantics and assumptions of the model, the proposed **methodology** to build the causal DAG that brings together expert knowledge and data consists of the following four **steps**. A graphical representation of these steps is shown in Fig. 3.1.

1. **Initial proposals**: Structure learning algorithms and the experts build their first proposals without transmitting information between these actors. The experts' proposals may be based on their experience, previous literature, or common sense. From this step, we obtain a series of potential causal DAGs.

2. **Consistency Causal Restrictions**: Together with their proposals, the experts forbid a set of causal connections that will be included as initial conditions to the algorithms. Naturally, one must be completely confident when prohibiting these connections. These restrictions come from common sense or widely accepted knowledge in the literature.

3. **Suggested Causal Arrows**: Conversely, the experts analyze the initial algorithms' proposals (without restrictions) and consider whether any causal relationships found should be incorporated into their second proposal.

4. **Quantitative comparison and consensus graph**: Once the new algorithmic proposals (with restrictions) and the second proposal of the experts have been obtained, a quantitative comparison (Sec. Metrics) is made to identify the model that best explains the available data. Finally, the best-performing models for the specified metrics are selected, and the implications of each proposal are discussed.

Automatic structure learning algorithms

Given a dataset, we can define two main BN structure learning approaches: (i) the constraint-based method, which induces the graph from the results of conditional independence tests on data over triplets of variables, and (ii) the score-based method, which maximizes a score function relative to data, measuring the goodness of each structure [252].

It is worth emphasizing that any automatic structure learning process has significant limitations. First, the algorithms rely heavily on the quality and quantity of data available and do not guarantee identifying an actual causal structure. Therefore, we must carefully evaluate the strengths and weaknesses of each structure proposed and compare their ability to generalize to new data. Second, selecting the final causal DAG is a critical step requiring statistical rigor, expert knowledge, and common sense. In general, learning a DAG using structure learning algorithms suffers from the following difficulties [88]:

- Involves solving expensive multi-variable optimization problems
- Requires numerous independence tests, the power of which decreases as dimensionality increases
- The solutions found are a Markov equivalence class of graphs involving the same conditional independencies, which would be insufficient when searching for a fitting set, since changes in the directionality of the edges or the presence or absence of certain edges could lead to an erroneous choice of covariates for the fit.

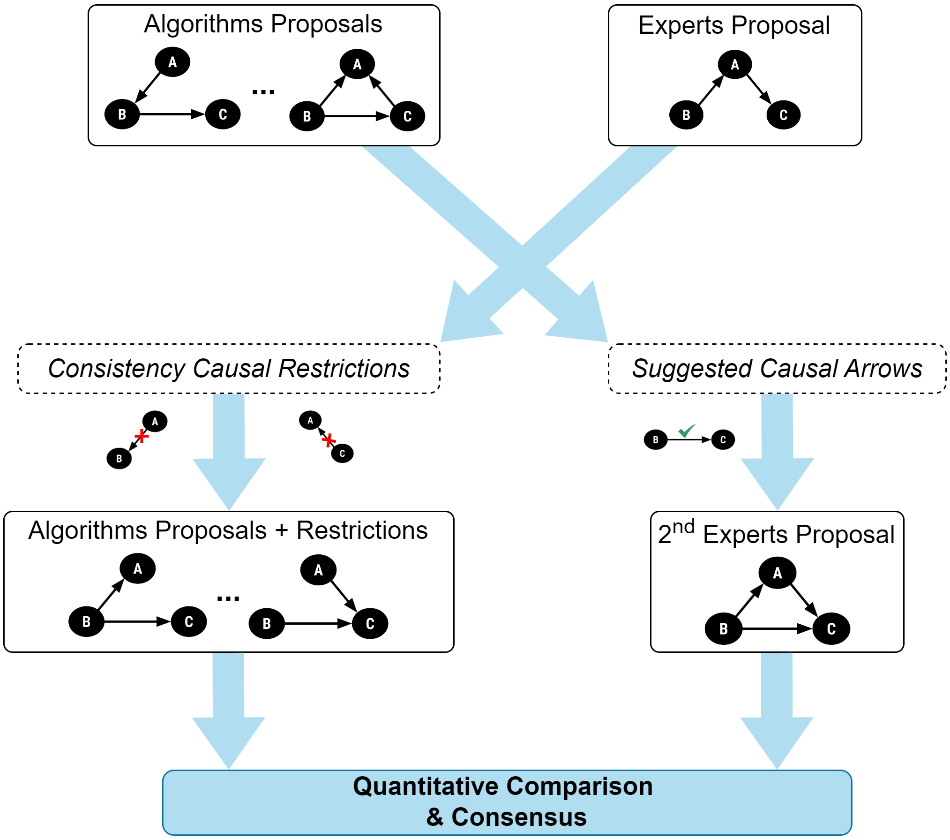


Figure 3.1: Conceptual graphical representation of the steps involved in the causal DAG construction. First, the experts and the structure learning algorithms make their proposals. Then, the experts impose restrictions and revise their initial proposal based on the relationships found by the algorithms. Finally, a quantitative comparison is performed and a consensus network is reached.

In this work, we use three structure learning algorithms from the GeNIe Modeler[a] software, among the many available in the literature. Specifically, we opt for the algorithms that allow us to include restrictions manually. The selected algorithms are explained below.

Bayesian Search: This score-based algorithm is one of the earliest and most widely used in causal discovery. It was introduced by Cooper and Herkovitz [46] and refined by Heckerman [95]. In essence, it follows a hill-climbing procedure guided by a scoring heuristic (which in GeNIe Modeler is the log-likelihood function) with random restarts.

PC: The PC algorithm is also one of the earliest and most popular in causal discovery. It was introduced by Spirtes [231]. It uses classical independence tests to analyze the available data and then infers the structure that has generated it. Thus, it belongs to the constraint-based category.

Greedy Thick Thinning (GTT): This score-based algorithm is based on the Bayesian Search and was proposed by Cheng et al. [42]. It starts with an empty graph and repeatedly adds the arc (without creating a cycle) that maximizes the marginal likelihood of the structure given the data until no arc addition produces a positive increase. Subsequently, it repeatedly deletes arcs until no arc deletion produces a positive increase in the marginal likelihood of the structure given the data.

[a] https://www.bayesfusion.com/genie/

Metrics

Each causal DAG candidate is evaluated using the metrics described below to perform the last step of the proposed methodology, the quantitative comparison of structures.

Log-Likelihood (LL) Score: The LL score measures how well the DAG fits the observed data. It calculates the logarithm of the likelihood function, which represents the probability of the observed data (x) given the network structure and parameters (Θ). Formally it is expressed as $log(\mathcal{L}(\Theta \mid x))$. Higher LL scores indicate a better fit.

Bayesian Information Criterion (BIC): The BIC is a widely used metric that balances the goodness-of-fit and model complexity. The BIC score is calculated using the LL score and penalizing the number of parameters in the model to avoid overfitting [39]. Formally, it is expressed as shown in Eq. (3.2) where k represents the number of parameters estimated, n is the number of data points, and $\widehat{L}$ is the maximized value of the likelihood function of the model. Lower BIC scores indicate a better trade-off between fit and complexity.

$$\mathrm{BIC} = k \ln(n) - 2\ln(\widehat{L}) \tag{3.2}$$

K2 score: The K2 score is a particular case of the Bayesian Dirichlet score and is commonly used in BN structure learning. It is based on the likelihood of the data given the network structure and parameters [46]. It incorporates prior probabilities and can handle small sample sizes. It also penalizes model complexity but to a lesser extent than BIC. Higher K2 scores indicate a better fit.

Correlation score: This score measures how well the DAG represents the correlations in the data using the d-separation property [79]. For each pair of variables in the data set, a correlation test (typically chi-square) is performed. Then, we test whether the same pair of variables are d-connected in DAG. Finally, a classification metric (e.g., F1 score) is calculated using the correlation test as the true value and the d connections as predicted values. Higher correlation scores indicate greater agreement between the correlations of the data and the DAG.

Overall, the choice of metric depends on the specific application and the goals of the analysis. The model's complexity, availability of prior knowledge, and sample size must be considered. Note, however, that the aim of designing the *best* DAG is not to achieve quantitative predictive accuracy but, more importantly, to capture the hypothetical causal relationships among variables in the most systematic and unbiased way.

Each metric has its advantages and disadvantages. The LL score tends to favor fully connected network structures, sometimes falling into overfitting. The K2 score is often used as a baseline score for comparison but relies on the choice of some hyperparameters. The BIC is commonly used when the goal is to find a model that fits the data well but is not too complex (Occam's razor approach). The correlation score tests only conditional independence assumptions but does not examine how well the data fits.

In the particular case of the LL score, we have used a k-fold cross-validation approach to compare the candidate models. This technique is widely used in the machine learning literature to avoid overfitting [111]. It consists of dividing the dataset into k segments and training the model k times with different segments as the test set. The LL score is the only proposed metric that requires the model parameters as input to perform the evaluation. Therefore, it is the only metric that justifies using k-fold cross-validation.

3.2.2 Performing Causal Tasks

Once we have selected a causal DAG with which we are satisfied and it has been trained with the available data, we can perform different analyses to interrogate and validate the model. For example, effect estimation (*If we change A, how much will it cause B to change?*), attribution (*Why did an event occur?*), counterfactual estimation (*What would have changed if we had measured a value in A different from the observed value?*), or prediction (*What will we get as a result of a new data entry?*) [182].

In this thesis, we are interested in the task of **effect estimation**. For instance, estimate the effect of different interventions on the risk of suffering CB. To this end, we computed the Average Causal Effect (ACE) [161], also known as Average Treatment Effect (ATE), to answer the following question: *How much does a certain target quantity differ under two different interventions?* Using the do-calculus notation [181], the ACE can be written as in Equation (3.3). The do operator symbolizes an intervention and can be defined as in Equation (3.4), where S is the sufficient adjustment set, the set of variables in the DAG that block all the confounding paths from treatment T to outcome Y and meet the requirements of the back-door criterion [180]. The key strength of this method is that it enables us to estimate the interventional probability distribution of the outcome from observational probability distributions. This method provides a single value representing the influence of specific interventions. Figure 3.2 illustrates the difference between conditioning ($P(Y \mid t)$) and intervening ($P(Y \mid do(t))$).

$$\mathrm{ACE} = \mathrm{ATE} = \mathbb{E}[Y \mid \mathrm{do}(T := A)] - \mathbb{E}[Y \mid \mathrm{do}(T := B)] \tag{3.3}$$

$$P(Y \mid \mathrm{do}(T := t)) = \sum_{s \in S} P(Y \mid S = s, T = t) P(S = s) \tag{3.4}$$

If the treatment variable T has more than two options, the ACE can be calculated for a pair of values of interest or the values that give the most extreme results. We expressed the ACE results as the percentage difference between treatment A and B ($\Delta P = P_A - P_B$), and equivalently as an Odds Ratio (OR) following Equation (3.5). Having obtained the results for each variable, we can rank them. Larger ACE values indicate a greater causal influence on the outcome.

$$\text{Odds Ratio } (OR) = \frac{P_A(1 - P_B)}{P_B(1 - P_A)} \tag{3.5}$$

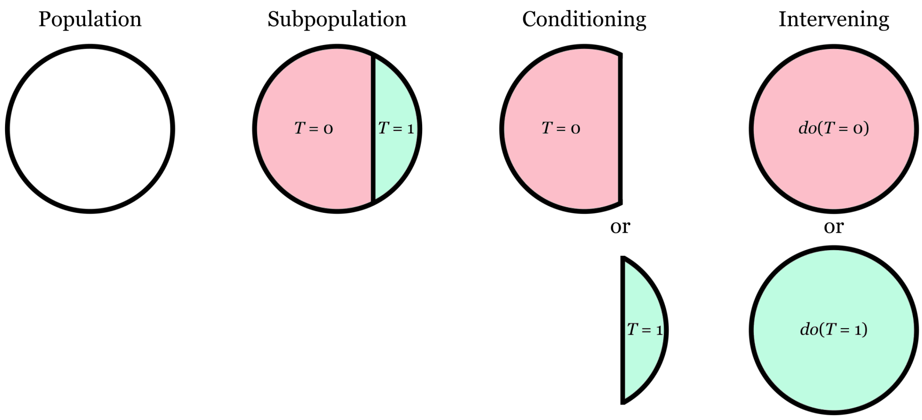


Figure 3.2: Illustration of the difference between conditioning and intervening. Adapted from [171].

3.3 Case Study: Cyberbullying Causal Model

This section presents a case study where we apply the proposed methodology for generating robust causal DAGs. Firstly, we will provide detailed information about the collaboration with RAYUELA's CB experts. Secondly, we will show the proposals from the experts and the structure learning algorithms. Then, we will apply the methodology to modify the proposals and compare them quantitatively. Finally, we will perform a causal analysis on the winning models to estimate the variables with greatest influence on CB victims and verify the robustness of the winning causal DAGs.

The data used in this case study were collected through a representative survey of children in schools in Madrid (Spain). Section 1.2.1 details the characteristics of the dataset and its collection. Table 1.1 provides the variable values and their percentage of occurrences (i.e., marginal probability).

It is worth mentioning again that, among all the cybercrimes analyzed in the RAYUELA serious game, we have decided to focus on CB since it is the only one for which a validated questionnaire [28] could be carried out on the participating minors. The results of this questionnaire represent a "ground truth" for the analyses carried out throughout this thesis.

3.3.1 Collaboration with Cyberbullying Experts from RAYUELA

In this case study the expert knowledge comes from members of the RAYUELA project consortium. The work package 1 in the project was concerned with creating a knowledge base on the drivers of cybercrime in young people. For this purpose, this team conducted research that sought to understand both the pathology and physiology of online behaviors, characterizing the victims and offenders of the forms of cybercrime considered, as well as the modus operandi.

This team included members from Universidad Pontificia Comillas (Spain), University of Ghent (Belgium), University of Tartu (Estonia), University College Limburg (Belgium), Bratislava Policy Institute (Slovakia), Ellinogermaniki Agogi (Greece), Polícia Judiciária (Portugal), Valencian Local Police (Spain), Police Service of Northern Ireland (United Kingdom), Estonian Police and Border Guard Board (Estonia).

Regarding the crime of CB this team conducted a total of 33 interviews (8 offenders, 12 victims, and 13 experts) [197] and analyzed 46 court sentences [198]. As a result, the team acquired a profound understanding of the issue, which has been used on several occasions throughout the remainder of the project and has been documented in the cited technical reports. To interact with the expert knowledge when constructing the causal DAGs, we held discussion sessions with some team members from Universidad Pontificia Comillas, as they were the leaders of this work package.

3.3.2 Causal DAG Construction and Comparison

In the first step of the proposed methodology, the experts and the structure learning algorithms create their initial proposals (Fig. 3.3) without transmitting information between these actors. That is, the experts do not have any information about the algorithms' results, and the algorithms are trained without using restrictions or forced connections from expert knowledge.

Figure 3.3a shows the structure we have used as a baseline where all nodes are directly connected to the target variable. We have called this structure "naive". That is, it is assumed that no interdependence exists between any of the nodes. This structure is similar to classical statistical analysis in the social sciences, examining each variable separately. Building upon the naive structure, Figure 3.3b shows the experts' first proposal. They added causal relationships from the *age* node to the *daily hours of Internet* and *CB awareness* and an additional arc from *gender* to *CB awareness*.

Figures 3.3c, 3.3d, and 3.3e correspond to those obtained through the structure learning algorithms. We can observe that the algorithms' proposals go against common sense on many occasions. For example, in Figure 3.3c, *sexual orientation* causes *migratory background*, and there are nodes causing *age*; in Figure 3.3d, numerous variables cause *gender*, etc. Moreover, in these algorithms' proposals some nodes are left unconnected in the network, sometimes even the target node (*Cyberbullying Related Situations*), which neither helps us in modeling nor in understanding the issue.

In the second and third steps of the methodology, the experts impose restrictions and forced connections on the training process of the algorithms, based on common sense and input from the experts. These expert impositions are shown in Table 3.1). For example, looking at the first row of the table, the experts assume that *gender* cannot have a causal effect on *age*, *sexual orientation*, or *migrant background*. And on the basis of the research, they force the link between *gender* and *CB-related situations*.

As a result, we obtain refined versions of the initial proposals (Fig. 3.4). For ease of visualization, the connections that differ between the proposals have been highlighted in blue. Figures 3.4a, 3.4b, and 3.4c correspond to those obtained from the structure learning algorithms, using restrictions and forced connections from the experts. The structure shown in Figure 3.4d corresponds to the second experts' proposal after analyzing in detail the initial results from the algorithms. This structure is the most densely connected, so it can be expected to be penalized more by some of the metrics used, such as BIC or K2.

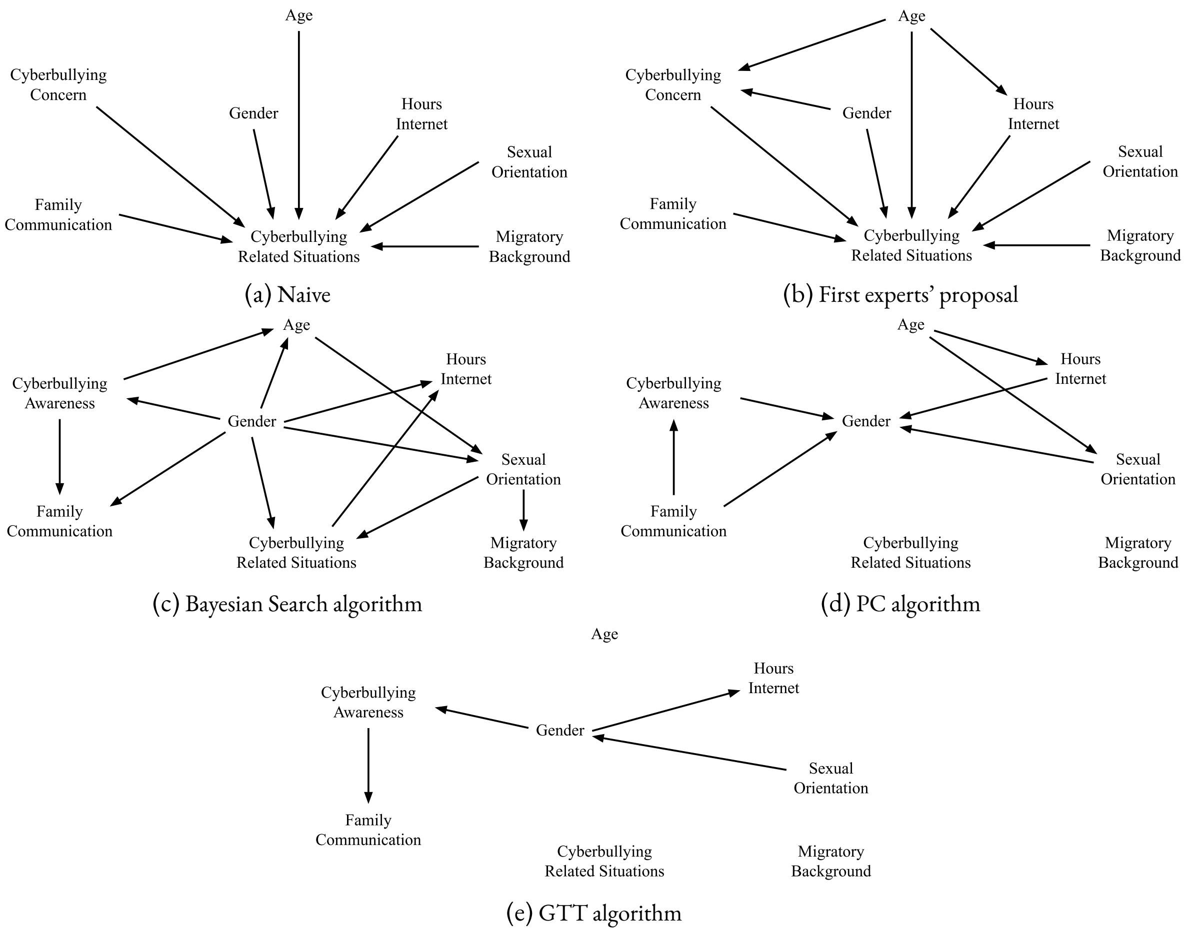
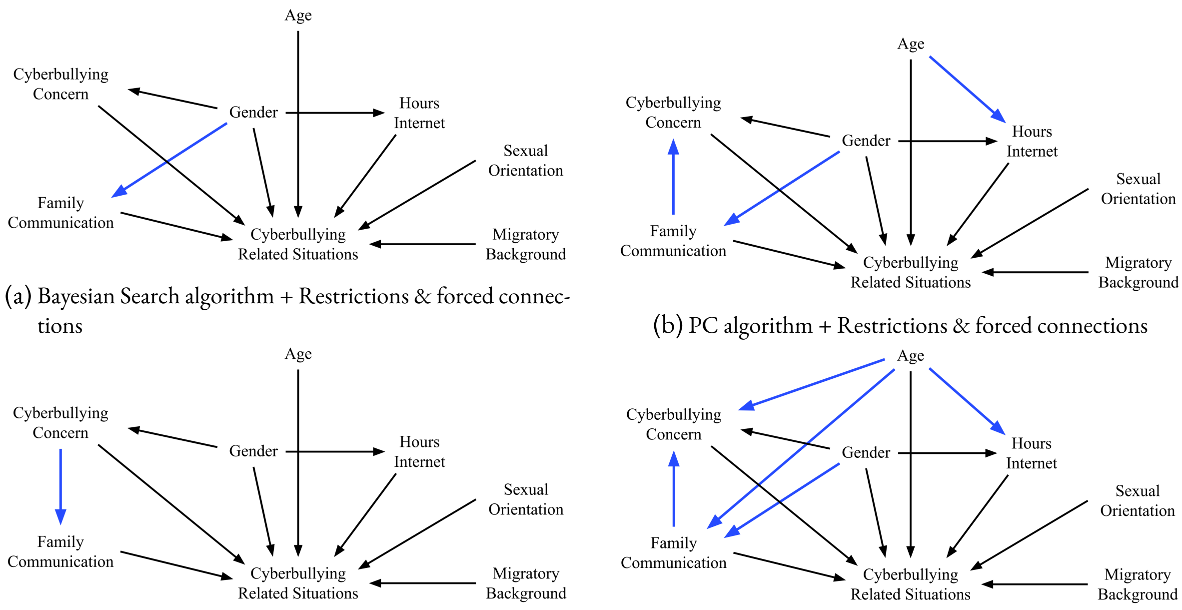


Figure 3.3: *Step 1*: Initial proposed causal DAGs. Structure (a) is a baseline where all variables are connected to the target variable. Structure (b) is the first experts' proposal. Structures (c), (d) and (e) were derived by the algorithms without using restrictions or forced connections.



(a) Bayesian Search algorithm + Restrictions & forced connections

(b) PC algorithm + Restrictions & forced connections

(c) GTT algorithm + Restrictions & forced connections

(d) Second experts' proposal

Figure 3.4: *Step 2* and *3*: Refined causal DAGs obtained through the methodology described in this chapter. Structures (a), (b) and (c) were derived by the algorithms using the restrictions and forced connections from expert knowledge (Table 3.1). Structure (d) is the second experts proposal. For ease of visualization, the connections that differ between the proposals have been highlighted in blue.

Table 3.1: *Step 2 and 3*: List of restrictions and forced connections imposed by the experts in the training of the structure learning algorithms.

Variable	**Forbidden connection**	**Forced connection**
Gender	[Age, sexual orientation, and migratory background]	[CB related situations]
Age	[Gender, sexual orientation, and migratory background]	[CB related situations]
Sexual Orientation	[Age, gender, and migratory background]	[CB related situations]
Migratory Background	[Age, gender, and sexual orientation]	[CB related situations]
Family Communication	[Age, gender, sexual orientation, and migratory background]	[CB related situations]
Daily Hours of Internet	[Age, gender, sexual orientation, and migratory background]	[CB related situations]
CB Awareness	[Age, gender, sexual orientation, and migratory background]	[CB related situations]
CB related situations	[Age, gender, sexual orientation, migratory background, family communication, hours of Internet, and CB awareness]	-

Table 3.2: *Step 4*: Quantitative comparison of causal DAGs. Cells highlighted in green indicate the best result in each of the metrics.

Metrics	Naive	1st Experts Proposal	Bayesian Search	PC	GTT	2nd Experts Proposal
Mean Log likelihood [test sets]	-1090.94	-1074.15	-1074.62	-1066.75	-1068.52	-1064.03
BIC	-20172.34	-20827.21	-20159.05	-20327.32	-20242.35	-20996.33
K2 score	-4856.9	-4885.6	-4810.8	-4824.04	-4817.65	-4891.73
Correlation score	0.64	0.62	0.64	0.67	0.64	0.57

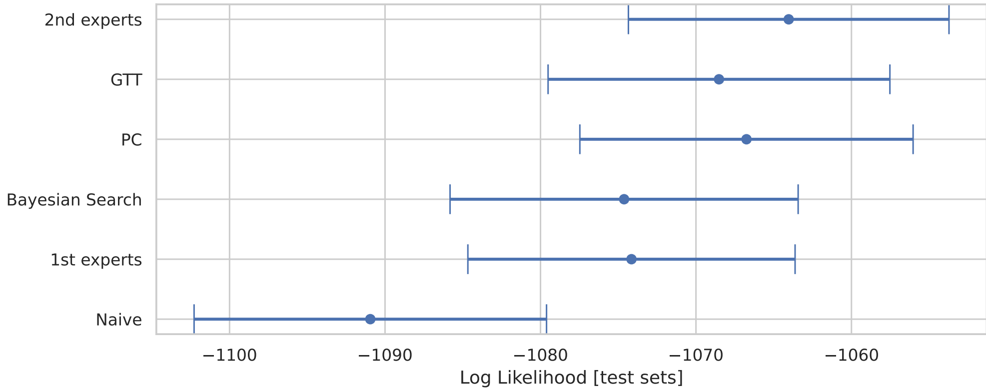


Figure 3.5: Results of the cross-validation ($k = 5$) evaluation using the Log Likelihood metric on the test sets. Standard deviations are included. The *second experts' proposal* is the winner, although it should be noted that the mean of the *GTT* and *PC* proposals are within its standard deviation.

Once we iterate and obtain a set of causal DAGs, the subsequent stage (*Step 4*) involves conducting a quantitative comparison using the metrics described in section 3.2.1. As discussed in that section, we used a k-fold cross-validation technique to reduce overfitting in the case of the LL score. We used $k = 5$ stratified random segments. The results shown in the first row of the Table 3.2 are the mean of the LL in the test segments. Figure 3.5 shows the results including the standard deviations of the LL score in the test sets.

Table 3.2 presents the findings from this quantitative comparison phase. It is important to note that all the structures compared quantitatively in the last phase are partially based on expert knowledge, as restrictions and forced connections are included in their training. Based on the results obtained, the top three structures are Bayesian Search (Fig. 3.4a), PC (Fig. 3.4b), and the second experts' proposal (Fig. 3.4d). Thus, these structures will be used for the following causal analysis.

3.3.3 Average Causal Effect Estimation

Having determined the winning structures, we computed the ACE explained in Sec. 3.2.2. This analysis allow us to estimate the causal influence that each network node has on the target, which in this case is suffering *Cyberbullying related situations*. This provides insights into the most relevant variables influencing the model output, considering them individually.

As a baseline for the experiments to be conducted, we have performed Chi-square statistical tests between all the variables in the survey dataset. Table 3.3 shows the results of the Chi-square test with a p-value of less than 0.05 between the profiling variables.

As outlined in the methodology, we express the ACE as percentage difference (ΔP) and the equivalent OR (Eq. 3.5). The results for the Bayesian search structure are shown in Table 3.4, those for the PC structure in Table 3.5, and those for the second experts' proposal in Table 3.6.

Figure 3.6 shows the aggregated ΔP results of all the variables. Across all the models, the variable with the greatest causal influence is *age*, indicating that it is the variable where it would be most effective to intervene. Logically, it is physically impossible to intervene in the age of adolescents. Nevertheless, these interventions can include, for example, awareness campaigns focused on certain ages or specific sectors of the population. This information and the resulting models can guide future interventions, policies, or educational programs. The remaining variables have a highly similar causal influence on the output of the models, except for *migratory background*, which has the most negligible influence in all the proposals.

The aggregate results in Figure 3.6 show a consensus among the different models. Therefore, any of these models could serve as a consensus graph depending on the metric used and the subsequent field of application

Table 3.3: Results of the Chi-square test between the profile variables of the dataset collected through a survey of Spanish minors within the RAYUELA project. Only the results of the tests with a p-value of less than 0.05 are shown.

	Age	**Gender**	**Sexual Orientation**	**Hours Internet**	**CB Awareness**	**Family Communication**	**Migratory Background**	**CB Related Situations**
Age	-							
Gender	-	-						
Sexual Orientation	18.52	68.47	-					
Hours Internet	24.4	27.47	-	-				
CB Awareness	-	80.8	-	-	-			
Family Communication	-	36.43	8.94	-	80.4	-		
Migratory Background	10.72	-	-	-	-	-	-	
CB Related Situations	-	7.12	7.24	12.43	-	-	-	-

of the model. The Bayesian search structure is the simplest model that explains the data (Occam's razor approximation). The PC model is the one that best meets the assumptions of conditional independence. Besides, the second experts' proposal is the best fit for the data, at the cost of having a more complex model.

Comparing these results with the baseline Chi-square test table (Tables 3.3), we observe that there are notable differences in which variables influence the outcome. For example, Age is the most important variable using the causal methodology, but it did not pass the correlation statistic test.

Table 3.4: Average Causal Effect estimation on the **Bayesian Search** structure (Fig. 3.4a). Larger values indicate greater causal influence in the outcome. ΔP represents the maximum difference in probabilities (estimated by the PGCM) of having suffered CB-related situations, by simulating interventions on all possible values of each variable. We also include the odds ratio related to ΔP (Eq. 3.5).

Variable	ΔP	Odds Ratio
Age	16%	1.97
Gender	11%	1.54
CB Awareness	10%	1.53
Daily Hours Internet	10%	1.5
Family Communication	10%	1.5
Sexual Orientation	10%	1.49
Migratory Background	7%	1.35

Table 3.5: Average Causal Effect estimation on the **PC** structure (Fig. 3.4b). Larger values indicate greater causal influence in the outcome. ΔP represents the maximum difference in probabilities (estimated by the PGCM) of having suffered CB-related situations, by simulating interventions on all possible values of each variable. We also include the odds ratio related to ΔP (Eq. 3.5).

Variable	ΔP	Odds Ratio
Age	16%	1.96
Gender	11%	1.6
CB Awareness	11%	1.58
Sexual Orientation	11%	1.56
Family Communication	10%	1.51
Daily Hours Internet	10%	1.5
Migratory Background	8%	1.38

Table 3.6: Average Causal Effect estimation on the **second experts' proposal** structure (Fig. 3.4d). Larger values indicate greater causal influence in the outcome. ΔP represents the maximum difference in probabilities (estimated by the PGCM) of having suffered CB-related situations, by simulating interventions on all possible values of each variable. We also include the odds ratio related to ΔP (Eq. 3.5).

Variable	ΔP	Odds Ratio
Age	16%	1.96
Gender	12%	1.66
CB Awareness	11%	1.57
Sexual Orientation	11%	1.56
Daily Hours Internet	11%	1.56
Family Communication	10%	1.5
Migratory Background	8%	1.42

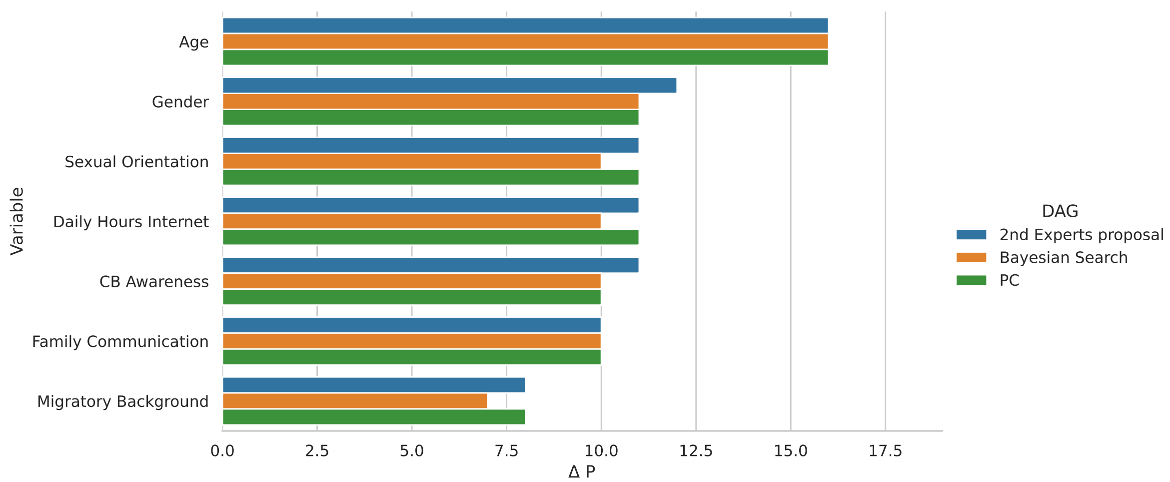


Figure 3.6: Compilation of Average Causal Effect estimations. ΔP represents the maximum difference of probabilities (estimated by the PGCM) of having suffered CB-related situations, by simulating interventions on all possible values of each variable and in each of the selected DAGs.

3.4 Discussion

To demonstrate the validity of our proposal, we discuss a case study investigating the complex dynamics of CB. It is acknowledged that while machine learning models excel in prediction, their interpretability is often compromised, leading to potential reliance on spurious correlations. Traditional correlation techniques may overlook confounding variables and colliders, further contributing to the risk of identifying misleading associations. In contrast, causality-based approaches, such as PGCM, prioritize understanding the underlying mechanisms of the phenomenon, fostering discussion and open science, making them particularly suitable for sensitive issues like CB.

While our work has yielded positive outcomes, it is essential to acknowledge the existence of unmeasured factors that influence CB, thereby limiting the scope of potential results. The complex nature of this social phenomenon implies that additional variables may contribute to its dynamics, warranting future research endeavors to capture a more comprehensive understanding of the problem. Moreover, the accuracy of our findings is closely related to the quality and variety of the data used.

Nevertheless, the use of PGCM offers notable advantages to our research. These include the ability to easily integrate expert knowledge and data, to simulate interventions, and to mitigate spurious correlations in the data. Additionally, this approach requires stating explicit hypotheses and assumptions, which encourages critical discussion and improves the transparency and accessibility of our findings.

It is also essential to acknowledge the inherent limitations of PGCM. The accuracy of our results is contingent upon the restrictions and forced connections imposed by experts, which may be biased by their knowledge or current paradigms. Furthermore, as the number of variables in the PGCM increases, the complexity and computational demands grow exponentially, potentially rendering the model challenging to interpret.

3.5 Conclusions

This chapter contributes to the study of obtaining robust causal models by mixing expert knowledge and data-driven algorithms. To demonstrate the validity of our proposal, we have analyzed a case study on CB using data from a representative survey of children in Spanish (Madrid) schools.

When using PGCM in CB, we take a causal analytical approach to better understand complex dynamics. We combine expert knowledge with data-driven algorithms to reach a consensus that overcomes each approach's inherent limitations. On the one hand, experts learn new causal connections proposed by the algorithms. On the other hand, the algorithms have to adjust their training to the constraints and forced connections imposed by the experts or by common sense.

Our findings in the case study highlight the significance of the variable *Age* in influencing CB victimization among those variables considered in the model. However, the rest of the variables studied also have a significant causal influence. From an intervention and policy perspective, our results suggest that efforts should focus on prevention strategies during critical ages. Moreover, in promoting acceptance of diverse sexual orientations and gender identities among children and implementing awareness campaigns.

However, this model includes a relatively limited number of variables, so it is likely that *Age* acts as a confounder for variables not captured in the survey or model. Thus, the conclusions drawn from the results may be limited. Despite these limitations, adopting PGCM and incorporating expert knowledge is extremely valuable in helping to understand complex phenomena, especially in situations where the sample size is limited or where the interpretation of the results may influence policies or protection strategies. This is often the case in the social sciences.

4 Generation of Probabilistic Synthetic Data for Serious Games

This chapter outlines a methodology for generating probabilistic synthetic data for Serious Games (SG). First, we discuss the motivation behind producing synthetic data in this field. Second, we analyze the state of the art in synthetic data to identify precisely where this work contributes. Third, we describe the proposed simulator and the design considerations addressed in detail. Fourth, a case study is carried out based on the SG of the H2020 RAYUELA project and survey data collected concerning cyberbullying (CB) in Spanish schools of minors. Finally, we conducted an identifiability and robustness analysis of the synthetic data generated and present the conclusions and limitations of the research. The content of this chapter has already been published in a peer-reviewed scientific journal [188]. The source code of the developed simulator is freely available on GitHub[a].

4.1 Motivation

Serious Games (SG) are tools designed for purposes beyond pure entertainment (e.g., educational, training, awareness, marketing) [3] and they have gained prominence in recent years in research, industry, and education [44, 271, 138], offering immersive and interactive experiences to users. 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 [114]. Although synthetic data is not a replacement for real data, it can accelerate the SG development process and facilitate advanced data modeling and analysis [189]. In recent years, interest in using synthetic data in social or behavioral science research has also notably increased [87, 195, 105].

We can highlight two essential benefits of data synthesis. First, providing more efficient access to data is essential for developing more reliable data-driven models. Accessing data to build and test AI models is a critical challenge for their adoption more broadly [16, 29]. A technical report from *Deloitte* [234] concludes that data-access issues are ranked in the top three companies' top challenges when implementing AI. Data synthesis can provide realistic data to work with, efficiently, and at a scale. Second, enabling better analytics when access to real data is too costly, dangerous, or unethical. Another valuable scenario is when actual data are not accessible yet. Hence, synthetic data are used to train an initial model, significantly accelerating the project convergence and potentially increasing the final models' robustness.

This chapter aims to present a methodology for generating probabilistic synthetic data for decision-based SGs, such as the one used in the RAYUELA[b] project where this thesis is framed. However, the methodology presented can be used to generate synthetic data in any decision-making scenario (e.g., multiple choice exams, political surveys, any questionnaire, or decision-based 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 Causal Models (PGCM)

[a] https://github.com/jaimeperezsanchez/Simulator_Synthetic-Data

[b] https://www.rayuela-h2020.eu/

[179]. PGCMs and their generalized form, Bayesian Networks (BN), have become a popular tool in recent years [162], including some examples in CB research [228, 148]. 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 [191, 38, 250].

4.2 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. The demand for data to build increasingly powerful AI and Deep Learning models has grown exponentially in recent years. Moreover, this data may be difficult, expensive, or unethical in many domains. 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.

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 [60].

4.2.1 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 starting point is a real dataset, and synthetic samples are obtained after modeling its structure and statistical distributions. The synthetic data should appear realistic in the separate distribution of each variable and the relationships between variables. 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 [114].

Classical statistical imputation methods (e.g., Synthetic Minority Over-sampling Technique - SMOTE, Adaptive Synthetic Sampling - 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. These include linear or geometric transformations to the data. For example, time series data may include noise addition techniques, rotations, scaling, warping, or permutations [248]; image data may include kernel filters, random blurring, or color space transformations [220].

In recent years, sophisticated machine and deep learning techniques have begun to capture particularly complex relationships between variables, allowing the data generative processes to be generalized better. Within this category, techniques such as Variational Autoencoder (VAE) [253, 109], Generative Adversarial Networks (GAN) [14, 186], or diffusion models [264, 55] are achieving the greatest success. Recent generative AI technologies (e.g., ChatGPT, Midjourney) promise significant synthetic text and image generation advances. However, special care must be taken to ensure that models do not collapse due to self-consuming loops [6, 223] and hallucinations [208]. Numerous updated surveys address the usefulness of data augmentation techniques depending on the data type, whether time series [256, 110], images [220], or text [221].

4.2.2 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 [227]. 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 [270]. The concept of Digital Twin is applied when the aim is to computationally mimic specific facilities, operational processes, or physical products [241].

4.2.3 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 [60].

Simulations also play a crucial role in this hybrid approach by generating scenarios that may not be well represented in the existing data [247, 176]. 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 [249].

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, it has been used to generate data in specialized industrial processes [192, 263] or to improve medical systems [219, 133]. Some proposals in the literature already propose using BNs, a generalized form of PGCM without causal implications, to generate synthetic data, as they are a convenient approach to merging expert knowledge and data [142, 120, 81].

The work developed in this chapter 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 SG. Unlike other agent-based simulations, the goal is not to "win" the game but to approximate realistic human behaviors in an approach more similar to [254] and [149].

4.3 Simulator

4.3.1 Design Considerations

Before detailing the proposed simulator architecture, we will review the design considerations and constraints that led to the decisions made. Firstly, although the proposed architecture can be applied to other environments where participants must complete a series of decisions or answer categorical questions, this work focuses on the specific problem of a decision-based SG.

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 generically so that it is easy to experiment and introduce additional data at any time and so that the proposed architecture can be used to address issues. To meet this design need, we propose using PGCM [179]. This model is a powerful visual and quantitative tool for expressing causal relationships among variables. The structure and parameters of a PGCM can be learned from data, manually constructed (usually with the help of experts in the specific problem being addressed), or a combination of both. Conveniently, a trained PGCM can generate synthetic data by sampling from the learned probability distributions.

The ultimate goal of RAYUELA's SG is to identify different groups/ clusters of players through the answers collected. In other words, to investigate whether the answers given in the SG 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 chapter aims to generate probabilistic synthetic data reflecting the internal states of the players and their cognitive decision-making process. To meet this design need, we propose using the IRT framework [62], 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 [63], 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 more complex models). Our work will use IRT to generate synthetic data rather than for statistical inference. With this approach, we aim to model the interactions between players and the in-game decisions they have to confront based on widely used psychological theories. Our proposed simulator probabilistically models players' decisions using IRT test theory, incorporating expert knowledge and external data through PGCM.

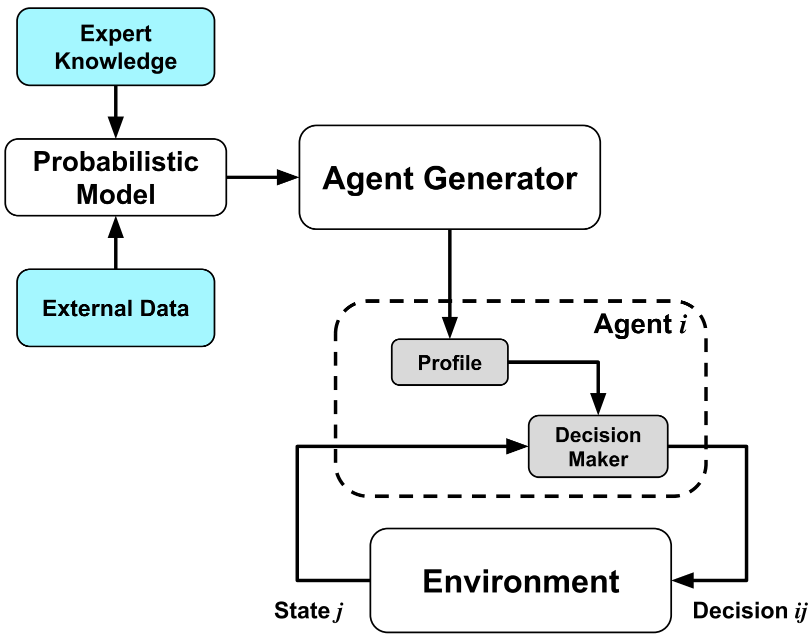


Figure 4.1: Simulator's overall architecture and components. The generated *agents* respond to the *environment* and decide according to its *profile* in a non-deterministic manner. The blue boxes represent external information that is fed into the simulation. The grey boxes represent the internal states and models of the synthetic agents.

4.3.2 Architecture

Considering the technical and design considerations outlined in the previous subsection, we summarize the simulator's proposed architecture in Figure 4.1. 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. 4.1), which we describe in the following subsections.

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 PGCMs, 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, PGCMs allow us to *interrogate* the model and perform causal tasks to obtain quantifiable responses for events for which we have little or no data. This can be useful for generating data on underrepresented events or population sectors.

We propose to use a PGCM trained with external data to generate synthetic data that the *Agent Generator* module will then use to produce synthetic players with a unique profile in a probabilistic manner. The final synthetic dataset consists of the data generated by the PGCM and the probabilistic responses each agent gives in the environment.

Agent Generator

This module generates synthetic agents with distinct internal parameters representing a variety of profiles. These internal parameters model the internal states of the player (e.g., that the player is more or less risky or tends to behave in a particular way). The data produced by the *Probabilistic Model* drives this generation process. 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. Section 4.4 will explain in detail our implementation for our case study on CB.

Agent

This module represents the profile of each synthetic player and recreates the interaction with the decisions to be taken in the simulator, obtaining as output the answers/decisions taken according to his profile (in a probabilistic way). Two main components constitute the *Agent* module:

1. *profile* (α_i): This is a fixed internal parameter, unique for each agent, representing its profile. This numerical value is inherited from the *Agent Generator* module. For instance, in our case study on CB, the profile parameter will represent the risk propensity of each agent in online situations. Positive α_i values 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.

2. *Decision maker*: This submodule will simulate the game's agents' decisions, 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 made to fit our case properly. As we explained before, the ultimate goal of RAYUELA's SG is to identify different groups/clusters of players through the answers collected. Therefore, there will not be correct or incorrect answers but answers representing greater or lesser alignment with certain profile traits.

In the simplest case, where agents will make dichotomous choices (i.e., two possible answers), it can be formally expressed as Equation (4.1). Assuming uncorrelated questions, 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. Equation (4.1) 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, such as the Graded Response Model [211].

$$\text{Answer}_{ij} \sim \text{Bernoulli}(p_{ij}), \text{ with } p_{ij} = \frac{1}{1 + e^{-\alpha_i \beta_j}} \tag{4.1}$$

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 a β close to 1 provides a good measurement of the agent's profile parameter (α_i). In actual SGs, 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).

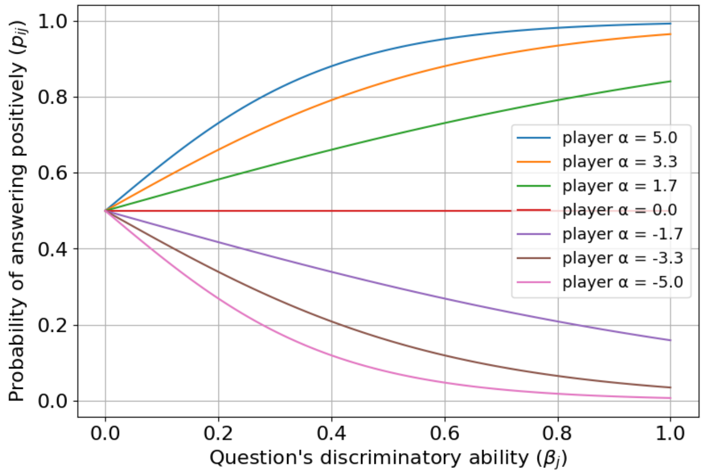


Figure 4.2: Visualization of the probability p_{ij} (Eq. 4.1) values as a function of α and β for the dichotomous/binary case. The values of the equation are shown for 7 values of α, represented in different colors. Positive values of α_i would represent agents more likely to respond positively to the questions posed to them. The opposite applies to negative values of α_i. Values of α_i around zero represent a random player.

Figure 4.2 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, irrespective of the question or β_j value.

Environment

This module simulates the SG's structure and is the component with which the synthetic agents interact, providing the β_j values modeling to the questions/decisions. In decision-based SG, 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. 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 or uniform distribution).

4.3.3 Generation Process

The proposed method to generate informed synthetic data using a PGCM can be summarized in the following steps. Figure 4.3 shows a graphical representation summarizing the proposed methodology.

1. Build the PGCM structure (i.e., DAG) from expert knowledge, a structure learning algorithm, or a hybrid approach.
2. Train the PGCM with external data to learn the parameters (i.e., CPTs) using a learning algorithm such as Maximum Likelihood Estimator or Expected Maximization [51].
3. Sample synthetic data from the PGCM using a sampling algorithm such as Bayesian Model Sampling or Gibbs sampling [80], yielding the characteristics that define each agent's profile.
4. Check the value of the variable of interest (in our case study, having experienced a CB-related situation) to determine the profile category of each synthetic player/agent.
5. Sample each agent's profile value (α_i) 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. 4.1).

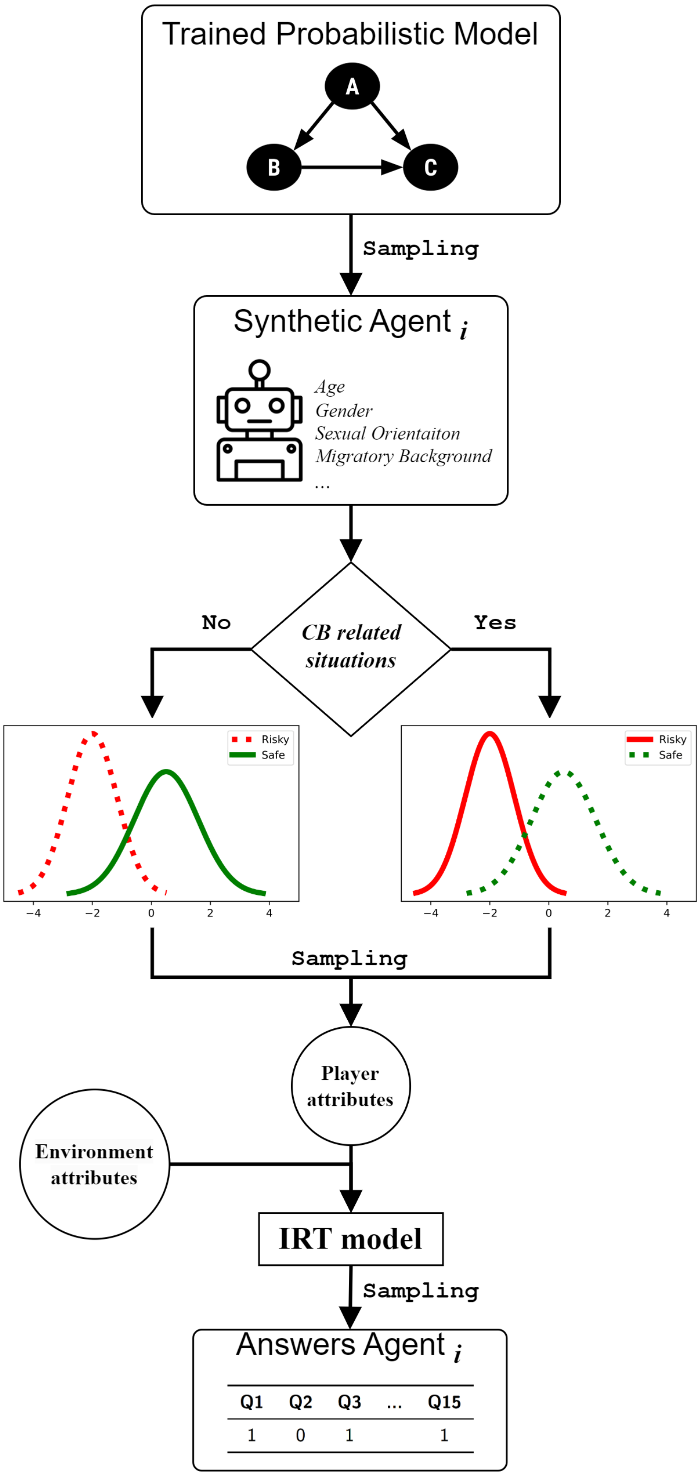


Figure 4.3: Conceptual graphical representation of the steps involved in the synthetic data generation. The elements of the generator architecture interact to produce a synthetic dataset representing decisions (0 or 1 in this case) representative of the modeled problem.

4.4 Case Study: Generating Synthetic Data for a Serious Game on Cyberbullying

This section will explain how we applied the proposed simulator architecture in the H2020 RAYUELA[c] project. The SG developed in the project is focused on CB, aiming to identify different groups/clusters of players through their responses to the situations they are presented with. Specifically, to differentiate between risky and safe players regarding their online behavior. The project was undertaken by an interdisciplinary team, including psychologists and anthropologists with expertise in CB and cybercrime. Following the proposed architecture (Fig. 4.1) and to generate synthetic data more faithful to reality, we will use a PGCM as the probabilistic model to introduce expert knowledge and external data into the simulation.

The external data consists of a survey of minors in Spanish schools (Madrid) in 2022. These are the same data used in the case study in Chapter 3. Section 1.2.1 details the characteristics of the dataset and its collection. This survey collected demographic data, questions about the participants' relationship with new technologies and the Internet, and inquiries about situations related to CB or cyber-harassment. Table 1.1 provides the variable values and their percentage of occurrences (i.e., marginal probability), and Table 4.1 shows a random sample of 5 survey participants. An ethics committee approved the procedure, and before conducting the survey, the researchers or professors explained the project and the data collection strategy to the participants.

[*Step 1*] The PGCM structure (i.e., causal DAG) employed in this case study (Fig. 4.4) results from the research presented in Chapter 3, where a consensus has been reached between expert knowledge and data-driven algorithms. 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.

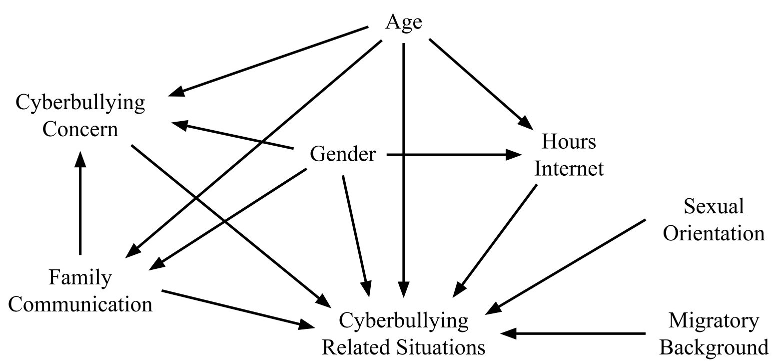


Figure 4.4: **Probabilistic Model**: PGCM structure which encodes the causal relationships among the variables collected in the survey to minors. This structure results from the research presented in Chapter 3, where a consensus has been reached between expert knowledge and data-driven algorithms.

[*Step 2*] The PGCM is trained using the Expected Maximization algorithm [139], a de-facto standard due to its ability to deal with missing data, being this a pervasive problem in SG or social science research. GeNIe Modeler[d] software was used to construct and train the PGCM. Uniform prior probability distributions were set in all the nodes to maintain a neutral stance and minimize possible biases.

[*Step 3*] Once the probabilistic model has been trained, we generate synthetic data using the Bayesian Model Sampling algorithm to obtain each synthetic agent characteristic. By checking the result of the variable of interest (i.e., *having experienced CB-related situations*), we know the category to which each synthetic agent belongs, and we can subsequently obtain its profile parameters (α_i).

[c] https://www.rayuela-h2020.eu/

[d] https://www.bayesfusion.com/genie/

Table 4.1: **External data**: Sample of five 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 3h and 4h	5/5 (very concerned)	2/4 (rarely)	No
Male	14	Heterosexual	No	Between 1h and 2h	2/5 (unconcerned)	3/4 (often)	No
Male	16	Bisexual	No	Between 3h and 4h	4/5 (concerned)	3/4 (often)	No
Female	14	Heterosexual	No	More than 4h	5/5 (very concerned)	4/4 (very often)	Yes
Female	16	Heterosexual	Yes	Between 3h and 4h	4/5 (concerned)	3/4 (often)	No

Table 4.2: **Synthetic Data**: Sample of five agents randomly selected from the dataset generated ($N = 500$ agents). Columns Q1 to Q15 indicate the questions of the simulation 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	...	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	Male	13	Heterosexual	No	Between 1h and 2h	4/5 (concerned)	3/4 (often)
1.69	1	0	...	1	0	Female	16	Heterosexual	No	More than 4h	2/5 (unconcerned)	1/4 (never)
0.42	1	0	...	1	0	Female	14	Heterosexual	No	Between 2h and 3h	4/5 (concerned)	3/4 (often)
-1.4	1	0	...	0	1	Female	14	Heterosexual	Yes	Between 2h and 3h	5/5 (very concerned)	4/4 (very often)
1.03	0	0	...	1	1	Non-binary	17	Non-heterosexual	No	Between 2h and 3h	5/5 (very concerned)	1/4 (never)

[*Step 4*] For the two possible groups of players, individuals who have experienced CB-related situations or not (i.e., risky or safe), we have assumed that their profile parameters (α_i) are samples from two different Gaussian distributions, allowing for some overlapping to account for the intrinsic uncertainty underlying human decision-making processes. For this case study, we have defined the hyperparameters for the Gaussian distributions as described in Equation (4.2). Note that the specific values we gave the hyperparameters are a somewhat realistic example. However, real players may behave differently.

$$\begin{aligned} \alpha_i|\text{safe} &\sim \text{Normal}(\mu = -2, \sigma = 0.7) \\ \alpha_i|\text{risky} &\sim \text{Normal}(\mu = 0.5, \sigma = 1.2) \end{aligned} \tag{4.2}$$

[*Step 5* and *6*] Once the PGCM has been trained with the survey data and we have defined the procedure to obtain the agents' profile parameters (α_i), 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 sampled from a uniform distribution.

[*Step 7*] Table 4.2 shows five samples of the generated synthetic dataset, where each agent is stored in a row. It includes the synthetic data obtained from the PGCM (*age*, *gender*, *sexual orientation*, etc.) and the answers to the posed "questions" (Q1 to Q15). In these answers, 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. Figure 4.5 shows a histogram of the generated agents' profile parameters (α_i). The bimodality of the figure mirrors the fact that α_i comes from two different Gaussian distributions, and the asymmetry arises because the incidence of risky profiles in the survey data is lower than that of safe profiles.

4.4.1 Identifiability Analysis

Once we have created the synthetic data, we will perform an empirical identifiability analysis to ensure it is possible to estimate back 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

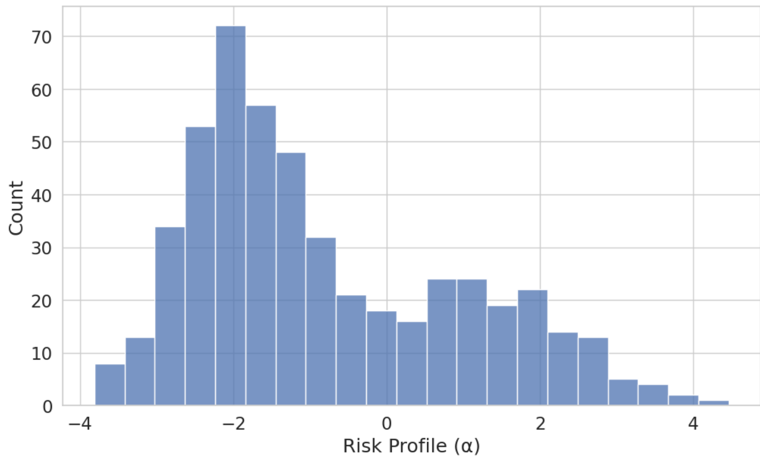


Figure 4.5: Histogram of the profiles parameters (α_i) from the synthetic generated dataset ($N = 500$ agents). Lower α values encode agents with lower risk propensity and vice versa. Values of α around 0 model the random players.

used to generate the data (described in the previous subsection) to estimate the hyperparameters defining the agents (α_i|safe and α_i|risky) and the questions' parameters (β_j). We will use only the synthetic responses ($Q_1 \ldots Q_{15}$) 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 directly through the PGCM (e.g., gender, age, sexual orientation, etc.) to reconstruct these parameters. Equation (4.3) describes the prior distributions introduced in the hierarchical Bayesian model, and Figure 4.6 shows the graphical representation of this model. The parameter p_{ij} is described in Equation (4.1).

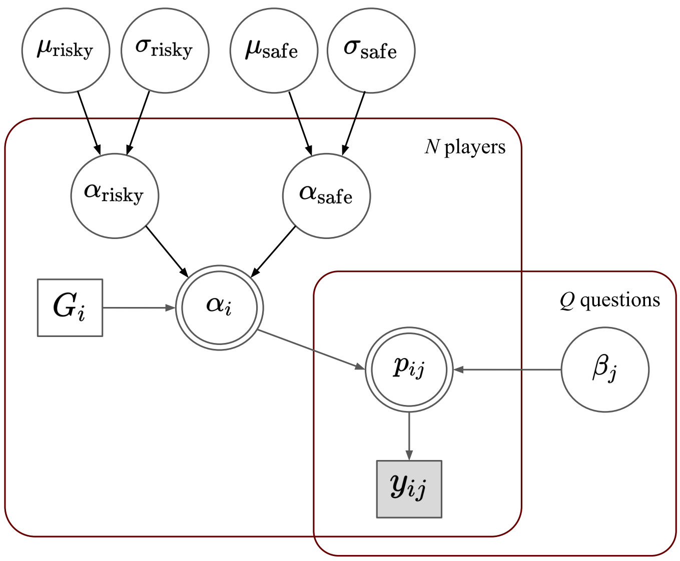


Figure 4.6: Graphical representation of the hierarchical Bayesian model. Circular nodes represent continuous random variables, and square nodes are discrete ones. Double-bordered nodes represent deterministic variables. Shaded nodes represent observed variables.

$$
\begin{aligned}
\mu_{safe} &\sim \text{Normal}(-1, 2) \\
\sigma_{safe} &\sim \text{Exponential}(1) \\
\mu_{\text{risky}} &\sim \text{Normal}(1, 2) \\
\sigma_{risky} &\sim \text{Exponential}(1) \\
G_i &\sim \text{Bernoulli}(0.5) \\
\alpha_i &\leftarrow \begin{cases} \text{Normal}(\mu_{safe}, \sigma_{safe}) \text{ if } G_i = 0 \\ \text{Normal}(\mu_{\text{risky}}, \sigma_{\text{risky}}) \text{ if } G_i = 1 \end{cases} \\
\beta_j &\sim \text{Beta}(1, 1) \\
y_{ij} &\sim \text{Bernoulli}(p_{ij})
\end{aligned}
\tag{4.3}
$$

As depicted in Figure 4.6, the profile parameters (α_i) were modeled under the paradigm of *Latent Mixture Models* [261, 239]. 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, the latent variable encodes the group membership of each player (risky or safe), assuming that players can solely originate from these two distributions. This modeling strategy enables the inference of the probability of each player belonging to the risky or safe group. The validation of this Latent Mixture Model consists of inferring the value of the distribution hyperparameters ($\mu_{safe}, \sigma_{safe}, \mu_{\text{risky}}, \sigma_{\text{risky}}$) 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.

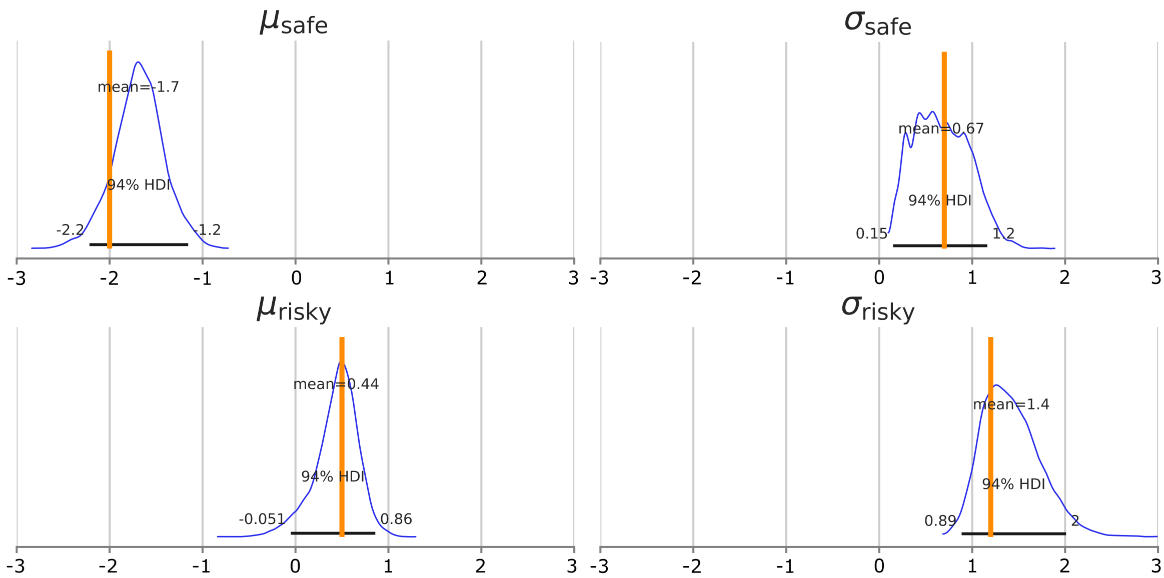


Figure 4.7: Posterior probability distributions of the hyperparameters of the Gaussian distributions generating the agents' profiles in the case study. The true values used in the generation process are shown in orange. The black line at the bottom of each plot represents the HDI (94%).

Unlike in the Machine or Deep Learning fields, we do not get a singular prediction value in Bayesian inference. Instead, 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 implemented the model with software using the open-source library *PyMC* [209], a state-of-the-art software tool for probabilistic programming and statistical computation.

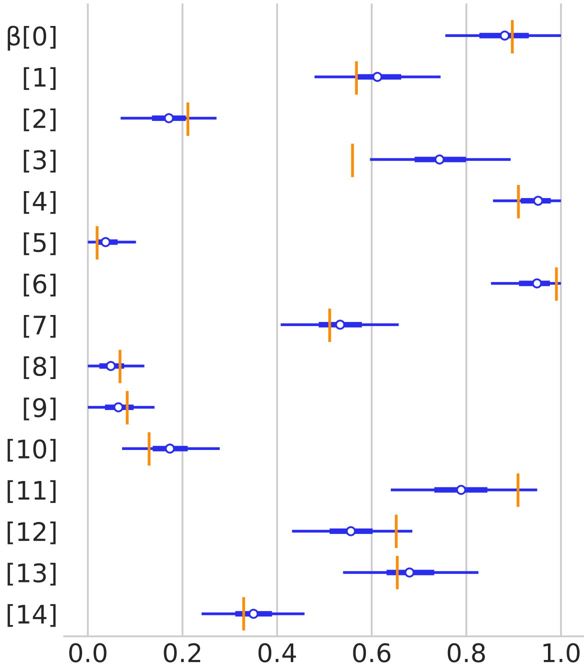


Figure 4.8: Posterior probability distributions of the questions' parameters (β_j). The blue line in each row indicates the HDI (94%) of the estimated distribution. The true values used in the generation process are shown in orange.

In Figure 4.7, we find the resulting posterior probability distributions of the hyperparameters of the Gaussian distributions that generate the agents' profiles. Figure 4.8 shows the posteriors of the β_j parameters of the questions. In both figures, the estimated probability distributions are shown in blue, and the orange lines indicate the actual value used to produce the synthetic data. We also display the High-Density Interval (HDI) of the posterior distributions, an interval containing the unobserved parameter's value with a certain probability [166]. In Bayesian inference, it is usually considered a "correct guess" if the actual parameter is within the 94% HDI of the posterior distribution [158].

As can be seen, the parameters were reconstructed quite accurately for the setting presented in this analysis, and all fell within the 94% HDI (except for the predictions on β_3). Therefore, in this sense, the synthetic data generation was successful, as the generated data encapsulates reconstructable information about the agent groups and the discriminative ability of the questions.

4.4.2 Robustness Analysis

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 same hierarchical Bayesian model shown in Figure 4.6.

The motivation for this analysis is that, as we have seen in the previous subsection, the parameters used to generate the synthetic data are reconstructible (i.e., the true values fall in the 94% HDI of the posteriors). However, if these posteriors are too broad (i.e., low confidence in the prediction), they will not be helpful, even if the actual 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.

We will use the entropy of the distributions to quantify the "width" of the posterior distributions with a single metric. This metric measures a random variable's average amount of information or uncertainty [216]. Given a discrete random variable X, which takes values in the range of $\mathcal{X}$ and is distributed according to the probabilities p, Equation (4.4) defines its entropy.

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) = \mathbb{E}[-\log_2 p(X)] \tag{4.4}$$

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. 4.9 and 4.10.

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. In Figure 4.11, we show two examples of posterior distributions of the α parameters to facilitate further interpreting the results obtained in the heatmaps, indicating the corresponding normalized entropy value.

The results of this robustness analysis indicate how many players or questions we will need, depending on the precision with which we want to estimate the latent parameters. Suppose that the proposed model reflects the behavior of real players sufficiently well. In that case, this will help us to design the SG 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.

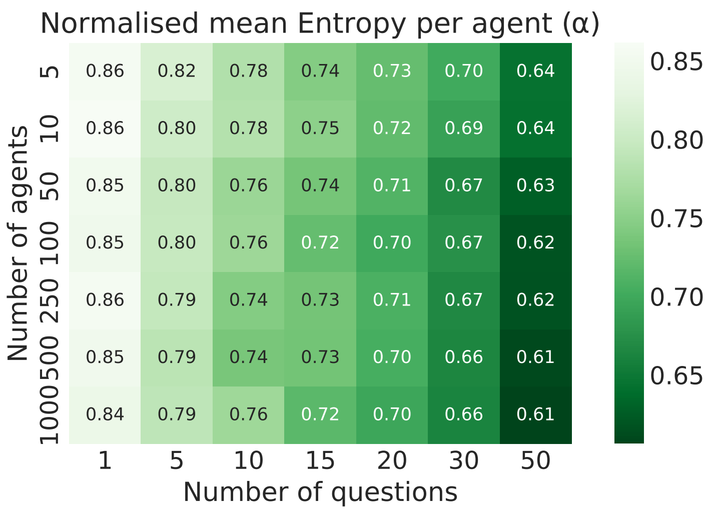


Figure 4.9: Robustness experiments using the hierarchical Bayesian model on the agents' parameters (α_i), varying the number of agents and questions. The results show the normalized entropy of the posterior distribution of the inferred parameters, being 1 maximum entropy (i.e., no valuable information about the parameters) and 0 minimum entropy (i.e., complete information about the true value of the parameters).

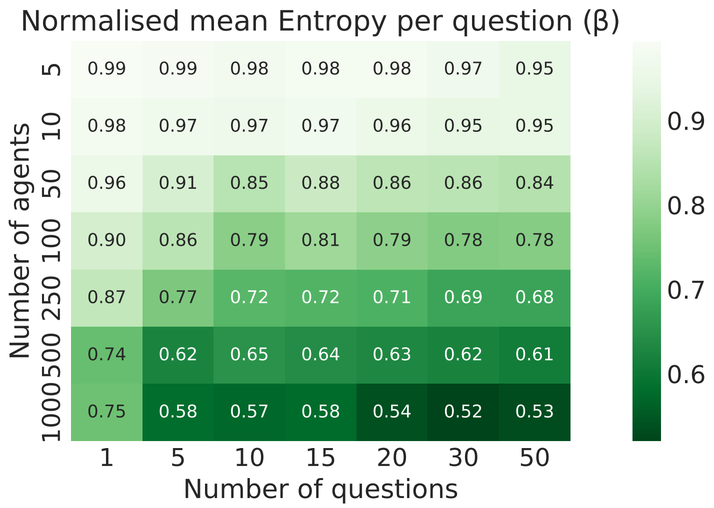


Figure 4.10: Robustness experiments using the hierarchical Bayesian model on the questions' parameters (β_j), varying the number of agents and questions. The results show the normalized entropy of the posterior distribution of the inferred parameters, being 1 maximum entropy (i.e., no valuable information about the parameters) and 0 minimum entropy (i.e., complete information about the true value of the parameters).

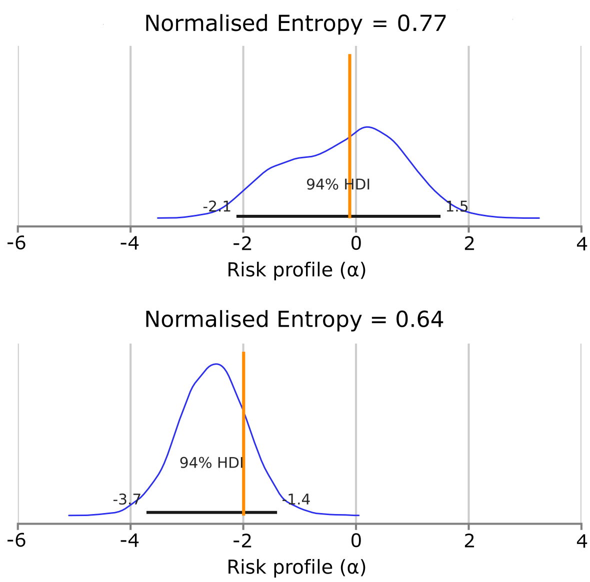


Figure 4.11: Examples of posterior probability distributions of the α parameters. The corresponding normalized entropy value is also indicated to facilitate interpreting the results obtained in the robustness experiments.

4.5 Conclusions and Limitations

In this chapter, we have introduced a novel approach for generating probabilistic synthetic data explicitly tailored for decision-based SG. 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 CB within the H2020 RAYUELA project.

The heart of our contribution lies in designing and implementing a simulator architecture conceived for generating synthetic probabilistic data (Fig. 4.1). This architecture provides flexibility for recreating decision-making scenarios within SG, capturing the complexity and uncertainty inherent in real-world situations. Our contribution fits into the hybrid synthetic data category since we employ PGCM to include external data and expert knowledge in the simulation. Furthermore, the model mimicking player behavior is based on the IRT, a cognitive modeling framework for test scoring. Thus contributing to increased realism in the generative process and, consequently, in the data produced.

Synthetically generated data has proven to be a strategic advantage in SG development. In the RAYUELA project, this methodology has allowed us to refine the number of questions necessary for the SG to achieve the desired results through robustness analysis (Figs. 4.9 and 4.10); to define the desired data structure, and to prepare the pipeline software ahead of actual data collection. Consequently, this work has accelerated project development and facilitated design, analysis, and data management.

Our work offers a valuable tool for game developers and researchers, holding immense potential to advance the development of SG and decision-based simulations and analysis. Combined with the consensus-based approach to DAG generation described in the preceding chapter, our methodology paves the ground for other studies in this and related fields.

As future lines of research, we suggest investigating more complex profile models consisting of more than one variable. Also, exploring the potential impact of introducing memory to the agents could lead to a more realistic model that allows past responses to influence future responses. Finally, we encourage researchers to use our methodology to generate synthetic data for other decision-making problems.

5 Causal Analysis of Serious Game Data

This chapter introduces a methodology for analyzing data collected with Serious Games (SG) using Probabilistic Graphical Causal Models (PGCM). First, we analyze the motivation underlying a causality-based analysis. Second, we describe in detail the proposed analytical methodology. Third, a case study is carried out based on the SG of the H2020 RAYUELA project and experimental data collected in European schools concerning cyberbullying (CB). Finally, the results are discussed, and the conclusions reached are presented.

5.1 Motivation

In recent years, SGs have emerged as a promising tool for conducting Social Science research [189]. Games are engaging by design what facilitates the exploration of complex social issues and presents a unique opportunity to access and engage with a broader and more diverse demographic of participants. SGs serve as a non-invasive research tool, offering participants a comfortable and interactive environment, which is critical when addressing sensitive and intricate social issues such as CB. This game-based research approach holds significant potential for enhancing the breadth and depth of social and behavioral research [153].

Using games as a research tool to investigate human behavior is not new but has recently increased in popularity due to the advances in digital technologies and the Internet. For example, an experiment embedded in a video game showed that the complexity of the city where a child lives influences their future navigation skills [48]. A video grammar game called "*Which English?*" probed the existence of a "critical period" for learning a second language that extends into adolescence [93]. *"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 [19].

However, the effectiveness of these games in contributing to meaningful social research hinges on the sample size and the robustness and reliability of the data analysis tools employed. Traditional methods, predominantly based on correlation techniques or black-box predictive models, are insufficient when researching sensitive issues [179] such as CB, where it is imperative to understand the phenomena' inner workings as they are aimed to maximizing some loss function, not necessarily to test causal relationships between variables.

In this chapter, we adapt existing ideas on PGCMs to analyzing SG-collected data. This approach provides a sophisticated yet explainable and intuitive framework for modeling complex causal relationships. These models allow researchers to incorporate expert knowledge (defining the network structure) or well-known quantitative knowledge (defining prior probability distributions). They also inherently include a causal perspective and quantification of uncertainty.

As mentioned in previous chapters, PGCMs have been successfully applied in diverse fields such as Biology [5], Psychology [203], Social Sciences [61], Econometrics [106], or Epidemiology [84]. The adoption of PGCMs in analyzing data derived from SGs promises a more nuanced and accurate understanding of the causal relationships at play. Furthermore, it paves the way for developing more effective policies and interventions. The presented research posits that the synergy between SG as a data collection tool and PGCMs as an analysis framework can drive forward the field of computational social science, leading to insights and solutions that are both profoundly informed and widely applicable.

5.2 Methodology

This section describes the method proposed to analyze SG data through a causal-based analysis. PGCMs enable researchers to answer various causal queries based on the research question posed by the issue. In the case of SGs employed as a social or behavioral research tool, we identify two main goals: (i) effect estimation analysis, used to estimate which interventions are the most influential on the event under study, and (ii) multi-factor profiling analysis, which investigates the combinations of factors that most influence the probability of observing the event under study.

5.2.1 Average Causal Effect Estimation

This methodology of causal analysis has already been described in section 3.2.2, since we have utilized it in chapter 3 to validate and extract information from the PGCMs under consideration. We performed an effect estimation analysis computing the Average Causal Effect (ACE), attempting to answer the following question: *How much does a certain target quantity differ under two different interventions?* Using this technique, we can quantify the causal influence of each variable on the outcome by simulating interventions using Do-calculus (Eq. 3.3) [181]. We expressed the ACE results as the percentage difference between treatment A and B ($\Delta P = P_A - P_B$), and equivalently as an odds ratio (OR) following Equation (3.5). Larger positive ACE values indicate a greater causal influence on the outcome.

5.2.2 Multi-Factor Profiling Analysis

This analysis seeks to answer the following question: *What combinations of factors make up a risk profile?* For this purpose, we examine how multiple simultaneous observations can impact the outcome. We do not simulate interventions but instead focus on observations to study profiles that significantly change the posterior probability of the outcome. Therefore, this analysis studies the difference between the prior and posterior probability after simulating an observation of specific profiling characteristics. Causality is considered when simulating the intervention through the DAG by selecting the appropriate adjustment variables.

We can study the relationship between the odds before and after the observations through the Bayes Factor (BF). It can be formally expressed as in Equation (5.1), where $P(Y)$ is the outcome probability, and E represents the observed evidence. The BF is commonly used to compare the likelihood of two models, but in this particular case, it is equivalent to the OR (Eq. 3.5).

$$\begin{aligned} \text{Prior odds } &= \frac{P(Y)}{1 - P(Y)} \\ \text{Posterior odds } &= \frac{P(Y \mid E)}{1 - P(Y \mid E)} \\ \text{Bayes Factor } (BF) = \frac{\text{Posterior odds}}{\text{Prior odds}} &= \frac{P(Y \mid E)(1 - P(Y))}{P(Y)(1 - P(Y \mid E))} \end{aligned} \tag{5.1}$$

The Jeffreys' scale [112] shown in Table 5.1 is used to decide which outcome values we consider to have changed significantly. This scale translates the order of magnitude of the BF into a qualitative judgment that allows us to decide the amount of evidence needed to support one hypothesis/model and not the other. For this research, we will consider that values of $BF > 10$ already form sufficient evidence to consider it a risk profile, which is a typically threshold to accept a hypothesis in psychological science [4].

Table 5.1: Jeffreys' scale of strength of evidence [112]. Translates the Bayes Factor value into a qualitative judgment.

Bayes Factor	Strength of evidence
10^0 to $10^{1/2}$	Barely worth mentioning
$10^{1/2}$ to 10^1	Substantial
10^1 to $10^{3/2}$	Strong
$10^{3/2}$ to 10^2	Very strong
$> 10^2$	Decisive

5.3 Case Study: Causal Analysis of Data from a Serious Game on Cyberbullying

This section briefly describes the SG developed within the RAYUELA project, defines the variables measured, and outlines the data collection procedure. Two types of variables are measured during a gaming session —those measured *in-game* and *out-of-game* through the registry or questionnaires. Finally, the experimental results of applying the proposed causal methodology are shown.

In this case study we intend to carry out an analysis of the variable *Previous CB Offending*. That is, to estimate the variables with the greatest influence on CB offenders and to find the profiling characteristics most common to these profiles.

It is worth mentioning again that, among all the cybercrimes analyzed in the RAYUELA serious game, we have decided to focus on CB since it is the only one for which a validated questionnaire [28] could be carried out on the participating minors. The results of this questionnaire represent a "ground truth" for the analyses carried out throughout this thesis.

5.3.1 Serious game description

The primary focus of the RAYUELA project is to study the drivers and human factors contributing to specific types of cybercrime affecting minors [155]. Specifically, to differentiate between risky and safe players regarding their online behavior. This goal is achieved through a unique approach that leverages gaming, providing a platform for learning and modeling behaviors in an engaging and non-invasive way. The project was undertaken by an interdisciplinary team, including psychologists and anthropologists with expertise in CB and cybercrime.

The SG was developed by the Tecnalia[a] team (part of the RAYUELA consortium). It is a point-and-click 3D graphic interactive narrative adventure where players make decisions that shape the narrative's progression and outcome. The game is set in a high school, presenting scenarios involving cybercrimes affecting young individuals. Figures 5.1 and 5.2 show screenshots of some of the decisions players have to make in the game. Figure 5.3 shows the final question related to the level of "honesty" during the gameplay. It is important to note that the situations presented in the SG are mainly from the point of view of a bystander, trying to understand how players would act in such situations and what risk they perceive. The design of the SG was subject to ethical constraints that have been discussed throughout the RAYUELA project and approved by ethics committees in each of the countries participating in the experimental pilots.

The *in-game* decisions were carefully designed and discussed by the RAYUELA team to analyze player patterns and identify those most likely to commit or experience specific cybercrimes, and were refined after testing in the earlier pilot studies. The cybercrimes considered are CB (aggressor, passive bystander, and active bystander), online grooming, fake news, and cybersecurity. The game comprises several adventures, each

[a] https://www.tecnalia.com/

addressing a cybercrime or the same cybercrime from a different perspective. Furthermore, demographic and psychological questionnaires were also administered during the gaming session to validate the researchers' hypotheses. By doing so, we could measure certain potential CB factors both *in-game* and *out-of-game* (Table 5.2), allowing us to verify the validity of the video game situation. Figures 5.1 and 5.2 are examples of *in-game* variables, where Figure 5.1 was designed to extract valuable information about cybercrimes, and Figure 5.2 was only included for the sake of playability and to improve the game flow. Appendix A shows all the game dilemmas/questions related to CB that players must answer.

The data used in this case study have been published openly [187], addressing the scarcity of shared data in this research domain. Section 1.2.1 details the characteristics of the dataset and its collection. Table 1.2 shows summarized statistics of the data collected, the possible values of each variable, and their percentage of occurrence (i.e., the marginal probability). Appendix C provides a more exhaustive exploratory data analysis.

Table 5.2: Factor/driver variables of CB identified, illustrating whether each variable was measured *in-game* or *out-of-game*.

Type	CB Factor/Driver	Measured in-game	Measured out-of-game
Environmental	Isolation/lack of social support	✓	✓
	Family communication	✓	✓
Personal	Previous CB Offending	✓	✓
	Previous CB Victimization	✓	✓
	Low self-esteem	✓	✓
	Difficulty in making friends face to face	✓	✗
	Age	✗	✓
	Gender	✗	✓
	Sexual orientation	✗	✓
	Migratory background	✗	✓
Technological	Public profile on social networks and publishing excessive information	✓	✗
	Time spent online	✓	✓
	Weak passwords or password sharing	✓	✗

As a baseline for the experiments to be conducted, we have performed Chi-square statistical tests between all the variables in the dataset. Table 5.3 shows the results of the Chi-square test with a p-value of less than 0.05 between the profiling variables. Table 5.4 shows the same results but between the game questions variables and the outcome or variable of interest (*Previous CB Offending*).

Table 5.3: Results of the Chi-square test between the profiling variables of the dataset collected in the experimental pilots of RAYUELA using a serious game. Only the results of the tests with a p-value lower than 0.05 are shown.

	Age	Gender	Sexual Orientation	Migratory Background	Self-esteem	Social support	Family support	Daily Hours of Internet	Previous CB Victimization	Previous CB Offending
Age	-									
Gender	-	-								
Sexual Orientation	30.60	61.93	-							
Migratory Background	41.20	-	-	-						
Self-esteem	-	54.65	35.24	-	-					
Social support	-	14.87	10.61	-	85.14	-				
Family support	-	30.43	22.46	12.70	171.17	272.25	-			
Daily Hours of Internet	-	-	9.57	-	16.33	38.90	23.08	-		
Previous CB Victimization	11.24	-	5.37	6.17	30.82	6.98	24.42	24.46	-	
Previous CB Offending	24.24	10.69	-	-	10.18	-	7.23	21.37	318.99	-

Table 5.4: Results of the Chi-square test between the game questions variables and the outcome or variable of interest (*Previous CB Offending*), from the data collected in the experimental pilots of RAYUELA using a serious game. Only the results of the tests with a p-value lower than 0.05 are shown.

	Previous CB Offending
Adventure 1 - Question 1 - Photo sharing	-
Adventure 1 - Question 2 - Sociable	-
Adventure 1 - Question 3 - Mathew Meme	26.73
Adventure 3 - Question 1 - Pirated Content	21.67
Adventure 3 - Question 2 - Pol or Pola	10.07
Adventure 3 - Question 3 - Time Overrun	38.94
Adventure 3 - Question 4 - Pol Bullied	32.74
Adventure 3 - Question 5 - Remind Mathew	35.7
Adventure 3 - Question 6 - Talk Pol	-
Adventure 3 - Question 7 - Help Pol	8.68

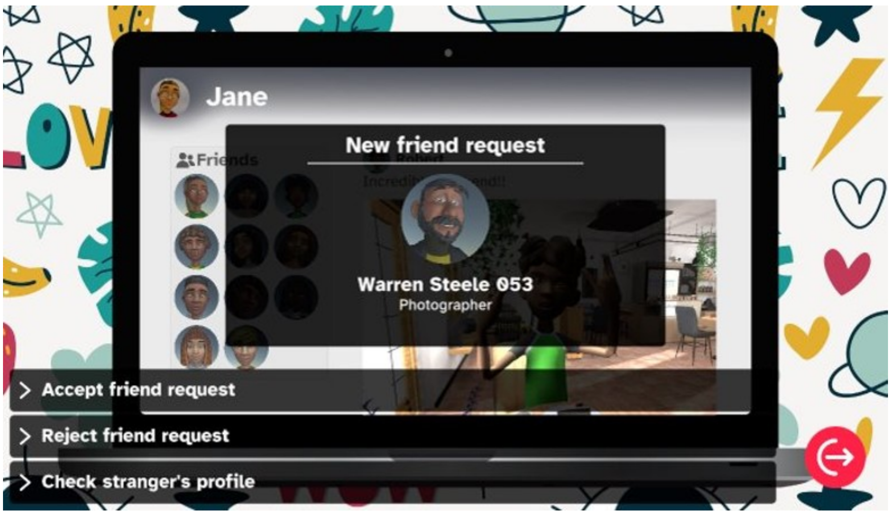


Figure 5.1: Screenshot of an informative decision players must make in RAYUELA's serious game.

Figure 5.2: Screenshot of an uninformative decision players must make in RAYUELA's serious game.

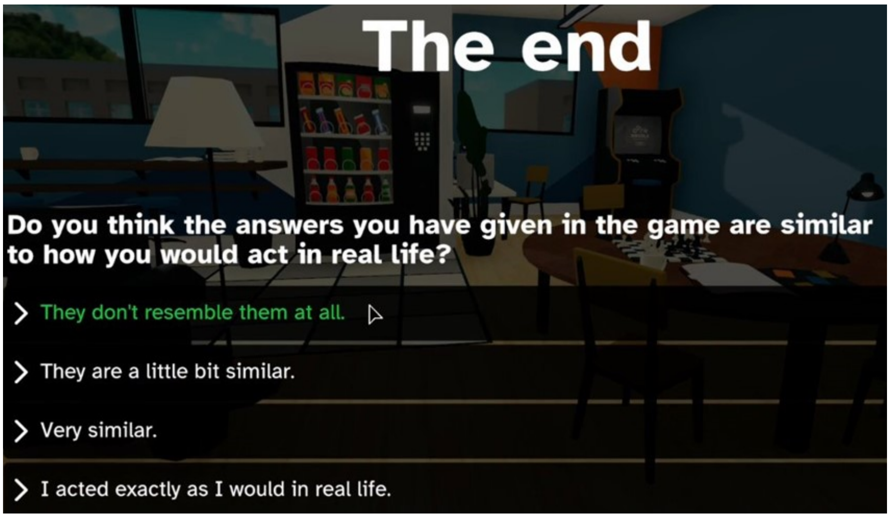


Figure 5.3: Screenshot of the "honesty" question in RAYUELA's serious game.

5.3.2 Results

This subsection shows the results of applying the proposed methodology to the described use case. The first step to performing any causal task is to reach a causal DAG that we are satisfied with and faithfully reflects the studied problem. To this end, following the methodology used in Chapter 3, we have reached a PGCM structure by joining expert knowledge with the data obtained in the experimental pilots of RAYUELA. Figure 5.4 shows the consensus DAG reached and with which we will perform the experiments. The model outcome, or variable of interest, is *Previous CB Offending*.

The prevalence of committing cyberbullying among European minors varies widely, with rates ranging from 3.0% to 30.6% [97]. With respect to cybervictimization, rates range from 5% to 37.3% [17]. We have set a 10% prior probability on both variables *Previous CB Victimization* and *Previous CB Offending* to perform the following experiments.

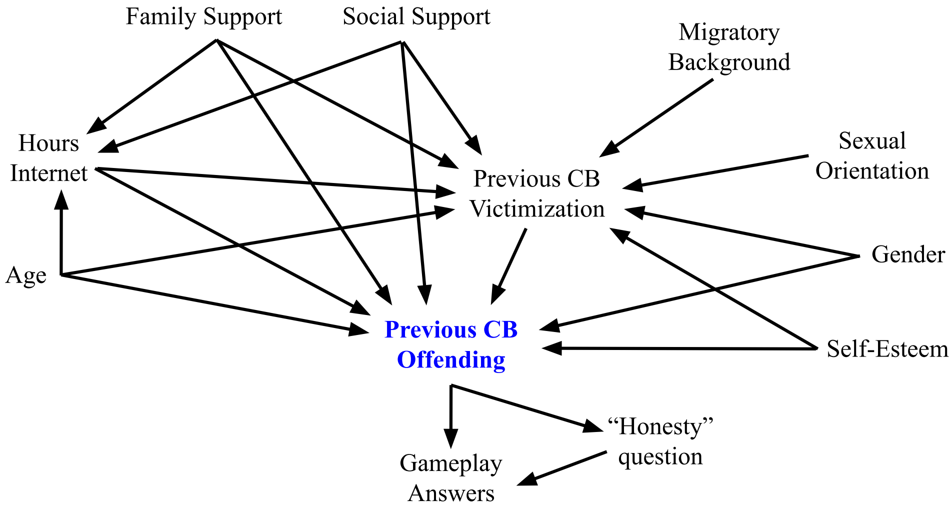


Figure 5.4: Causal DAG proposed by the CB experts from the RAYUELA project. The variable of interest (*Previous CB Offending*) is highlighted in blue.

Once we have created a DAG that faithfully represents the causal relationships of the problem, we can identify the adjustment sets, which are covariates such that stratification, adjustment, or selection (e.g., by restriction or matching) will minimize bias in estimating the causal effects of each variable on the outcome. Table 5.5 shows the total and direct adjustment sets for the DAG of this case study (Fig. 5.4), computed with the software tool *DAGitty* [242]. For example, following the first line of this table, if we wanted to analyze the direct effect of the variable *Gender* on the outcome (*Previous CB Victimization*) we should adjust the variables *Age*, *Family Support*, *Self Esteem*, *Social Support*, and *Previous CB Victimization* in the model.

Average Causal Effect Estimation

This analysis estimates the causal influence that each network node has on the target (*Previous CB Offending*). This provides insights into the most relevant variables influencing the model outcome and considering them individually. As the methodology outlines, we express the ACE as a percentage difference (ΔP) and the equivalent OR (Eq. 3.5).

Table 5.6 presents the analysis results concerning the variable of interest (*Previous CB Offending*). Figure 5.5 shows the aggregated ΔP results of all the variables. The first game question (*Adventure 1 - Question 1 - Photo Sharing*) acts as a control variable, generating random responses from players. Hence, variables with similar results are considered irrelevant for practical purposes.

The results indicate that the most relevant variables are primarily derived from the game questions, except for the variable *Previous CB Victimization*. This suggest that the SG could be a reliable tool for detecting or predicting *Previous CB offending*. As an aid to interpreting the significance of the results obtained, Figure 5.6 shows the conditional probabilities of *Previous CB Offending* estimated by the PGCM, given the possible answers to the question *Adventure 3 - Question 5 - Remind Mathew*, which is the most influential according to the results (Table 5.6). This figure illustrates that players choosing answer 3 have a 24.2% likelihood of engaging in cyber-aggression in the past, independent of other variables. Conversely, those selecting answer 2 have a mere 8.4% chance. This particular game question deals with a situation where a character is being cyberbullied. The players are asked if they have experienced a similar CB-related situation in the past or if they have ever been the perpetrators.

Table 5.5: Adjustment sets for total and direct effects in the consensus DAG (Fig. 5.4) taking the variable *Previous CB Offending* as outcome.

Variable	**Adjustment direct effect**	**Adjustment total effect**
Gender	[Age, Family Support, Self Esteem, Social Support, Previous CB Victimization]	—
Age	[Family Support, Gender, Hours Internet, Self Esteem, Previous CB Victimization, Social Support]	—
Sexual Orientation	[Age, Family Support, Gender, Self Esteem, Social Support, Previous CB Victimization]	—
Migratory Background	[Age, Family Support, Gender, Self Esteem, Social Support, Previous CB Victimization]	—
Self-esteem	[Age, Family Support, Gender, Social Support, Previous CB Victimization]	—
Social support	[Age, Family Support, Gender, Hours Internet, Previous CB Victimization, Self Esteem]	—
Family support	[Age, Gender, Hours Internet, Self Esteem, Social Support, Previous CB Victimization]	—
Daily Hours of Internet	[Age, Family Support, Social Support]	[Age, Family Support, Social Support]
Previous CB Victimization	[Age, Family Support, Gender, Self Esteem, Social Support]	[Age, Family Support, Gender, Self Esteem, Social Support]

By examining the remaining questions that most influence past CB aggression, we can extract relevant information for the elaboration of intervention strategies and new prevention policies: *Adventure 3 Question 4 (Pol Bullied)* and *Adventure 1 Question 3 (Matthew Meme)* address the importance the player places on situations in the early stages of CB; *Adventure 3 Question 3 (Time Overrun)* covers being online too long without taking breaks; and *Adventure 3 Question 1 (Pirated Content)* examines the perceived risk of downloading pirated content from the Internet. All the game questions related to CB and its possible answers can be found in Appendix A.

On the other hand, the demographic variables, when examined individually, exhibit minimal significance. Suggesting that, in order to study CB, how each subject responds to specific situations is more relevant than their personal or environmental characteristics.

Comparing these results with the baseline Chi-squared test tables (Tables 5.3 and 5.4), we observe that there are differences when we order the influence of each variable from highest to lowest. This indicates that causal modeling has helped us to calibrate the sizes of the influence effects. In addition, causal modeling allowed us to make a fair comparison between the profiling variables and the game questions, which was not possible by looking at correlation statistics alone.

Table 5.6: ACE estimation on the variable *Previous CB Offending*. Larger values indicate greater causal influence in the outcome. ΔP represents the maximum difference in probabilities (estimated by the PGCM), by simulating interventions on all possible values of each variable. Odds Ratio is computed using (Eq. 3.5). Note how, besides "Previous Victimization", the game questions contain higher causal strength. Also, *control* questions (the last two on the table) do not have any causal content, as expected by design.

Variable	ΔP	Odds Ratio
Adventure 3 - Question 5 - Remind Mathew	16%	3.47
Adventure 3 - Question 4 - Pol Bullied	14%	3.15
Adventure 3 - Question 3 - Time Overrun	10%	2.72
Adventure 1 - Question 3 - Mathew Meme	10%	2.55
Previous Victimization	9%	2.24
Adventure 3 - Question 1 - Pirated Content	9%	2.17
Adventure 3 - Question 7 - Help Pol	8%	1.95
Adventure 3 - Question 2 - Pol or Pola	6%	1.74
Honesty	5%	1.62
Adventure 3 - Question 6 - Talk Pol	4%	1.46
Age	3%	1.4
Hours of Internet	3%	1.36
Gender	3%	1.3
Self-esteem	2%	1.18
Family Support	2%	1.18
Social Support	2%	1.18
Sexual Orientation	1%	1.12
Migratory Background	1%	1.1
Adventure 1 - Question 2 - Sociable	0%	1.04
Adventure 1 - Question 1 - Photo Sharing	0%	1.01

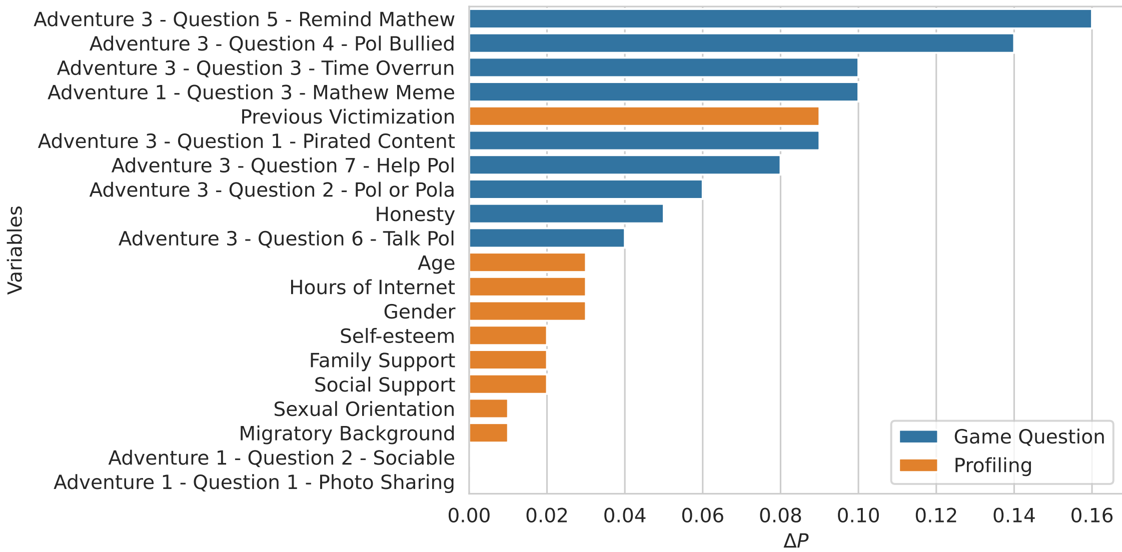


Figure 5.5: Compilation of ACE estimations. ΔP represents the maximum difference of probabilities (estimated by the PGCM) of having committed CB-related situations, by simulating interventions on all possible values of each variable. As shown, when considered individually, the game questions (blue) have a higher ACE than the variables related to the players' profiles (orange). The only exception, interestingly enough, is *Previous CB Victimisation*.

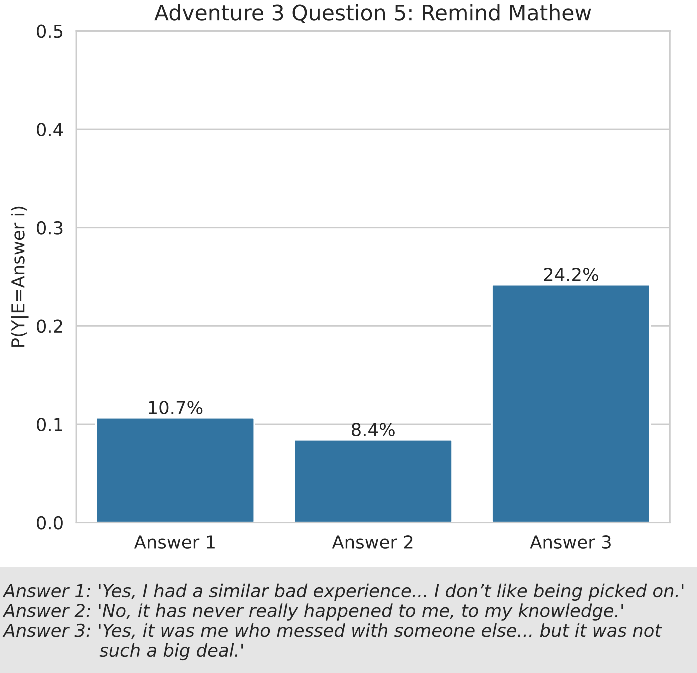


Figure 5.6: Conditional probabilities of *Previous CB Offending* estimated by the PGCM, given the possible answers to the question *Adventure 3 - Question 5 - Remind Mathew*. This question has the highest ACE in the experiments (Fig. 5.5). The graph shows that if we have no further information from the players and they choose answer 3, the estimated probability that he has committed CB in the past is 24.2%. Contrary to the other two answers, which keep this estimate around 10%, which corresponds to the prior probability.

5.3.3 Multi-Factor Profiling Analysis

This analysis seeks to find the combinations of factors that make up the risk profiles identified by the PGCM. For this purpose, we simulate multiple observations on the model variables and estimate the conditional probabilities of the variable of interest (*Previous CB Offending*).

The variables that do not come directly from the gameplay are encompassed in the term *profiling*. Those are demographic variables and those collected through psychological questionnaires (*age, gender, sexual orientation, migratory background, daily hours of Internet use, social support, family support, self-esteem, previous CB victimization*, and *previous CB offending*).

The method applied consists of inserting, by brute force, all possible combinations of observations on the variables in the PGCM. We then filter the profiles considered to be at risk according to some criteria and analyze the characteristics they have in common. We start by fixing a single piece of evidence, incrementally adding more pieces of evidence, and testing all combinations. This is done on the one hand for the game questions and on the other for profiling variables.

As explained in the methodology, the criterion to select the risky profiles is to have values of $BF > 10$ when comparing prior and posterior probabilities, and we have set a 10% prior probability on the variables *Previous CB Victimization* and *Previous CB Offending*. Therefore, a *substantial* and *strong* difference will occur at posterior probabilities of $\sim 26\%$ and $\sim 53\%$, respectively (Equation 5.2).

$$\begin{aligned}
BF &= \frac{\text{Posterior odds}}{\text{Prior odds}} = \frac{P(Y \mid E)(1-0.1)}{0.1(1-P(Y \mid E))} \\
P(Y) &= 0.1 \text{ (prior probability)} \\
BF &= 10^{1/2} \text{ (\textit{Substantial} evidence)} \Rightarrow P(Y \mid E) \approx 0.26 \\
BF &= 10 \text{ (\textit{Strong} evidence)} \Rightarrow P(Y \mid E) \approx 0.53
\end{aligned} \tag{5.2}$$

Fig. 5.7 illustrates the result of this analysis showing the maximum posterior probability of observing *Previous CB Offending* ($P(Y \mid E)$) varying the number of fixed evidences. In other words, we are depicting the probability of the riskiest profile setting one variable, then 2, then 3, and so on. These riskier profiles according to the number of fixed evidences are detailed in Table 5.7. This analysis is done both for the variables coming from game questions and profiling.

Fig. 5.7 also shows the two thresholds obtained with the Jeffreys' criterion (Equation 5.2) that delineate at what probabilities a profile starts to be risky, with substantial and strong certainty, respectively. We can observe that as the number of fixed evidences increases, the posterior probability of the riskiest profile rises progressively. However, it should also be noted that as the number of fixed evidences increases, the number of players who meet these criteria also gradually decreases. Therefore, we have less evidence about the riskiest profiles.

In contrast to the first analysis (Section 5.3.2), the findings suggest that combinations of profiling variables are equally or more informative than combinations of game questions. These results confirm that both the game questions and the profiling variables are effective tools in distinguishing players with a history of CB offenses from those without it. Therefore, SGs could be an effective alternative for measuring how players would behave in real-life situations, subject to ethical and technical constraints.

To identify common risk profile characteristics, we must choose a fixed number of evidence to detect risky profiles. Analyzing Figure 5.7, we decided to perform the profiling by fixing 7 evidences, since it is the point where the maximum posterior probability saturates. We proceed to search by brute force for all combinations of 7 variables and filter out those combinations with a posterior probability greater than 53% ($BF > 10$). Given the list of risky profiles, we perform a count of shared characteristics *prevalence*, shown in Figure 5.8. Note that these prevalences are not commensurable with the probabilities in Figure 5.5, since they show which of the players' attributes are more common among the risky profiles.

We can see (Fig 5.8) that the most common characteristic among the selected risk profiles is *Previous CB Victimisation*, observed in more than 90% of them. In other words, individuals who had been victims of CB showed a marked propensity to cyberbully in the future. This relationship has long been known in the scientific literature [225]. We can also state that *male*, *high self-esteem*, *high social support*, *high family support*, *14* and *16 years old*, and *heterosexual* are also quite prevalent characteristics within the identified risky profiles.

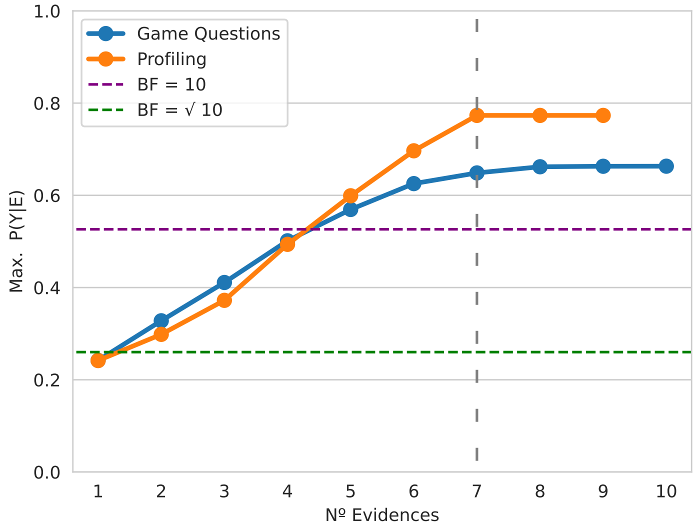


Figure 5.7: Multi-factor profiling analysis: Maximum posterior conditional probability of *Previous CB offending* varying the number of fixed pieces of evidence. This is done for the game questions and the profiling variables. The figure also shows the posterior probabilities corresponding to the relevant thresholds of the posterior probability, according to Jeffreys' criterion (Equation 5.2). It can be seen that the probability saturates in both curves from 7 fixed evidences onwards, marked by a vertical grey line.

Table 5.7: Multi-factor profiling analysis: The most risky profiles as a function of the number of fixed evidences. These profiles are shown in Figure 5.7 in the "Profiling" curve.

Nº Evidences	**Riskiest Profile**	**Estimated CB Offending Risk**
1	[Previous CB Victimization]	24.2%
2	[Previous CB Victimization + Male]	29.8%
3	[Previous CB Victimization + Male + 16 years old]	37.3%
4	[Previous CB Victimization + Male + 16 years old + High Self-Esteem]	49.4%
5	[Previous CB Victimization + Male + 16 years old + High Self-Esteem + 4-5h Internet]	59.9%
6	[Previous CB Victimization + Male + 14 years old + High Self-Esteem + 3-4h Internet + High Family Support]	69.7%
7	[Previous CB Victimization + Male + 14 years old + High Self-Esteem + 3-4h Internet + High Family Support + High Social Support]	77.3%

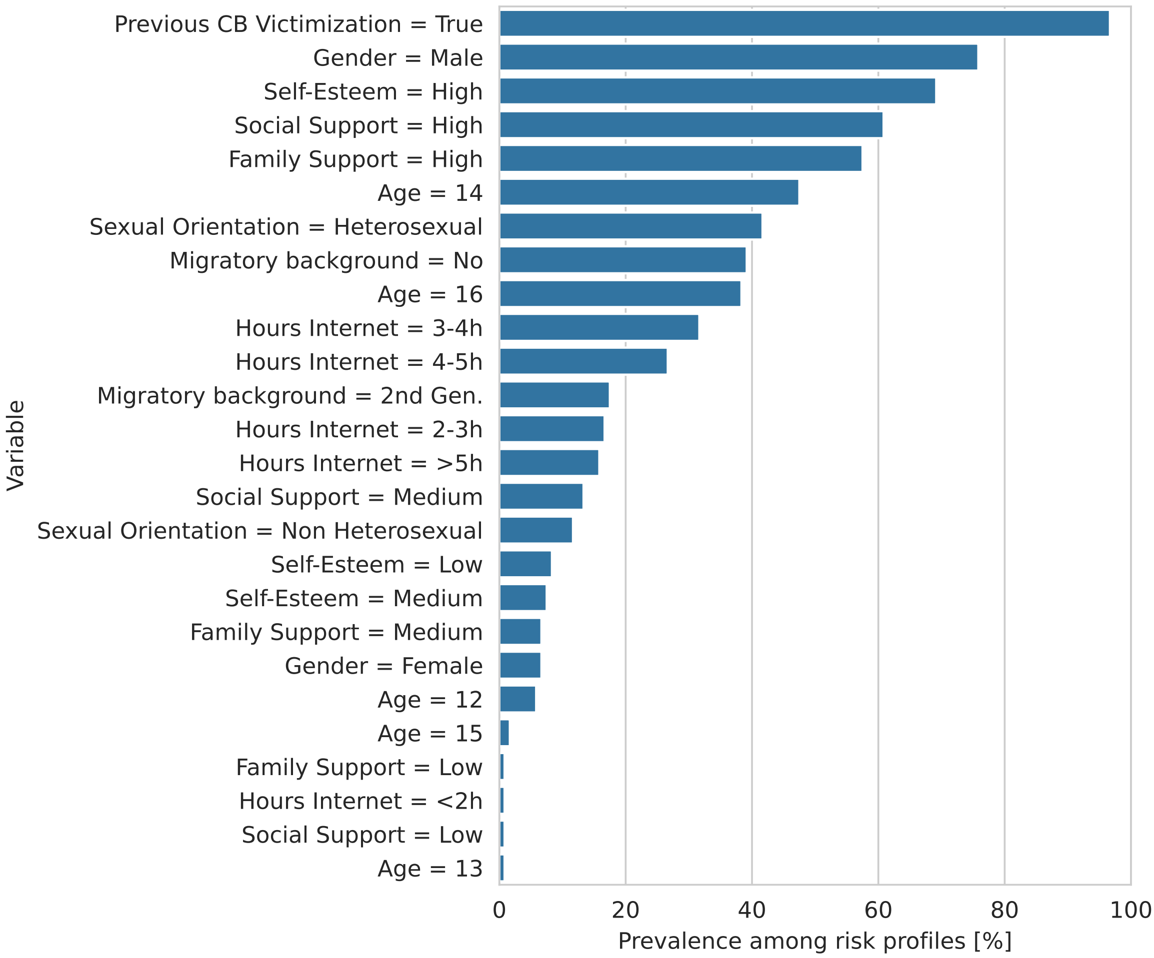


Figure 5.8: Multi-factor profiling analysis: Characteristics shared by risk profiles for 7 fixed evidences. The figure displays the prevalence of each characteristic among the identified risk profiles. We consider a profile to be risky when the posterior probability of CB offending is greater than 53% (Equation 5.2). Notably, *Previous CB Victimization = True* has a prevalence of over 90%.

Figure 5.9 intuitively illustrates the importance of a player's actions compared to their profile characteristics. The left side of the figure shows how the probability of committing CB (estimated by the PGCM) increases as we observe certain profile characteristics considered risky (previous victimization, male, high number of hours spent on the Internet, etc.). Once a sufficiently high-risk profile is reached (60%), we simulate this person playing Adventure 3 of RAYUELA's SG (the entire set of questions can be found in Appendix A). Depending on their responses to the situations presented, their estimated risk of CB will increase to 94% on the riskiest path or decrease to 18% on the safest path. This figure illustrates that profiling alone seems too narrow for estimating risk, and we need to examine how the person would behave in risky situations (simulated in the SG).

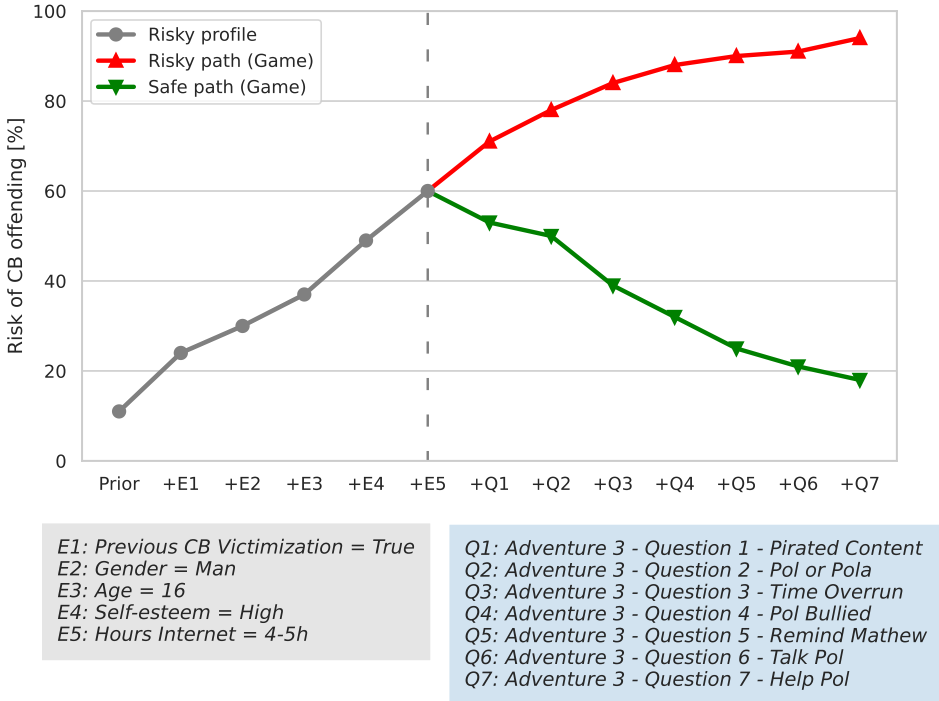


Figure 5.9: Analysis of the risk of committing cyberbullying (estimated by the PGCM) by comparing a high-risk profile that makes risky decisions in the SG with a profile that makes safe decisions in the SG. First, a set of profile characteristics (i.e., evidences) that increase risk by up to 60% are set as observations. Then, two response paths are simulated in Adventure 3 of RAYUELA's SG. Depending on how the person reacts to the situations presented, their risk can increase or decrease significantly, despite the fact that starts from a risky profile.

5.3.4 Offender Profile: Comparison with Previous Research

- **Previous Victimization:** STRONG AGREEMENT

  Previous research suggests that individuals who have been victims of traditional bullying or previous CB are more likely to become cyberbullies themselves [21, 152]. Our results show that this variable is of fundamental importance in explaining the occurrence of CB offending.

- **Gender:** AGREEMENT

  Historically, studies of traditional bullying have consistently shown that men are more likely than women to engage in bullying behavior [230]. In CB, gender differences may not be as prominent but it appears that males commit more direct cyber-aggressions, while females tend to engage in more indirect harassment (e.g., spreading false rumours about the victim) [22]. A recent study with Spanish minors shows a higher prevalence of CB offending among males [190]. Our results are consistent with males committing more CB offences.

- **Hours spent on the Internet:** AGREEMENT

  Previous research has linked greater amounts of time spent on the Internet with increased CB activities [144]. Our results also show that more hours online leads to an increase in CB offending, although it does not appear as one of the most relevant variables for offender profiling (moderate relevance).

- **Age:** MODERATE AGREEMENT

  Research on the role of age in CB offending has yielded mixed results. Most studies agree that age is an influential factor, but its exact role is complex and may be influenced by other variables [23]. Some studies report a peak around ages of 14 and 15 before decreasing in later years [143]. Our results also show that age is a moderately relevant factor with a complex relationship, with ages 14 and 16 being relatively prominent.

- **Self-esteem:** MODERATE AGREEMENT

  Studies have shown that individuals with low or high self-esteem (i.e., extreme values) are more likely to engage in CB offenses [146]. In our results we have only identified risk profiles with high self-esteem, not with low self-esteem. One possible explanation is that the game failed to correctly identify risk-takers with low self-esteem, either due to the design of the questions or the scenario in which the pilots take place.

- **Social and Family support:** DISAGREEMENT

  Recent studies found that perceived support from family and friends can serve as protective factors against CB perpetration [96, 15]. Additionally, poor perceived social and family support are strongly associated with CB perpetration [15]. Our results are contradictory to previous research evidence, suggesting that risky profiles have high social and family support. A possible explanation for this discrepancy is that our results identify only 'popular' offenders with a strong support network, but that the data entries for low social and family support are sparse and noisy and therefore difficult to identify in this scenario. Nonetheless, this disagreement invites the scientific community to investigate this particular factors further.

5.4 Discussion and Limitations

Our research has shown that data obtained from SGs provide valuable insights into CB-related behaviors. These findings exemplify the effectiveness of SGs as a novel tool for social research that can be used as a non-invasive and enjoyable way to gather information from hard-to-reach populations, such as minors.

Despite these promising findings, we should be cautious in interpreting and generalizing the results, and acknowledge the possible limitations of the work. In particular, by identifying experimental limitations (e.g., biases in data collection), methodological (e.g., algorithms or techniques used), or interpretative ones.

The data were collected solely in European countries, so further experiments in different populations would be needed to confirm the results' reliability. It should also be noted that questionnaire and video game data are often noisy and heterogeneous, especially when dealing with minors (e.g., some participants may have played randomly or deliberately answered questions incorrectly).

As discussed in Chapter 2, PGCMs mitigate data biases (e.g., selection, colliders, confounders) [222, 259]. This approach also forces us to make our assumptions and hypotheses explicit, leading to discussions and critical questions about the issues we are trying to address and contributing to more accessible and comprehensible science. However, at the same time, causal models rely heavily on assumptions such as the no unmeasured confounders (e.g., socioeconomic status, time spent on social networks), the stability of causal relationships over time and populations, and the correct specification of the model. The causal estimates may be biased or misleading if these assumptions are violated or incorrect.

The results show that variables obtained through the SG are relevant to explaining and predicting the risk of committing CB. Demographic variables and those obtained through psychological questionnaires do not have significant relevance when analyzed individually, except for the variable *Previous CB Victimization* (5.3.2). This means that individuals who have previously been victims of CB showed a significantly higher propensity to commit it. However, when analyzing combinations, the profiling variables (demographic and psychological questionnaire data) are of equal or greater importance than the game questions (Section 5.3.3).

5.5 Conclusions

This chapter proposed a methodology to analyze data from SGs using causal-based analysis. For this purpose, we propose using PGCM, which provides a sophisticated framework for graphically modeling causal relations. Furthermore, it allows the introduction of expert knowledge intuitively into the network structure and prior probabilities. This research aims to demonstrate the validity and potential of using SGs as practical research tools in social science. We applied a case study on a SG about CB with data collected through experimental pilots within the H2020 RAYUELA project. Therefore, this research also aims to provide novel knowledge about the underlying causal mechanisms of CB.

The causal analysis results from the case study confirm the relevance of the game questions in identifying and understanding the CB phenomenon, which could be used to devise effective prevention and training strategies for CB among minors. The game questions with the most weight in explaining previous CB aggression are related to the importance players place on early stages of harassment, spending too much time online without taking breaks, and the perceived risk of downloading pirated content from the Internet.

It is worth noting that the demographic variable with the highest relevance is *Previous CB Victimization*, which means that minors who had been victims of CB were more likely to become offenders themselves. This cycle of victimization and perpetration underscores the importance of addressing past victimization experiences in understanding and preventing future CB behaviors.

SGs are postulated as an effective alternative to measure how players would act in certain real-life situations, provided that there are ethical constraints and the game is well-designed for the desired function. Using a SG contributes to better immersion and disinhibition of players, which is crucial for studying sensitive issues or hard-to-measure segments of the population, such as minors. Moreover, the causal analysis proves particularly helpful in the in-depth understanding of sensitive issues, contributing to developing more effective preventive measures to reduce cybercrime among minors and ensure their well-being online.

6 Conclusions, Limitations and Future Work

This chapter serves to conclude the work carried out. First, the main conclusions are presented. Secondly, the original contributions to scientific research are outlined. Finally, promising future lines of work are proposed.

6.1 Conclusions

In this doctoral dissertation, we have examined the intersection of Serious Games (SG) and Artificial Intelligence (AI), unraveling the challenges and promising future research opportunities, with particular emphasis on potential applications in social science research. This exploration started with a background review in Chapter 2, where we identify synthetic data generation, data sharing, and causality as key emerging fields of interest. These areas are not only fundamental in today's academic environment, but also have the potential to influence the trajectory of SG development and AI integration significantly. To help contextualize this work, Chapter 2 also identified and illustrated the major applications of SGs in industry and research, as well as the applications that AI has historically had in SGs.

Chapter 3 tackles the construction of robust causal network structures or causal Direct Acyclic Graphs (DAG), an essential issue when working with Probabilistic Graphical Causal Models (PGCM). We present a novel methodology that combines expert knowledge with algorithmic approaches to construct consensus-based DAGs. This method was empirically validated using data from a survey among Spanish minors, conducted within the RAYUELA project. Our analysis revealed that of the variables analyzed in this case study, *Age* plays a significant role in influencing cyberbullying (CB) victimization. However, this model includes a relatively limited number of variables, so it is likely that *Age* acts as a confounder for variables not captured in the survey or model. Nevertheless, from an intervention and policy development perspective, these findings suggest that efforts should be focused on prevention efforts at critical ages.

In Chapter 4, we design and implement a simulator capable of generating probabilistic synthetic data for decision-making based SGs. Although the proposed architecture could be used to generate data in any type of decision-making scenario (e.g., multiple-choice exams, political polls, psychological questionnaires). We propose to use models based on Item Response Theory (IRT) to realistically mimic human behavior in these environments. In addition, we used PGCMs to inform the generation with real data and guide it through an explicit data generation process, increasing realism and control over the synthetic data produced. With this contribution, we demonstrate the great potential of synthetic data generation to accelerate development times, standardize data structures and pipelines, and address data deficiencies such as under-representation of certain populations.

Chapter 5 converges the central themes of the thesis by proposing a causality-based methodology for analyzing data derived from SG research, which enables combining profiling (e.g., demographic, environmental) with gameplay data. We demonstrate the validity of our proposal using data from the experimental pilots of the SG from the RAYUELA project, specifically addressing the CB issue. Notably, our research indicates that individuals previously victimized by CB are more predisposed to become offenders themselves. Examining the game questions that most influence past CB aggression: *Adventure 3 Question 4 (Pol Bullied)* and *Adventure 1 Question 3 (Matthew Meme)* address the importance the player places on situations in the early stages of CB; *Adventure 3 Question 3 (Time Overrun)* covers being online too long without taking breaks; and *Adventure 3 Question 1 (Pirated Content)* examines the perceived risk of downloading pirated content from the Internet. Our findings illustrate the significant role of SGs in observing CB dynamics, suggesting their potential in crafting impactful prevention and intervention strategies among minors.

This thesis argues for the integration of SGs as a valuable research tool, especially within social science disciplines. Having particularly beneficial advantages such as the disinhibition effect or enhancing player immersion in simulated environments, thus facilitating access to hard-to-reach populations like minors. Moreover, the potential for large-scale experimentation offered by SGs could address the challenges of data scarcity and variability.

Furthermore, we advocate for applying causal AI, particularly PGCM, in analyzing data derived from SG experiments. This approach enables the examination of extensive datasets to uncover patterns that traditional statistical methods may overlook, albeit in a more robust manner than proposals based on Machine Learning due to their tendency to rely on data biases and spurious relationships to make predictions. Remarkably, PGCMs also allow for the simulation of interventions, essential for developing effective prevention strategies and policies. The application of causal analysis is particularly valuable in addressing complex issues such as CB, providing a more nuanced understanding and potential solutions, encouraging critical discussion, and improving transparency.

In summary, this dissertation not only demonstrates the significant potential of using SGs in research but also advances the methodological discourse by integrating causal AI techniques. The implications of this research extend beyond academic boundaries, offering practical insights and tools for tackling urgent societal challenges.

The objectives of this doctoral thesis described at the beginning of this document (Section 1.3) have been successfully achieved:

1. **Develop a methodology to generate robust causal model structures (DAG) that unify expert knowledge and results of automatic structure learning algorithms.**

   This objective has been addressed in Chapter 3 and this methodology has been applied in subsequent chapters to achieve reasonable DAGs to work with.

2. **Develop a general probabilistic framework to produce synthetic data that models human behavior in decision-based serious games.**

   In Chapter 4 we have designed and implemented a generalist simulator to generate synthetic data in decision-based SGs utilizing PGCM and IRT models.

3. **Build a causality-based computational methodology for analyzing Serious Game data.**

   Chapter 5 converges the themes discussed throughout this thesis, proposing a methodology for analyzing SG data with a causal perspective.

4. **Identify relevant risk factors for cyberbullying using causality-based techniques.**

   Throughout Chapters 3, 4, and 5, a series of use cases have been presented within the framework of the RAYUELA project, specifically addressing the problem of CB. Through this thesis, in addition to proposing novel methodologies in the field of SGs and AI, we also provide new insights into the intricate mechanisms of CB.

6.2 Original contributions of the thesis

Articles published in peer-reviewed academic journals as first author

- J. Pérez, M. Castro and G. López, "Serious Games and AI: Challenges and Opportunities for Computational Social Science," in *IEEE Access*, vol. 11, 2023, pp. 62051-62061. DOI: 10.1109/ACCESS.2023.3286695. JCR: 3,900 Q2 (2022) - SJR: 0,926 Q1 (2022).

- J. Pérez, M. Castro, E. Awad and G. López, "Generation of probabilistic synthetic data for serious games: A case study on cyberbullying," in *Knowledge-Based Systems*, Volume 286, 2024, pp. 111440, ISSN 0950-7051. DOI: 10.1016/j.knosys.2024.111440. JCR: 8,800 Q1 (2022) - SJR: 2,065 Q1 (2022).

Articles presented at international conferences as first author

- J. Pérez, E. Awad, M. Castro, G. López, N. Bueno-Guerra, M. Reneses, M. Riberas-Gutiérrez and A. Gómez-Dorado, "A computational framework for understanding risk factors in cybercrime," *8th International Conference on Computational Social Science - IC2S2 2022*, Chicago (USA). 19-22 july 2022.

- J. Pérez, V. Balmaseda, A.L. Urbistondo, E. Awad, M. Castro, G. López , "A child's play: an agent-based simulator to protect minors online," *8th International Conference on Computational Social Science - IC2S2 2022*, Chicago (USA). 19-22 july 2022.

- J. Pérez, E. Awad, M. Castro and G. López, "Causality guiding survey analysis: a use case on cyberbullying," *9th International Conference on Computational Social Science - IC2S2 2023*, Copenhague (Denmark). 17-20 july 2023. URL: `https://openreview.net/forum?id=99EHn8T0r3L`.

Collaborative articles published in peer-reviewed academic journals

- S. Solera-Cotanilla, M. Vega-Barbas, J. Pérez, G. López, J. Matanza, and M. Álvarez-Campana, "Security and Privacy Analysis of Youth-Oriented Connected Devices," in *Sensors* 22, no. 11: 3967. DOI: 10.3390/s22113967. JCR: 3,900 Q2 (2022); - SJR: 0,764 Q1 (2022).

- J. Fúster, S. Solera-Cotanilla, J. Pérez, M. Vega-Barbas, R. Palacios, M. Álvarez-Campana and G. López, "Analysis of security and privacy issues in wearables for minors," in *Wireless Networks*, 2023. DOI: 10.1007/s11276-022-03211-6 . JCR: 3,000 Q2 (2022) - SJR: 0,706 Q2 (2022).

- C. Valero, J. Pérez, S. Solera-Cotanilla, M. Vega-Barbas, G. Suarez-Tangil, M. Alvarez-Campana and G. López, "Analysis of security and data control in smart personal assistants from the user's perspective", in *Future Generation Computer Systems*, Volume 144, 2023, Pages 12-23, ISSN 0167-739X, 10.1016/j.future.2023.02.009. JCR: 7,500 Q1 (2022) - SJR: 2,043 Q1 (2022).

Collaborative articles presented at international conferences

- G. López, N. Bueno-Guerra, M. Castro, M. Reneses, J. Pérez, M. Riberas-Gutiérrez, M. Álvarez-Campana, M. Vega-Barbas, S. Solera-Cotanilla, L. Bastida, A. Moya, R. Fernández, V. Vázquez, G. Zango and P. Vicente, "The H2020 project RAYUELA: a fun way to fight cybercrime," *Jornadas Nacionales de Investigación en Ciberseguridad - JNIC 2021*, Cuenca (Spain) Online. 09-10 june 2021. DOI: 10.18239/jornadas_2021.34.27

- S. Solera-Cotanilla, J. Fúster de la Fuente, J. Pérez, R. Palacios, M. Vega-Barbas, M. Álvarez-Campana and G. López , "Análisis de problemas de seguridad y privacidad en wearables usados por menores," *VII Jornadas Nacionales de Investigación en Ciberseguridad - JNIC 2022*, Bilbao (España). 27-29 june 2022. ISBN 978-84-88734-13-6, pp. 209-215. URL: https://2022.jnic.es/Actas_JNIC_2022_v11.pdf.

- C. Valero, J. Pérez, S. Solera-Cotanilla, M. Vega-Barbas, G. Suárez, G. López and M. Álvarez-Campana "Evaluando la seguridad y privacidad de los asistentes personales inteligentes: ¡Ojo con el juguete!," *VII Jornadas Nacionales de Investigación en Ciberseguridad - JNIC 2022*, Bilbao (España). 27-29 junio 2022. ISBN 978-84-88734-13-6, pp. 220-227. URL: https://2022.jnic.es/Actas_JNIC_2022_v11.pdf.

6.3 Limitations

One of the main challenges we face in the field of causal inference is that is still emerging. Unlike machine learning, there is a lack of uniformity in causal inference due to the various adaptations of methodologies and terminologies across different domains. This makes it difficult to compare model performance and generalize results. While this heterogeneity enriches the field, it also highlights the need for a cohesive framework that will allow causal inference to evolve and be more widely adopted.

When using PGCM, we encounter certain limitations that originate from the DAG itself. As the number of variables increases, the interpretability of DAGs becomes more challenging, and the computational cost increases significantly. This limitation restricts the scalability of the proposed method to create robust causal DAGs proposed in Chapter 3 of this thesis. Additionally, we depend on expert assumptions to identify forbidden connections, which may introduce bias from either the experts or the current paradigms.

The conclusions we can draw from the CB case studies are limited to the variables included in the models. That is, there may be unmeasured factors that also significantly influence the outcome (e.g., socioeconomic status, time spent on social networks). However, this limitation is shared with any other scientific study addressing this issue. Furthermore, the validity of our conclusions is closely related to the quality of our data, which in this case is limited by its relatively small size, the noisy nature of the data (due to the fact that the participants were minors playing a video game), and its geographical limitation to European countries.

Likewise, causal models are based on fundamental assumptions, such as the stability of causal relationships over time and between populations. Any deviation from these assumptions can lead to biased or misleading causal estimates, highlighting the delicate balance required in constructing and interpreting causal models.

Despite these challenges, it is imperative to view these limitations not as undermining the value of our work but as integral components of the scientific research process. They provide a roadmap for future research and point to areas where further methodological advances and empirical studies are needed. By presenting and considering these limitations, we gain a deeper understanding of causal inference and its application to social issues like CB.

6.4 Future Work

Throughout the development of this dissertation, AI has made impressive strides forward, capturing the public's attention. These advances have been particularly revolutionary in Large Language Models (LLMs). One of the most promising prospects that emerged from our research lies in the convergence of LLMs and causality [129, 90]. For instance, discovering causal DAGs [113] or creating intelligent agents [135]. This new field has the potential to transform our view of LLMs from mere predictors to "intelligent" models capable of deciphering the causal relationships of the world through language.

The field of SGs has much potential, and it is in a great position to benefit from the revolution brought by generative AI. Using AI to speed up the design and development of video games opens up new possibilities [257]. It imagines a future where the time from coming up with an idea to actually creating it is significantly reduced, making it possible to quickly turn conceptual frameworks into tangible educational and training tools. This acceleration may improve the field of SGs and make advanced resources more accessible. Moreover, causality-based methodologies can be applied in diverse fields, such as education and training, transforming how we teach and develop skills.

The intersection of machine learning and causality in the context of SGs also offers fertile ground for future research [212, 200] . Moving beyond conditional probabilities to adopt more nuanced relationship models can significantly improve the effectiveness of SG-based research.

Regarding the synthetic player simulator proposed in Chapter 4 of this thesis, there are numerous potential improvements. For example, we suggest exploring the use of more complex profile models consisting of more than one variable. Also, to explore the potential impact of introducing memory to the agents, which could lead to a more realistic model that allows past responses to influence future responses.

The findings of this research are fraught with challenges but also with promise and potential discoveries. The suggested areas for future research offer a base for other researchers to expand upon, investigate, and ultimately revolutionize the field of SGs and causal AI.

A Game Questions

Adventure 1

Question 1: Photo Sharing
Context: [Talking to Matthew after taking a selfie.]
Dialogue: *Now we only have to share and tag the photos. Jane, do you want to share them, or do you prefer me to do it?*

(a) *I will do it.*
(b) *You can do it.*

Question 2: Sociable
Context: [Talking to Robert after sharing the selfie. Dialog depends on the previous answers.]
Dialogue: *It seems you like to upload many photos and share stuff on your social network.*

(a) *I would say I am sociable.*
(b) *I consider myself kind of shy.*

Question 3: Matthew Meme
Context: [After receiving a message from Patty with the meme about Matthew.]

(a) *Hehe, it's funny, I will share the meme.*
(b) *I am not going to share it.*
(c) *I won't share the meme and I'm going to try to end up with this.*

Adventure 3

Question 1: Pirated Content
Context: [Playing video games in your room.]
Dialogue: *I know of some sites that pirate the content and then you can download the update for free.*

(a) *I will download the pirated update through a website.*
(b) *I will wait until I have money or until my parents give me the money to buy the new expansion.*

Question 2: Pol or Paula
Context: [Playing video games in your room. Your friends joke about Pol's appearance.]

(a) *I don't like this kind of jokes.*

(b) *That's funny.*

(c) *Say nothing.*

Question 3: Time Overrun
Context: [Playing video games in your room. A warning pops up about the number of hours you have been online]

(a) *4 hours are not that much. So, I can keep chatting a bit longer.*

(b) *It's time to stop and disconnect for a while, although I might miss some juicy gossiping.*

Question 4: Pol Bullied
Dialogue: *I heard that some guys are messing with Paul. I am worried that he may be bullied. What do you think?*

(a) *They are just having fun; I would not call that bullying.*

(b) *I think it's not right... but calling that bullying is a bit of a stretch.*

(c) *I think that's unacceptable; we should do something about it.*

Question 5: Remind Matthew
Context: [Your friends start messing with Pol.]
Dialogue: *Has anything like this ever happened to you?*

(a) *Yes, I had a similar bad experience... I don't like being picked on.*

(b) *No, it has never really happened to me, to my knowledge.*

(c) *Yes, it was me who messed with someone else... but it was not such a big deal.*

Question 6: Talk to Pol
Dialogue: *So, shall we talk to Pol to see how he is?*

(a) *It is better to let him be.*

(b) *Of course, we should try to help.*

Question 7: How to Help Pol
Dialogue: *We can help you. You are not alone in this. I think...*

(a) *We should go to tell the teacher, he should know what to do.*

(b) *We should report the comments to the social network, so that it doesn't happen again.*

(c) *We should not report it, because I don't want to get picked on for being a snitch...*

(d) *We should not report it as reporting is usually useless.*

B Exploratory Data Analysis (I): Survey of Spanish minors

The following is a set of descriptive statistics on the data through a representative survey of children in schools in Madrid (Spain) within the RAYUELA project. The survey collected responses from 665 students.

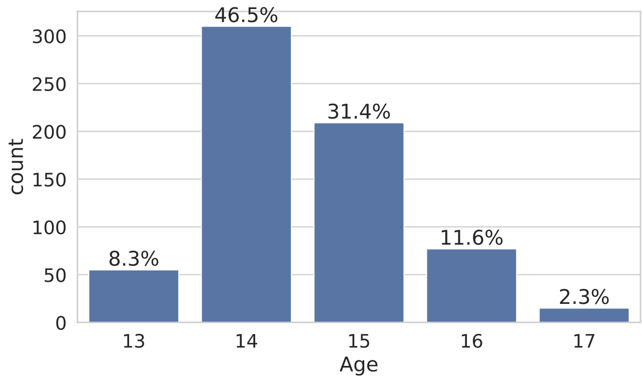


Figure B.1: Age distribution in the dataset.

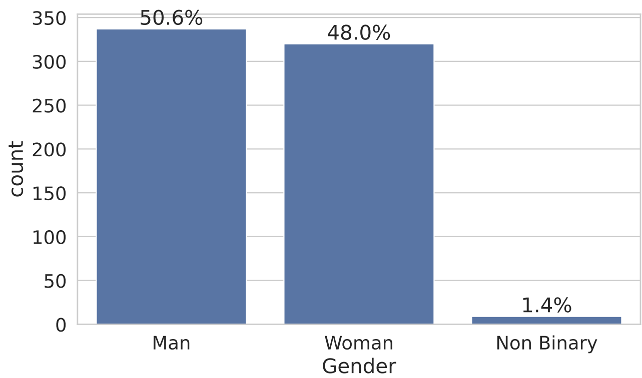


Figure B.2: Gender distribution in the dataset.

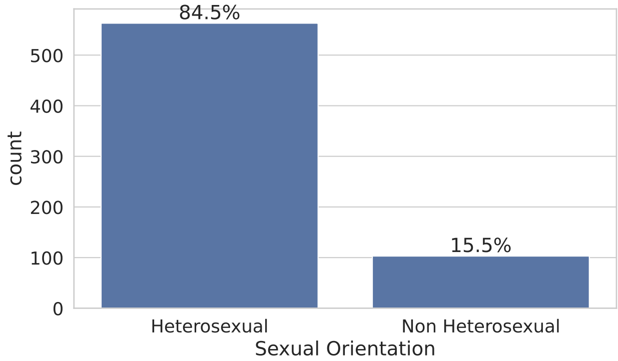


Figure B.3: Sexual Orientation distribution in the dataset. The *Non Heterosexual* bar aggregates the values *Bisexual*, *Homosexual*, *I do not know yet*, and *other*.

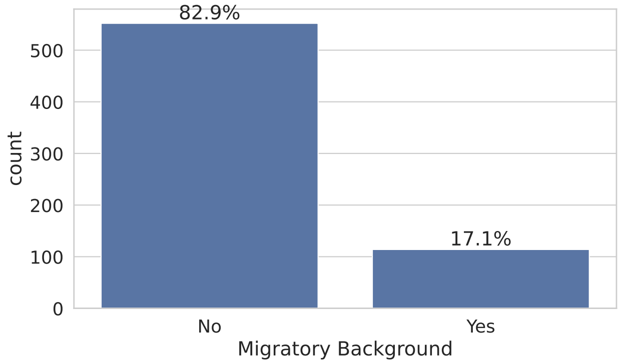


Figure B.4: Migratory Background distribution in the dataset.

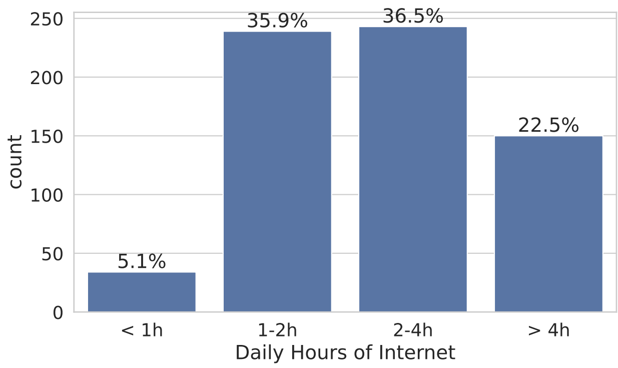


Figure B.5: Self-reported hours of Internet distribution (aggregated) in the dataset.

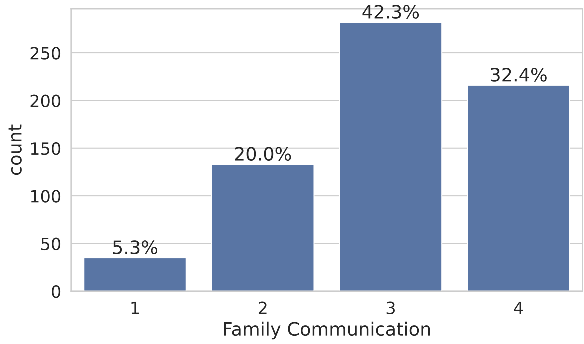


Figure B.6: Self-reported family communication distribution in the dataset. Ranked from 1 (never) to 4 (very frequently).

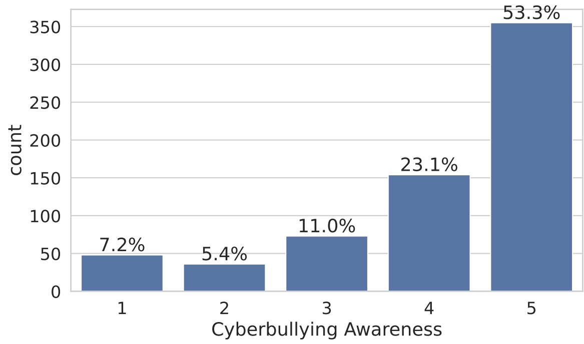


Figure B.7: Self-reported cyberbullying awareness distribution in the dataset. Ranked from 1 (not worried) to 5 (very worried).

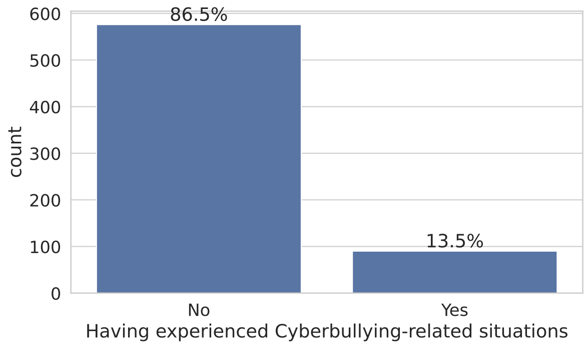


Figure B.8: Self-reported "having experienced cyberbullying-related situations" distribution in the dataset.

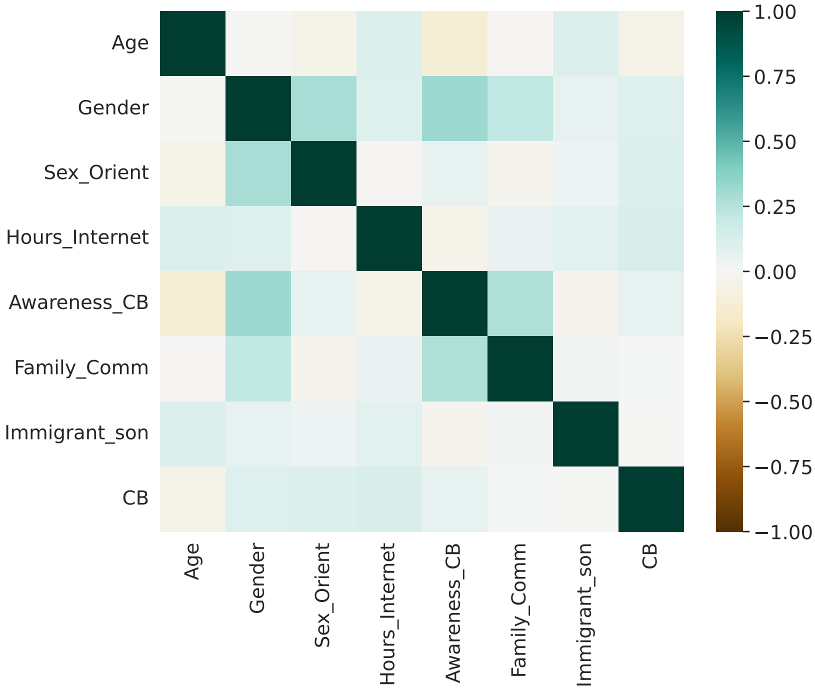


Figure B.9: Heatmap of *Pearson* correlations between variables.

C Exploratory Data Analysis (II): Experimental pilots using RAYUELA's serious game

The following is a set of descriptive statistics on the data that was collected during the experimental pilots of RAYUELA in European schools and institutes. We gathered responses from 1055 students. In this exploratory analysis, we did not consider the game data. We have only considered demographic data and the psychological/sociological questionnaires that students completed before and after playing the Serious Game.

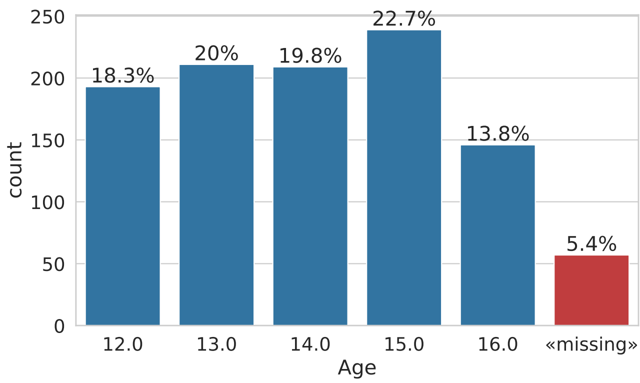


Figure C.1: Age distribution in the dataset. The «missing» bar (red) indicates the proportion of erroneous values or values that participants did not want to fill in.

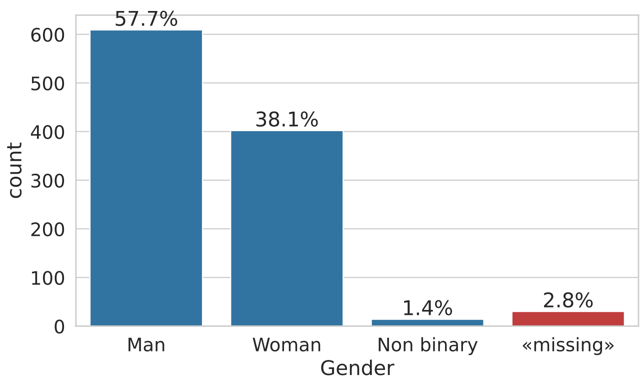


Figure C.2: Gender distribution in the dataset. The «missing» bar (red) indicates the proportion of erroneous values or values that participants did not want to fill in.

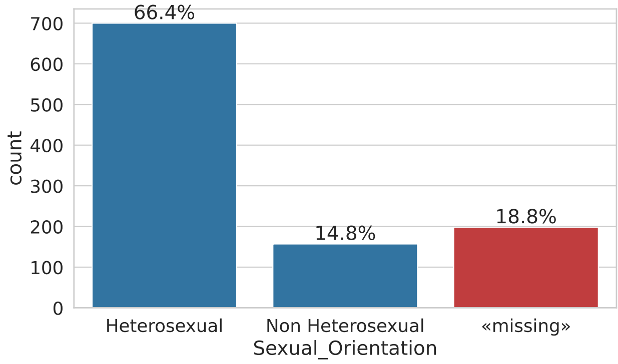


Figure C.3: Sexual Orientation distribution in the dataset. The *Non Heterosexual* bar aggregates the values *Bisexual*, *Homosexual*, *I do not know yet*, and *other*. The «missing» bar (red) indicates the proportion of erroneous values or values that participants did not want to fill in.

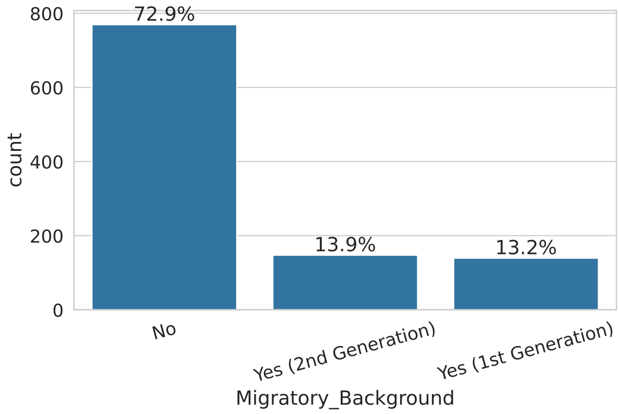


Figure C.4: Migratory Background distribution in the dataset. The *2nd generation* bar shows participants who were born in the country but whose parents were born abroad. The *1st generation* bar shows participants who were not born in the country.

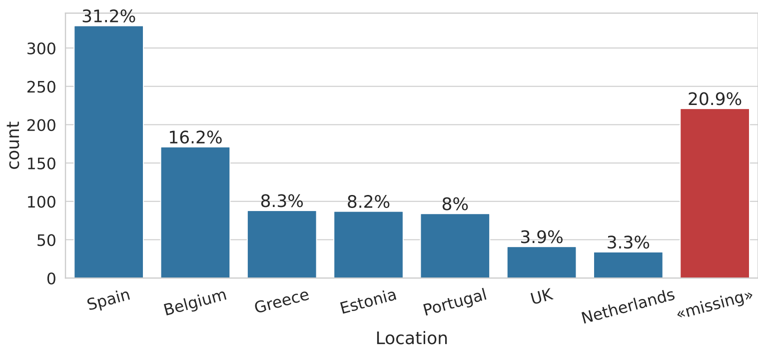


Figure C.5: Country distribution in the dataset. The «missing» bar (red) indicates the proportion of erroneous values or values that participants did not want to fill in.

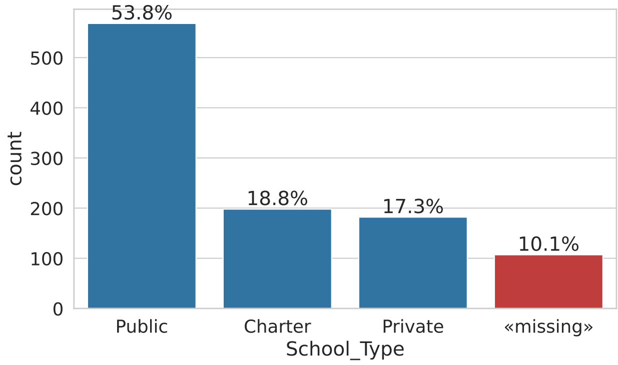


Figure C.6: School type distribution in the dataset. The «missing» bar (red) indicates the proportion of erroneous values or values that participants did not want to fill in.

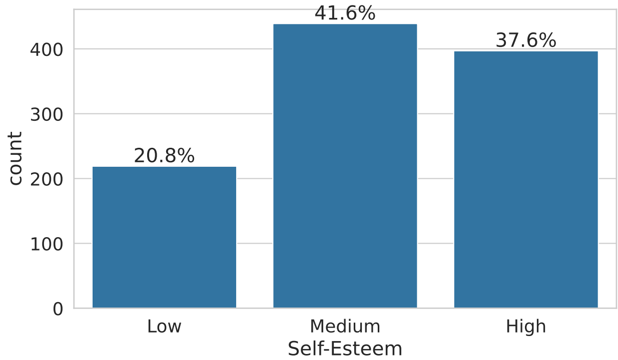


Figure C.7: Self-Esteem distribution in the dataset, obtained though the *Rosenberg Self-Esteem Scale* [204].

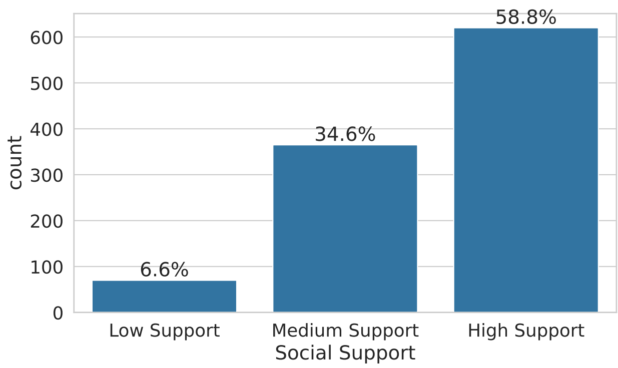


Figure C.8: Social support distribution in the dataset, obtained though *The Multidimensional Scale of Perceived Social Support* [272].

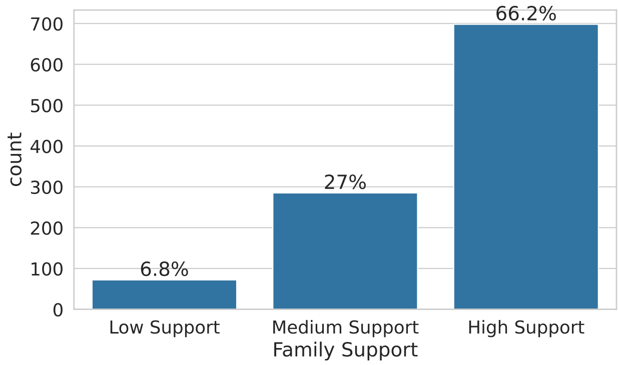


Figure C.9: Family support distribution in the dataset, obtained though *The Multidimensional Scale of Perceived Social Support* [272].

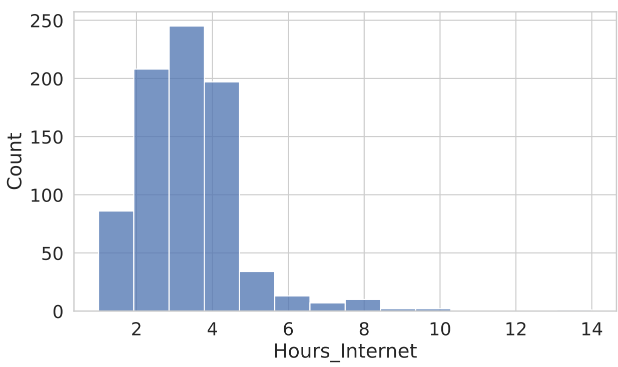


Figure C.10: Self-reported hours of Internet distribution (histogram) in the dataset.

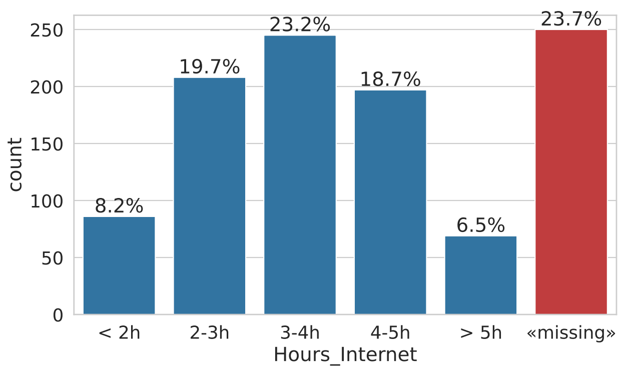


Figure C.11: Self-reported hours of Internet distribution (aggregated) in the dataset. The «missing» bar (red) indicates the proportion of erroneous values or values that participants did not want to fill in.

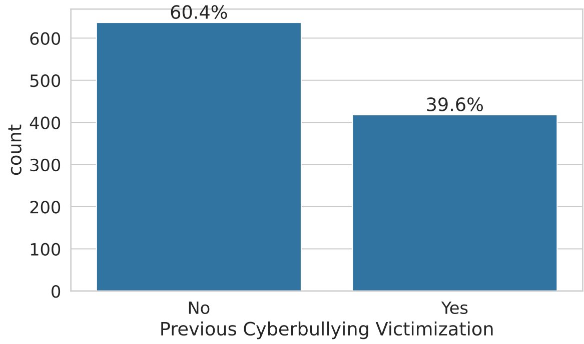


Figure C.12: Previous cyberbullying victimization distribution in the dataset, obtained though the *European Cyberbullying Intervention Project Questionnaire* [28].

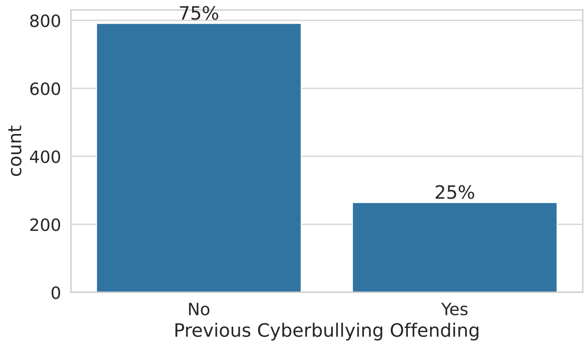


Figure C.13: Previous cyberbullying offending distribution in the dataset, obtained though the *European Cyberbullying Intervention Project Questionnaire* [28].

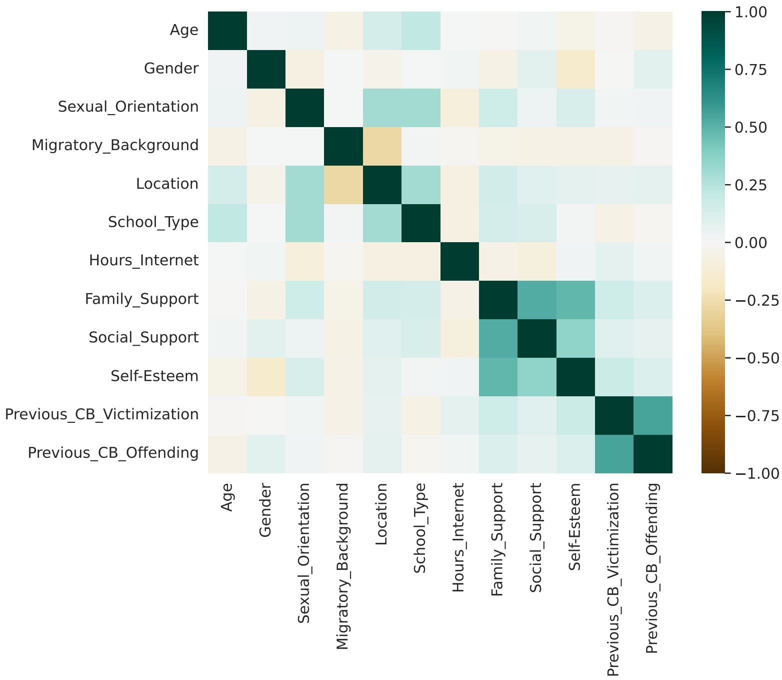


Figure C.14: Heatmap of *Pearson* correlations between variables.

Bibliography

1. *6 Favorite Business Simulations to Teach and Why*. Harvard University. Harvard Business Publishing - Education. 2021. URL: https://hbsp.harvard.edu/inspiring-minds/6-favorite-business-simulations-to-teach-and-why.
2. E. Abi-Jaoude, K. T. Naylor, and A. Pignatiello. “Smartphones, social media use and youth mental health”. *Canadian Medical Association Journal (CMAJ)* 192:6, 2020, E136–E141. ISSN: 0820-3946. DOI: 10.1503/cmaj.190434.
3. C. C. Abt. *Serious games*. University press of America, 1987. ISBN: 0819161489.
4. B. Aczel, B. Palfi, and B. Szaszi. “Estimating the evidential value of significant results in psychological science”. *PLoS One* 12:8, 2017, e0182651. DOI: 10.1371/journal.pone.0182651.
5. E. M. Airoldi. “Getting started in probabilistic graphical models”. *PLoS Comput Biol* 3:12, 2007, e252. DOI: 10.1371/journal.pcbi.0030252.
6. S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, and R. G. Baraniuk. “Self-Consuming Generative Models Go MAD”. *arXiv preprint arXiv:2307.01850*, 2023. DOI: 10.48550/ARXIV.2307.01850.
7. C. Alonso-Fernández, A. Calvo-Morata, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón. “Applications of data science to game learning analytics data: A systematic literature review”. *Computers & Education* 141, 2019, p. 103612. ISSN: 0360-1315. DOI: 10.1016/j.compedu.2019.103612.
8. C. Alonso-Fernández, A. R. Cano, A. Calvo-Morata, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón. “Lessons learned applying learning analytics to assess serious games”. *Computers in Human Behavior* 99, 2019, pp. 301–309. ISSN: 0747-5632. DOI: 10.1016/j.chb.2019.05.036.
9. C. Alonso-Fernández, I. Martínez-Ortiz, R. Caballero, M. Freire, and B. Fernández-Manjón. “Predicting students’ knowledge after playing a serious game based on learning analytics data: A case study”. *Journal of Computer Assisted Learning* 36:3, 2020, pp. 350–358. DOI: 10.1111/jcal.12405.
10. H. Alvari and P. Shakarian. “Causal inference for early detection of pathogenic social media accounts”. *arXiv preprint arXiv:1806.09787*, 2018. DOI: 10.48550/ARXIV.1806.09787.
11. *America’s Army*. U.S. Army. PC game. 2002.
12. H. Amirkhani, M. Rahmati, P. J. F. Lucas, and A. Hommersom. “Exploiting Experts’ Knowledge for Structure Learning of Bayesian Networks”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39:11, 2017, pp. 2154–2170. DOI: 10.1109/TPAMI.2016.2636828.
13. K. Andersen, S. J. Gaab, J. Sattarvand, and F. C. Harris. “METS VR: Mining Evacuation Training Simulator in Virtual Reality for Underground Mines”. In: *17th International Conference on Information Technology–New Generations (ITNG 2020)*. Springer International Publishing, Cham, 2020, pp. 325–332. ISBN: 978-3-030-43020-7. DOI: 10.1007/978-3-030-43020-7_43.
14. A. Antoniou, A. Storkey, and H. Edwards. “Augmenting Image Classifiers Using Data Augmentation Generative Adversarial Networks”. In: *Lecture Notes in Computer Science*. Springer International Publishing, 2018, pp. 594–603. ISBN: 9783030014247. DOI: 10.1007/978-3-030-01424-7_58.

15. N. Arató, A. N. Zsidó, A. Rivnyák, B. Péley, and B. Lábadi. "Risk and Protective Factors in Cyberbullying: the Role of Family, Social Support and Emotion Regulation". *International Journal of Bullying Prevention* 4:2, 2022, pp. 160–173. ISSN: 2523-3661. DOI: 10.1007/s42380-021-00097-4.

16. *Artificial Intelligence: Emerging Opportunities, Challenges, and Implications for Policy and Research*. Technical report. U.S. Government Accountability Office, 2018. URL: https://www.gao.gov/assets/gao-18-644t.pdf.

17. K. Athanasiou, E. Melegkovits, E. K. Andrie, C. Magoulas, C. K. Tzavara, C. Richardson, D. Greydanus, M. Tsolia, and A. K. Tsitsika. "Cross-national aspects of cyberbullying victimization among 14–17-year-old adolescents across seven European countries". *BMC Public Health* 18:1, 2018, p. 800. ISSN: 1471-2458. DOI: 10.1186/s12889-018-5682-4.

18. A. H. Aubert, R. Bauer, and J. Lienert. "A review of water-related serious games to specify use in environmental Multi-Criteria Decision Analysis". *Environmental Modelling & Software* 105, 2018, pp. 64–78. ISSN: 1364-8152. DOI: 10.1016/j.envsoft.2018.03.023.

19. E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. "The Moral Machine experiment". en. *Nature* 563:7729, 2018, pp. 59–64. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0637-6.

20. I. Ayed, A. Ghazel, A. Jaume-i-Capó, G. Moyà-Alcover, J. Varona, and P. Martínez-Bueso. "Vision-based serious games and virtual reality systems for motor rehabilitation: A review geared toward a research methodology". *International Journal of Medical Informatics* 131, 2019, p. 103909. ISSN: 1386-5056. DOI: 10.1016/j.ijmedinf.2019.06.016.

21. S.-M. Bae. "The relationship between exposure to risky online content, cyber victimization, perception of cyberbullying, and cyberbullying offending in Korean adolescents". *Children and Youth Services Review* 123, 2021, p. 105946. ISSN: 0190-7409. DOI: 10.1016/j.childyouth.2021.105946.

22. C. Barlett and S. M. Coyne. "A meta-analysis of sex differences in cyber-bullying behavior: The moderating role of age". *Aggressive Behavior* 40:5, 2014, pp. 474–488. DOI: 10.1002/ab.21555.

23. C. P. Barlett and K. Chamberlin. "Examining cyberbullying across the lifespan". *Computers in Human Behavior* 71, 2017, pp. 444–449. ISSN: 0747-5632. DOI: 10.1016/j.chb.2017.02.009.

24. A. G. Barrera Yañez, C. Alonso-Fernandez, and B. Fernandez Manjon. "Review of serious games to educate on gender equality". In: *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*. TEEM'20. Association for Computing Machinery, 2021, pp. 662–668. ISBN: 9781450388504. DOI: 10.1145/3434780.3436592.

25. *Board Games Market - Global Outlook and Forecast 2021-2026*. Technical report. Arizton, 2020. URL: https://www.arizton.com/market-reports/global-board-games-market-industry-analysis-2024.

26. I. Bogost. *Persuasive games: The expressive power of videogames*. MIT Press, 2010. ISBN: 9780262268912. DOI: 10.7551/mitpress/5334.001.0001.

27. D. Bossen, A. Broekema, B. Visser, A. Brons, A. Timmerman, F. van Etten-Jamaludin, K. Braam, and R. Engelbert. "Effectiveness of Serious Games to Increase Physical Activity in Children With a Chronic Disease: Systematic Review With Meta-Analysis". *Journal of Medical Internet Research* 22:4, 2020, e14549. ISSN: 1438-8871. DOI: 10.2196/14549.

28. A. Brighi, R. Ortega, J. Pyzalski, H. Scheithauer, P. K. Smith, H. Tsormpatzoudis, H. Tsorbatzoudis, and et al. *European Cyberbullying Intervention Project Questionnaire*. 2012. DOI: 10.1037/t66195-000.

29. J. Bughin, E. Hazan, S. Ramaswamy, M. Chui, T. Allas, P. Dahlström, N. Henke, and M. Trench. *Artificial Intelligence: The Next Digital Frontier?* Technical report. McKinsey Global Institute, 2017. URL: https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx.

30. J. P. Burgard, J.-P. Kolb, H. Merkle, and R. Münnich. "Synthetic data for open and reproducible methodological research in social sciences and official statistics". *AStA Wirtschafts- und Sozialstatistisches Archiv* 11:3, 2017, pp. 233–244. ISSN: 1863-8163. DOI: 10.1007/s11943-017-0214-8.

31. Z. Buzady and F. Almeida. "FLIGBY — A Serious Game Tool to Enhance Motivation and Competencies in Entrepreneurship". In: *Informatics*. Vol. 6. Multidisciplinary Digital Publishing Institute. 2019, p. 27. DOI: 10.3390/informatics6030027.

32. A. Calvo-Morata, C. Alonso-Fernández, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón. "Creating awareness on bullying and cyberbullying among young people: Validating the effectiveness and design of the serious game Conectado". en. *Telematics and Informatics* 60, 2021, p. 101568. ISSN: 0736-5853. DOI: 10.1016/j.tele.2021.101568.

33. A. Calvo-Morata, C. Alonso-Fernández, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón. "Serious games to prevent and detect bullying and cyberbullying: A systematic serious games and literature review". *Computers & Education* 157, 2020, p. 103958. ISSN: 0360-1315. DOI: 10.1016/j.compedu.2020.103958.

34. A. Calvo-Morata, D. C. Rotaru, C. Alonso-Fernandez, M. Freire-Moran, I. Martinez-Ortiz, and B. Fernandez-Manjon. "Validation of a Cyberbullying Serious Game Using Game Analytics". en. *IEEE Transactions on Learning Technologies* 13:1, 2020, pp. 186–197. ISSN: 1939-1382, 2372-0050. DOI: 10.1109/TLT.2018.2879354.

35. A. R. Cano, B. Fernández-Manjón, and Á. J. García-Tejedor. "Downtown, a Subway Adventure: Using Learning Analytics to Improve the Development of a Learning Game for People with Intellectual Disabilities". In: *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*. 2016, pp. 125–129. DOI: 10.1109/ICALT.2016.46.

36. A. R. Cano, B. Fernández-Manjón, and Á. J. García-Tejedor. "Using game learning analytics for validating the design of a learning game for adults with intellectual disabilities". *British Journal of Educational Technology* 49:4, 2018, pp. 659–672. DOI: 10.1111/bjet.12632.

37. A. Cano, A. R. Masegosa, and S. Moral. "A Method for Integrating Expert Knowledge When Learning Bayesian Networks From Data". *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41:5, 2011, pp. 1382–1394. DOI: 10.1109/TSMCB.2011.2148197.

38. J. C. Cappelleri, J. J. Lundy, and R. D. Hays. "Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures". *Clinical therapeutics* 36:5, 2014, pp. 648–662. DOI: 10.1016/j.clinthera.2014.04.006.

39. A. M. Carvalho. "Scoring functions for learning Bayesian networks". *Inesc-id Tec. Rep* 12, 2009, pp. 1–48. URL: http://www.lx.it.pt/~asmc/pub/talks/09-TA/ta_pres.pdf.

40. D. Checa and A. Bustillo. "A review of immersive virtual reality serious games to enhance learning and training". *Multimed Tools Appl* 79, 2020, pp. 5501–5527. DOI: 10.1007/s11042-019-08348-9.

41. Y. Chen, K. Sherren, M. Smit, and K. Y. Lee. "Using social media images as data in social science research". *New Media & Society* 25:4, 2023, pp. 849–871. DOI: 10.1177/14614448211038761.

42. J. Cheng, D. A. Bell, and W. Liu. "An Algorithm for Bayesian Network Construction from Data". In: *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*. Vol. R1. Proceedings of Machine Learning Research. Reissued by PMLR on 30 March 2021. PMLR, 1997, pp. 83–90. URL: https://proceedings.mlr.press/r1/cheng97a.html.
43. L. Cheng, R. Guo, and H. Liu. "Robust Cyberbullying Detection with Causal Interpretation". In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW '19. Association for Computing Machinery, New York, NY, USA, 2019, pp. 169–175. ISBN: 9781450366755. DOI: 10.1145/3308560.3316503.
44. S. Çiftci. "Trends of Serious Games Research from 2007 to 2017: A Bibliometric Analysis." *Journal of Education and Training Studies* 6:2, 2018. ISSN-2324-805X., pp. 18–27. DOI: 10.11114/jets.v6i2.2840.
45. *Coding Apps*. LightBot Inc. Android and iOS Games. 2017. URL: https://lightbot.com.
46. G. F. Cooper and E. Herskovits. "A Bayesian method for the induction of probabilistic networks from data". *Machine Learning* 9:4, 1992, pp. 309–347. ISSN: 1573-0565. DOI: 10.1007/BF00994110.
47. *Corbyn Run*. Digital Liberties. Android game. 2019. URL: https://play.google.com/store/apps/details?id=com.digitalliberties.corbynrun.
48. A. Coutrot, E. Manley, S. Goodroe, C. Gahnstrom, G. Filomena, D. Yesiltepe, R. C. Dalton, J. M. Wiener, C. Hölscher, M. Hornberger, and H. J. Spiers. "Entropy of city street networks linked to future spatial navigation ability". *Nature* 604:7904, 2022, pp. 104–110. DOI: 10.1038/s41586-022-04486-7.
49. E. S. de Lima, B. Feijó, and A. L. Furtado. "Player behavior and personality modeling for interactive storytelling in games". *Entertainment Computing* 28, 2018, pp. 32–48. ISSN: 1875-9521. DOI: 10.1016/j.entcom.2018.08.003.
50. J. A. DeFalco, J. P. Rowe, L. Paquette, V. Georgoulas-Sherry, K. Brawner, B. W. Mott, R. S. Baker, and J. C. Lester. "Detecting and Addressing Frustration in a Serious Game for Military Training". *International Journal of Artificial Intelligence in Education* 28:2, 2018, pp. 152–193. ISSN: 1560-4306. DOI: 10.1007/s40593-017-0152-1.
51. A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data Via the EM Algorithm". *Journal of the Royal Statistical Society: Series B (Methodological)* 39:1, 1977, pp. 1–22. DOI: 10.1111/j.2517-6161.1977.tb01600.x.
52. D. Demszky, D. Yang, D. Yeager, et al. "Using large language models in psychology". *Nat Rev Psychol* 2, 2023, pp. 688–701. DOI: 10.1038/s44159-023-00241-5.
53. R.-J. Den Haan and M. C. Van der Voort. "On Evaluating Social Learning Outcomes of Serious Games to Collaboratively Address Sustainability Problems: A Literature Review". *Sustainability* 10:12, 2018. ISSN: 2071-1050. DOI: 10.3390/su10124529.
54. M. Denden, A. Tlili, F. Essalmi, and M. Jemni. "Implicit modeling of learners' personalities in a game-based learning environment using their gaming behaviors". *Smart Learning Environments* 5:1, 2018, p. 29. ISSN: 2196-7091. DOI: 10.1186/s40561-018-0078-6.
55. P. Dhariwal and A. Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

56. J. M. Digman. “Personality Structure: Emergence of the Five-Factor Model”. *Annual Review of Psychology* 41:1, 1990, pp. 417–440. DOI: 10.1146/annurev.ps.41.020190.002221.
57. *DragonBox & Kahoot! products*. Kahoot DragonBox AS. Fridtjof Nansens Plass 7, 0160 Oslo, Norway. Organization number: 922 693 021. 2019. URL: https://dragonbox.com/products.
58. M. Dreier, K. Wölfling, E. Duven, S. Giralt, M. Beutel, and K. Müller. “Free-to-play: About addicted Whales, at risk Dolphins and healthy Minnows. Monetarization design and Internet Gaming Disorder”. *Addictive Behaviors* 64, 2017, pp. 328–333. ISSN: 0306-4603. DOI: 10.1016/j.addbeh.2016.03.008.
59. Y. Y. Dyulicheva and A. O. Glazieva. “Game based learning with artificial intelligence and immersive technologies: an overview”. In: *Ceur Workshop Proceedings*. Vol. 3077. 2022, pp. 146–159. URL: http://ceur-ws.org/Vol-3077/paper05.pdf.
60. K. El Emam, L. Mosquera, and R. Hoptroff. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O’Reilly Media, 2020. ISBN: 9781492072744.
61. F. Elwert. “Graphical Causal Models”. In: *Handbook of Causal Analysis for Social Research*. Springer Netherlands, Dordrecht, 2013, pp. 245–273. ISBN: 978-94-007-6094-3. DOI: 10.1007/978-94-007-6094-3_13.
62. S. E. Embretson and S. P. Reise. *Item response theory*. Psychology Press, 2013. DOI: 10.4324/9781410605269.
63. S. E. Embretson and S. P. Reise. *Item response theory for psychologists*. Multivariate Applications Books Series. Lawrence Erlbaum Associates Publishers, 2000, pp. xi, 371–xi, 371. ISBN: 0-8058-2818-4. DOI: 10.4324/9781410605269.
64. *EteRNA*. Carnegie Mellon University and Stanford University. Web, Android, and iOS game. 2010. URL: https://eternagame.org/.
65. European Commission, J. R. Centre, B. Lobe, A. Velicu, E. Staksrud, S. Chaudron, and 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. ISBN: 978-92-76-29763-5. DOI: 10.2760/066196.
66. *EyeWire*. Sebastian Seung’s Lab at Princeton University. 2012. URL: https://eyewire.org/.
67. H. Farley. “Promoting self-efficacy in patients with chronic disease beyond traditional education: A literature review”. *Nursing Open* 7:1, 2020, pp. 30–41. DOI: 10.1002/nop2.382.
68. C. Ferguson, R. Lewis, C. Wilks, and R. Picard. “The Guardians: Designing a Game for Long-term Engagement with Mental Health Therapy”. In: *2021 IEEE Conference on Games (CoG)*. 2021, pp. 1–8. DOI: 10.1109/CoG52621.2021.9619026.
69. P. C. Ferreira, A. M. Veiga Simão, A. Paiva, C. Martinho, R. Prada, A. Ferreira, and F. Santos. “Exploring empathy in cyberbullying with serious games”. *Computers & Education* 166, 2021, p. 104155. ISSN: 0360-1315. DOI: 10.1016/j.compedu.2021.104155.
70. *Fight Against Opium*. Code To Inspire. 2018. URL: https://www.codetoinspire.org.
71. R. Fischer and Y. H. Poortinga. “Addressing Methodological Challenges in Culture-Comparative Research”. *Journal of Cross-Cultural Psychology* 49:5, 2018, pp. 691–712. DOI: 10.1177/0022022117738086.
72. T. M. Fleming, L. Bavin, K. Stasiak, E. Hermansson-Webb, S. N. Merry, C. Cheek, M. Lucassen, H. M. Lau, B. Pollmuller, and S. Hetrick. “Serious games and gamification for mental health: current status and promising directions”. *Frontiers in psychiatry* 7, 2017, p. 215. DOI: 10.3389/fpsyt.2016.00215.

73. S. Flood, N. A. Cradock-Henry, P. Blackett, and P. Edwards. “Adaptive and interactive climate futures: systematic review of ‘serious games’ for engagement and decision-making”. *Environmental Research Letters* 13:6, 2018, p. 063005. DOI: 10.1088/1748-9326/aac1c6.

74. *Foldit: Solve Puzzles for Science*. University of Washington. PC game. 2008. URL: https://fold.it/.

75. M. Freire, A. Serrano-Laguna, B. Manero, I. Martinez-Ortiz, P. Moreno-Ger, and B. Fernandez-Manjon. “Game Learning Analytics: Learning Analytics for Serious Games”. In: *Learning, Design, and Technology*. Springer Nature Switzerland AG, 2016, pp. 1–29. DOI: 10.1007/978-3-319-17727-4_21-1.

76. M. Freire, Á. Serrano-Laguna, B. M. Iglesias, I. Martínez-Ortiz, P. Moreno-Ger, and B. Fernández-Manjón. “Game Learning Analytics: Learning Analytics for Serious Games”. In: *Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy*. Springer International Publishing, Cham, 2016, pp. 1–29. ISBN: 978-3-319-17727-4. DOI: 10.1007/978-3-319-17727-4_21-1.

77. S. Fridenson-Hayo, S. Berggren, A. Lassalle, S. Tal, D. Pigat, N. Meir-Goren, H. O’Reilly, S. Ben-Zur, S. Bölte, S. Baron-Cohen, et al. “‘Emotiplay’: a serious game for learning about emotions in children with autism: results of a cross-cultural evaluation”. *European child & adolescent psychiatry* 26:8, 2017, pp. 979–992. DOI: 10.1007/s00787-017-0968-0.

78. *Garfield’s Count Me In*. Grendel Games. Blokhuisplein 40, 8911 LJ Leeuwarden, Netherlands. 2021. URL: https://www.garfieldscountmein.nl/.

79. D. Geiger, T. Verma, and J. Pearl. “d-Separation: From Theorems to Algorithms”. In: *Uncertainty in Artificial Intelligence*. Vol. 10. Machine Intelligence and Pattern Recognition. North-Holland, 1990, pp. 139–148. DOI: 10.1016/B978-0-444-88738-2.50018-X.

80. S. Geman and D. Geman. “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6:6, 1984, pp. 721–741. DOI: 10.1109/TPAMI.1984.4767596.

81. G. Gogoshin, S. Branciamore, and A. S. Rodin. “Synthetic data generation with probabilistic Bayesian Networks”. *Mathematical Biosciences and Engineering* 18:6, 2021, pp. 8603–8621. ISSN: 1551-0018. DOI: 10.3934/mbe.2021426.

82. M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. G. Clemente. “A Systematic Literature Review of Game-Based Assessment Studies: Trends and Challenges”. *IEEE Transactions on Learning Technologies* 16:4, 2023, pp. 500–515. DOI: 10.1109/TLT.2022.3226661.

83. I. Gorbanev, S. Agudelo-Londoño, R. A. González, A. Cortes, A. Pomares, V. Delgadillo, F. J. Yepes, and Ó. Muñoz. “A systematic review of serious games in medical education: quality of evidence and pedagogical strategy”. *Medical Education Online* 23:1, 2018, p. 1438718. ISSN: 1087-2981. DOI: 10.1080/10872981.2018.1438718.

84. S. Greenland, J. Pearl, and J. M. Robins. “Causal Diagrams for Epidemiologic Research”. *Epidemiology* 10:1, 1999, pp. 37–48. ISSN: 10443983. URL: http://www.jstor.org/stable/3702180 (visited on 02/20/2024).

85. J. Gripsrud. “The Impact of Video Games on Culture”. In: *Understanding media culture*. Bloomsbury Publishing, 2017. Chap. 10.4. DOI: 10.24926/8668.2601.

86. M. P. Grosz, J. M. Rohrer, and F. Thoemmes. “The Taboo Against Explicit Causal Inference in Nonexperimental Psychology”. *Perspectives on Psychological Science* 15:5, 2020. PMID: 32727292, pp. 1243–1255. DOI: 10.1177/1745691620921521.

87. S. Grund, O. Lüdtke, and A. Robitzsch. "Using synthetic data to improve the reproducibility of statistical results in psychological research." *Psychological Methods*, 2022. DOI: 10.1037/met0000526.

88. L. Gultchin. "Casual and trustworthy machine learning: methods and applications". PhD thesis. University of Oxford, 2023.

89. R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu. "A Survey of Learning Causality with Data: Problems and Methods". *ACM Comput. Surv.* 53:4, 2020. ISSN: 0360-0300. DOI: 10.1145/3397269.

90. T. Gupta, W. Gong, C. Ma, N. Pawlowski, A. Hilmkil, M. Scetbon, A. Famoti, A. J. Llorens, J. Gao, S. Bauer, D. Kragic, B. Schölkopf, and C. Zhang. "The Essential Role of Causality in Foundation World Models for Embodied AI". *arXiv preprint arXiv:2402.06665*, 2024. DOI: 10.48550/arXiv.2402.06665.

91. Z. Halim, M. Atif, A. Rashid, and C. A. Edwin. "Profiling Players Using Real-World Datasets: Clustering the Data and Correlating the Results with the Big-Five Personality Traits". *IEEE Transactions on Affective Computing* 10:4, 2019, pp. 568–584. DOI: 10.1109/TAFFC.2017.2751602.

92. G. Haoran, E. Bazakidi, and N. Zary. "Serious games in health professions education: review of trends and learning efficacy". *Yearbook of medical informatics* 28:1, 2019, p. 240. DOI: 10.1055%2Fs-0039-1677904.

93. J. K. Hartshorne, J. B. Tenenbaum, and S. Pinker. "A critical period for second language acquisition: Evidence from 2/3 million English speakers". *Cognition* 177, 2018, pp. 263–277. DOI: 10.1016/j.cognition.2018.04.007.

94. H. van Hasselt, A. Guez, and D. Silver. "Deep Reinforcement Learning with Double Q-Learning". *Proceedings of the AAAI Conference on Artificial Intelligence* 30:1, 2016. DOI: 10.1609/aaai.v30i1.10295.

95. D. Heckerman, D. Geiger, and D. M. Chickering. "Learning Bayesian networks: The combination of knowledge and statistical data". *Machine Learning* 20:3, 1995, pp. 197–243. ISSN: 1573-0565. DOI: 10.1007/BF00994016.

96. K. Hellfeldt, L. López-Romero, and H. Andershed. "Cyberbullying and Psychological Well-being in Young Adolescence: The Potential Protective Mediation Effects of Social Support from Family, Friends, and Teachers". *International Journal of Environmental Research and Public Health* 17:1, 2020, p. 45. DOI: 10.3390/ijerph17010045.

97. J. Henares-Montiel, V. Benítez-Hidalgo, I. Ruiz-Pérez, G. Pastor-Moreno, and M. Rodríguez-Barranco. "Cyberbullying and Associated Factors in Member Countries of the European Union: A Systematic Review and Meta-Analysis of Studies with Representative Population Samples". *International Journal of Environmental Research and Public Health* 19:12, 2022. ISSN: 1660-4601. DOI: 10.3390/ijerph19127364.

98. M. Hendrix, T. Bellamy-Wood, S. McKay, V. Bloom, and I. Dunwell. "Implementing Adaptive Game Difficulty Balancing in Serious Games". *IEEE Transactions on Games* 11:4, 2019, pp. 320–327. ISSN: 2475-1510. DOI: 10.1109/tg.2018.2791019.

99. M. Hendrix, A. Al-Sherbaz, and B. Victoria. "Game based cyber security training: are serious games suitable for cyber security training?" *International Journal of Serious Games* 3:1, 2016. ISSN: 2384-8766. DOI: 10.17083/ijsg.v3i1.107.

100. A. B. Hernández-Lara, A. Perera-Lluna, and E. Serradell-López. "Applying learning analytics to students' interaction in business simulation games. The usefulness of learning analytics to know what students really learn". *Computers in Human Behavior* 92, 2019, pp. 600–612. ISSN: 0747-5632. DOI: 10.1016/j.chb.2018.03.001.

101. D. Hicks, M. Eagle, E. Rowe, J. Asbell-Clarke, T. Edwards, and T. Barnes. "Using Game Analytics to Evaluate Puzzle Design and Level Progression in a Serious Game". In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. LAK '16. Association for Computing Machinery, 2016, pp. 440–448. ISBN: 9781450341905. DOI: 10.1145/2883851.2883953.

102. N. Hocine, A. Gouaïch, S. A. Cerri, D. Mottet, J. Froger, and I. Laffont. "Adaptation in serious games for upper-limb rehabilitation: an approach to improve training outcomes". *User Modeling and User-Adapted Interaction* 25:1, 2015, pp. 65–98. ISSN: 1573-1391. DOI: 10.1007/s11257-015-9154-6.

103. J. M. Hofman, D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, A. Vespignani, and T. Yarkoni. "Integrating explanation and prediction in computational social science". *Nature* 595:7866, 2021, pp. 181–188. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03659-0.

104. D. Hooshyar, M. Yousefi, and H. Lim. "Data-Driven Approaches to Game Player Modeling: A Systematic Literature Review". *ACM Comput. Surv.* 50:6, 2018. ISSN: 0360-0300. DOI: 10.1145/3145814.

105. B. Howe, J. Stoyanovich, H. Ping, B. Herman, and M. Gee. "Synthetic Data for Social Good". *arXiv preprint arXiv:1710.08874*, 2017. DOI: 10.48550/ARXIV.1710.08874.

106. P. Hünermund and E. Bareinboim. "Causal Inference and Data Fusion in Econometrics". *arXiv preprint arXiv:1912.09104*, 2023. DOI: 10.48550/arXiv.1912.09104.

107. Z. Hussain, M. Binz, R. Mata, and D. U. Wulff. *A tutorial on open-source large language models for behavioral science*. Preprint in PsyArXiv. 2023. DOI: 10.31234/osf.io/f7stn.

108. A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne. "Imitation Learning: A Survey of Learning Methods". *ACM Comput. Surv.* 50:2, 2017. ISSN: 0360-0300. DOI: 10.1145/3054912.

109. Z. Islam, M. Abdel-Aty, Q. Cai, and J. Yuan. "Crash data augmentation using variational autoencoder". *Accident Analysis & Prevention* 151, 2021, p. 105950. ISSN: 0001-4575. DOI: 10.1016/j.aap.2020.105950.

110. B. K. Iwana and S. Uchida. "An empirical survey of data augmentation for time series classification with neural networks". *Plos one* 16:7, 2021, e0254841. DOI: 10.1371/journal.pone.0254841.

111. H. K. Jabbar and R. Z. Khan. "Methods to Avoid Over-Fitting and Under-Fitting in Supervised Machine Learning (Comparative Study)". In: *Computer Science, Communication and Instrumentation Devices*. AET 2014. Research Publishing Services, 2014. DOI: 10.3850/978-981-09-5247-1_017.

112. H. Jeffreys. *The theory of probability*. OuP Oxford, 1998. ISBN: 9780198503682.

113. T. Jiralerspong, X. Chen, Y. More, V. Shah, and Y. Bengio. "Efficient Causal Graph Discovery Using Large Language Models". *arXiv preprint arXiv:2402.01207*, 2024. DOI: 10.48550/arXiv.2402.01207.

114. J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller. "Synthetic Data – what, why and how?" *arXiv preprint arXiv:2205.03257*, 2022. DOI: 10.48550/ARXIV.2205.03257.

115. M. Julia Flores, A. E. Nicholson, A. Brunskill, K. B. Korb, and S. Mascaro. "Incorporating expert knowledge when learning Bayesian network structure: A medical case study". *Artificial Intelligence in Medicine* 53:3, 2011, pp. 181–204. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2011.08.004.

116. J. Kang, M. Liu, and W. Qu. "Using gameplay data to examine learning behavior patterns in a serious game". *Computers in Human Behavior* 72, 2017, pp. 757–770. ISSN: 0747-5632. DOI: 10.1016/j.chb.2016.09.062.

117. R. Karamians, R. Proffitt, D. Kline, and L. V. Gauthier. "Effectiveness of Virtual Reality- and Gaming-Based Interventions for Upper Extremity Rehabilitation Poststroke: A Meta-analysis". *Archives of Physical Medicine and Rehabilitation* 101:5, 2020, pp. 885–896. ISSN: 0003-9993. DOI: 10.1016/j.apmr.2019.10.195.

118. P. M. Kato and S. de Klerk. "Serious Games for Assessment: Welcome to the Jungle". *Journal of Applied Testing Technology* 18:S1, 2017. URL: http://www.jattjournal.net/index.php/atp/article/view/118669.

119. A. Katsaounidou, L. Vrysis, R. Kotsakis, C. Dimoulas, and A. Veglis. "MAthE the Game: A Serious Game for Education and Training in News Verification". *Education Sciences* 9:2, 2019. ISSN: 2227-7102. DOI: 10.3390/educsci9020155.

120. D. Kaur, M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta, and N. Markuzon. "Application of Bayesian networks to generate synthetic health data". *Journal of the American Medical Informatics Association* 28:4, 2020, pp. 801–811. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa303.

121. A. Kawrykow, G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, E. Zarour, L. Sarmenta, M. Blanchette, and J. Waldispühl. "Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment". en. *PLoS ONE* 7:3, 2012. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0031362.

122. B. Keles, N. McCrae, and A. Grealish. "A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents". *International Journal of Adolescence and Youth* 25:1, 2020, pp. 79–93. DOI: 10.1080/02673843.2019.1590851.

123. K. Khowaja and S. S. Salim. "Serious game for children with autism to learn vocabulary: an experimental evaluation". *International journal of human–computer interaction* 35:1, 2019, pp. 1–26. DOI: 10.1080/10447318.2017.1420006.

124. K. Kiili, K. Moeller, and M. Ninaus. "Evaluating the effectiveness of a game-based rational number training - In-game metrics as learning indicators". *Computers & Education* 120, 2018, pp. 13–28. ISSN: 0360-1315. DOI: 10.1016/j.compedu.2018.01.012.

125. R. Kim, M. Kleiman-Weiner, A. Abeliuk, E. Awad, S. Dsouza, J. B. Tenenbaum, and I. Rahwan. "A Computational Model of Commonsense Moral Decision Making". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '18. Association for Computing Machinery, New York, NY, USA, 2018, pp. 197–203. ISBN: 9781450360128. DOI: 10.1145/3278721.3278770.

126. Y. J. Kim and D. Ifenthaler. "Game-Based Assessment: The Past Ten Years and Moving Forward". In: *Game-Based Assessment Revisited*. Springer International Publishing, Cham, 2019, pp. 3–11. ISBN: 978-3-030-15569-8. DOI: 10.1007/978-3-030-15569-8_1.

127. Y. J. Kim and J. A. Ruipérez-Valiente. "Data-Driven Game Design: The Case of Difficulty in Educational Games". In: *Addressing Global Challenges and Quality Education*. Springer International Publishing, Cham, 2020, pp. 449–454. ISBN: 978-3-030-57717-9. DOI: 10.1007/978-3-030-57717-9_43.

128. N. K. Kitson, A. C. Constantinou, Z. Guo, Y. Liu, and K. Chobtham. "A survey of Bayesian Network structure learning". *Artificial Intelligence Review* 56:8, 2023, pp. 8721–8814. ISSN: 1573-7462. DOI: 10.1007/s10462-022-10351-w.

129. E. Kıcıman, R. Ness, A. Sharma, and C. Tan. "Causal Reasoning and Large Language Models: Opening a New Frontier for Causality". *arXiv preprint arXiv:2305.00050*, 2023. DOI: 10.48550/arXiv.2305.00050.

130. J. Koivisto and J. Hamari. "The rise of motivational information systems: A review of gamification research". *International Journal of Information Management* 45, 2019, pp. 191–210. ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2018.10.013.

131. D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. ISBN: 9780262013192.

132. K.-F. Kowalewski, J. D. Hendrie, M. W. Schmidt, T. Proctor, S. Paul, C. R. Garrow, H. G. Kenngott, B. P. Müller-Stich, and F. Nickel. "Validation of the mobile serious game application Touch Surgery™ for cognitive training and assessment of laparoscopic cholecystectomy". *Surgical Endoscopy* 31:10, 2017, pp. 4058–4066. ISSN: 1432-2218. DOI: 10.1007/s00464-017-5452-x.

133. C.-E. Kuo, G.-T. Chen, and P.-Y. Liao. "An EEG spectrogram-based automatic sleep stage scoring method via data augmentation, ensemble convolution neural network, and expert knowledge". *Biomedical Signal Processing and Control* 70, 2021, p. 102981. ISSN: 1746-8094. DOI: 10.1016/j.bspc.2021.102981.

134. I. Kuo. *L'Oreal uses Serious Games for Employee Recruitment*. Gamification Blog. 2014. URL: https://www.gamification.co/2014/10/01/loreal-uses-serious-games-employee-recruitment/.

135. A. K. Lampinen, S. C. Y. Chan, I. Dasgupta, A. J. Nam, and J. X. Wang. "Passive learning of active causal strategies in agents and language models". *arXiv preprint arXiv:2305.16183*, 2023. DOI: 10.48550/arXiv.2305.16183.

136. R. N. Landers and D. R. Sanchez. "Game-based, gamified, and gamefully designed assessments for employee selection: Definitions, distinctions, design, and validation". *International Journal of Selection and Assessment* 30:1, 2022, pp. 1–13. DOI: 10.1111/ijsa.12376.

137. G. van Lankveld, P. Spronck, J. van den Herik, and A. Arntz. "Games as personality profiling tools". In: *2011 IEEE Conference on Computational Intelligence and Games (CIG'11)*. 2011, pp. 197–202. DOI: 10.1109/CIG.2011.6032007.

138. K. Larson. "Serious Games and Gamification in the Corporate Training Environment: a Literature Review". *TechTrends* 64:2, 2020, pp. 319–328. ISSN: 1559-7075. DOI: 10.1007/s11528-019-00446-7.

139. S. L. Lauritzen. "The EM algorithm for graphical association models with missing data". *Computational Statistics & Data Analysis* 19:2, 1995, pp. 191–201. ISSN: 0167-9473. DOI: 10.1016/0167-9473(93)E0056-A.

140. D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. V. Alstyne. "Computational Social Science". *Science* 323:5915, 2009, pp. 721–723. DOI: 10.1126/science.1167742.

141. D. M. J. Lazer, A. Pentland, D. J. Watts, S. Aral, S. Athey, N. Contractor, D. Freelon, S. Gonzalez-Bailon, G. King, H. Margetts, A. Nelson, M. J. Salganik, M. Strohmaier, A. Vespignani, and C. Wagner. "Computational social science: Obstacles and opportunities". *Science* 369:6507, 2020, pp. 1060–1062. DOI: 10.1126/science.aaz8170.

142. G. Lederrey, T. Hillel, and M. Bierlaire. "DATGAN: Integrating expert knowledge into deep learning for synthetic tabular data". *arXiv preprint arXiv:2203.03489*, 2022. DOI: 10.48550/ARXIV.2203.03489.

143. K. Leduc, L. Conway, C. Gomez-Garibello, and V. Talwar. "The influence of participant role, gender, and age in elementary and high-school children's moral justifications of cyberbullying behaviors". *Computers in Human Behavior* 83, 2018, pp. 215–220. ISSN: 0747-5632. DOI: 10.1016/j.chb.2018.01.044.

144. M. Lee, M. Kaur, V. Shaker, A. Yee, R. Sham, and C. Siau. "Cyberbullying, Social Media Addiction and Associations with Depression, Anxiety, and Stress among Medical Students in Malaysia". *International Journal of Environmental Research and Public Health* 20:4, 2023, p. 3136. DOI: 10.3390/ijerph20043136.

145. *Left 4 Dead*. Valve Corporation. 2008. URL: https://www.valvesoftware.com.

146. H. Lei, W. Mao, C. M. Cheong, Y. Wen, Y. Cui, and Z. Cai. "The relationship between self-esteem and cyberbullying: A meta-analysis of children and youth students". *Current Psychology* 39:3, 2020, pp. 830–842. ISSN: 1936-4733. DOI: 10.1007/s12144-019-00407-6.

147. R. Levy. "Dynamic Bayesian Network Modeling of Game-Based Diagnostic Assessments". *Multivariate Behavioral Research* 54:6, 2019. PMID: 30942094, pp. 771–794. DOI: 10.1080/00273171.2019.1590794.

148. J. Li, Y. Jin, S. Xu, A. Wilson, C. Chen, X. Luo, Y. Liu, X. Ling, X. Sun, and Y. Wang. "Effects of Bullying on Anxiety, Depression, and Posttraumatic Stress Disorder Among Sexual Minority Youths: Network Analysis". *JMIR Public Health and Surveillance* 9, 2023, e47233. ISSN: 2369-2960. DOI: 10.2196/47233.

149. B. Lin, G. Cecchi, D. Bouneffouf, J. Reinen, and I. Rish. "A story of two streams: Reinforcement learning models from human behavior and neuropsychiatry". *arXiv preprint arXiv:1906.11286*, 2019. DOI: 10.48550/arXiv.1906.11286.

150. M. Liu, J. Kang, S. Liu, W. Zou, and J. Hodson. "Learning Analytics as an Assessment Tool in Serious Games: A Review of Literature". In: *Serious Games and Edutainment Applications : Volume II*. Springer International Publishing, Cham, 2017, pp. 537–563. ISBN: 978-3-319-51645-5. DOI: 10.1007/978-3-319-51645-5_24.

151. T. Liu, L. Ungar, and K. Kording. "Quantifying causality in data science with quasi-experiments". en. *Nature Computational Science* 1:1, 2021. Number: 1 Publisher: Nature Publishing Group, pp. 24–32. ISSN: 2662-8457. DOI: 10.1038/s43588-020-00005-8.

152. G. Livazović and E. Ham. "Cyberbullying and emotional distress in adolescents: the importance of family, peers and school". *Heliyon* 5:6, 2019, e01992. DOI: 10.1016/j.heliyon.2019.e01992.

153. B. Long, J. Simson, A. Buxó-Lugo, D. G. Watson, and S. A. Mehr. "How games can make behavioural science better". *Nature* 613:7944, 2023, pp. 433–436. DOI: 10.1038/d41586-023-00065-6.

154. S. Lopes, P. Magalhães, A. Pereira, J. Martins, C. Magalhães, E. Chaleta, and P. Rosário. "Games used with serious purposes: a systematic review of interventions in patients with cerebral palsy". *Frontiers in psychology* 9, 2018, p. 1712.

155. G. López, N. Bueno, M. Castro, M. Reneses, J. Pérez, M. Riberas, M. Á. Campana, M. Vega-Barbas, S. Solera-Cotanilla, L. Bastida, A. Moya, R. Fernández, V. Vázquez, G. Zango, and P. Vicente. "The H2020 project RAYUELA: A fun way to fight cybercrime". In: *Nº 34: Investigación en Ciberseguridad*. Ediciones de la Universidad de Castilla-La Mancha, 2021. DOI: 10.18239/jornadas_2021.34.27.

156. S. Luma-Osmani, F. Ismaili, P. Pathak, and X. Zenuni. "Identifying Causal Structures from Cyberstalking: Behaviors Severity and Association". *Journal of Communications Software and Systems* 18:1, 2022, pp. 1–8. ISSN: 1846-6079. DOI: 10.24138/jcomss-2021-0139.

157. L. Ma and B. Sun. "Machine learning and AI in marketing – Connecting computing power to human insights". *International Journal of Research in Marketing* 37:3, 2020, pp. 481–504. ISSN: 0167-8116. DOI: 10.1016/j.ijresmar.2020.04.005.

158. D. Makowski, M. Ben-Shachar, and D. Lüdecke. "bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework". *Journal of Open Source Software* 4:40, 2019, p. 1541. DOI: 10.21105/joss.01541.

159. "Machine Learning and Evidence-Based Medicine". *Annals of Internal Medicine* 169:1, 2018, pp. 44–46. DOI: 10.7326/M18-0115.

160. *MalariaSpot Bubbles*. Universidad Politécnica de Madrid (UPM) and Campus of International Excellence. 2016. URL: https://malariaspot.org/.

161. G. Maldonado and S. Greenland. "Estimating causal effects". *International Journal of Epidemiology* 31:2, 2002, pp. 422–429. ISSN: 0300-5771. DOI: 10.1093/ije/31.2.422.

162. B. G. Marcot and T. D. Penman. "Advances in Bayesian network modelling: Integration of modelling technologies". *Environmental Modelling & Software* 111, 2019, pp. 386–393. ISSN: 1364-8152. DOI: 10.1016/j.envsoft.2018.09.016.

163. M. P. Martín. *Minecraft como herramienta para el Ayuntamiento de Madrid*. Blog IGN España. 2017. URL: https://es.ign.com/minecraft/120226/news/minecraft-como-herramienta-para-el-ayuntamiento-de-madrid.

164. A. R. Masegosa and S. Moral. "An interactive approach for Bayesian network learning using domain/expert knowledge". *International Journal of Approximate Reasoning* 54, 8 2013, pp. 1168–1181. ISSN: 0888-613X. DOI: 10.1016/j.ijar.2013.03.009.

165. J.-L. McCord, J. L. Harman, and J. Purl. "Game-like personality testing: An emerging mode of personality assessment". *Personality and Individual Differences* 143, 2019, pp. 95–102. ISSN: 0191-8869. DOI: 10.1016/j.paid.2019.02.017.

166. R. McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2020. ISBN: 9780367139919.

167. H. A. Meijer, M. Graafland, J. C. Goslings, and M. P. Schijven. "Systematic Review on the Effects of Serious Games and Wearable Technology Used in Rehabilitation of Patients With Traumatic Bone and Soft Tissue Injuries". *Archives of Physical Medicine and Rehabilitation* 99:9, 2018, pp. 1890–1899. ISSN: 0003-9993. DOI: 10.1016/j.apmr.2017.10.018.

168. *Microsoft Flight Simulator for Windows 10 | Xbox*. Asobo Studio. 2023. URL: https://www.flightsimulator.com.

169. *Minecraft: Education Edition*. Mojang AB. TM Microsoft Corporation. Söder Mälarstrand 43, SE-11825, Stockholm, Sweden. Organization number: 556819-2388. 2018. URL: https://education.minecraft.net.

170. M. S. El-Nasr, T.-H. D. Nguyen, A. Canossa, and A. Drachen. *Game Data Science*. Oxford University Press, 2021. ISBN: 9780192897879. DOI: 10.1093/oso/9780192897879.001.0001.

171. B. Neal. *Introduction to causal inference*. Course Lecture Notes (draft). 2020.

172. T. T. H. Nguyen, D. Ishmatova, T. Tapanainen, T. N. Liukkonen, N. Katajapuu, T. Makila, and M. Luimula. "Impact of serious games on health and well-being of elderly: a systematic review". In: *Proceedings of the 50th Hawaii International Conference on System Sciences*. 2017. DOI: 10.24251/HICSS.2017.447.

173. H. Ning, R. Li, X. Ye, Y. Zhang, and L. Liu. "A Review on Serious Games for Dementia Care in Ageing Societies". *IEEE Journal of Translational Engineering in Health and Medicine* 8, 2020, pp. 1–11. DOI: 10.1109/JTEHM.2020.2998055.

174. T. R. d. Oliveira, T. F. Martinelli, B. P. Bello, J. D. Batista, M. M. d. Silva, B. B. Rodrigues, R. A. N. Spinassé, R. V. Andreão, M. Mestria, and M. Q. Schimidt. "Virtual Reality System for Industrial Motor Maintenance Training". In: *2020 22nd Symposium on Virtual and Augmented Reality (SVR)*. 2020, pp. 119–128. DOI: 10.1109/SVR51698.2020.00031.

175. S.-V. Oprea, A. Bâra, D. Preotescu, R. A. Bologa, and L. Coroianu. "A Trading Simulator Model for the Wholesale Electricity Market". *IEEE Access* 8, 2020, pp. 184210–184230. DOI: 10.1109/ACCESS.2020.3029291.

176. B. Osiński, A. Jakubowski, P. Zięcina, P. Miłoś, C. Galias, S. Homoceanu, and 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. DOI: 10.1109/ICRA40945.2020.9196730.

177. J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. "Generative Agents: Interactive Simulacra of Human Behavior". *arXiv preprint arXiv:2304.03442*, 2023. DOI: 10.48550/arXiv.2304.03442.

178. J. Pearl. "Causal diagrams for empirical research". *Biometrika* 82:4, 1995, pp. 669–688. DOI: 10.1093/biomet/82.4.669.

179. J. Pearl. *Causality*. Cambridge university press, 2009. ISBN: 9780521895606. DOI: 10.1017/CBO9780511803161.

180. J. Pearl. "Direct and Indirect Effects". In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 1st ed. Association for Computing Machinery, 2022, pp. 373–392. ISBN: 9781450395861. DOI: 10.1145/3501714.3501736.

181. J. Pearl. "The Do-Calculus Revisited". *arXiv preprint arXiv:1210.4852*, 2012. DOI: 10.48550/arXiv.1210.4852.

182. J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. ISBN: 978-1-119-18684-7.

183. J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018. ISBN: 9780141982410.

184. Z. Peddycord-Liu, C. Cody, S. Kessler, T. Barnes, C. F. Lynch, and T. Rutherford. "Using Serious Game Analytics to Inform Digital Curricular Sequencing: What Math Objective Should Students Play Next?" In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '17. Association for Computing Machinery, 2017, pp. 195–204. ISBN: 9781450348980. DOI: 10.1145/3116595.3116620.

185. J. W. Pellegrino, N. Chudowsky, R. Glaser, and N. R. Council. *Knowing what students know: The science and design of educational assessment*. National Academies Press, 2001. ISBN: 978-0-309-29322-8. DOI: 10.17226/10019.

186. J. Pérez, P. Arroba, and J. M. Moya. "Data augmentation through multivariate scenario forecasting in Data Centers using Generative Adversarial Networks". *Applied Intelligence* 53:2, 2023, pp. 1469–1486. ISSN: 1573-7497. DOI: 10.1007/s10489-022-03557-6.

187. J. Pérez, G. A. V. Castilla, G. López, M. Castro, M. R. Botija, M. Riberas-Gutiérrez, and N. Bueno-Guerra. *[Dataset] RAYUELA - Open Data - Data collected through a serious game created to identify patterns and profiles of young potential victims/perpetrators of cybercrimes.* Zenodo. 2024. DOI: 10.5281/zenodo.10604760.

188. J. Pérez, M. Castro, E. Awad, and G. López. "Generation of probabilistic synthetic data for serious games: A case study on cyberbullying". *Knowledge-Based Systems*, 2024, p. 111440. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2024.111440.

189. J. Pérez, M. Castro, and G. López. "Serious Games and AI: Challenges and Opportunities for Computational Social Science". *IEEE Access* 11, 2023, pp. 62051–62061. DOI: 10.1109/ACCESS.2023.3286695.

190. S. Pérez de Viñaspre, D. Díaz, and E. Toledano. *III Estudio sobre acoso escolar y ciberbullying según los afectados: informe del teléfono ANAR*. Technical report. Fundación ANAR and Fundación Mutua Madrileña, 2018. URL: https://www.observatoriodelainfancia.es/oia/esp/descargar.aspx?id=5621&tipo=documento.

191. J. Petrillo, S. J. Cano, L. D. McLeod, and 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 in Health* 18:1, 2015, pp. 25–34. ISSN: 1098-3015. DOI: 10.1016/j.jval.2014.10.005.

192. O. Petrovic, D. L. D. Duarte, S. Storms, and W. Herfs. "Towards Knowledge-based Generation of Synthetic Data by Taxonomizing Expert Knowledge in Production". *Intelligent Human Systems Integration (IHSI 2023): Integrating People and Intelligent Systems* 69:69, 2023. DOI: 10.54941/ahfe1002915.

193. *Political Misinformation Game: Harmony Square*. University of Cambridge. 2021. URL: https://harmonysquare.game/.

194. *Quantum Moves 2*. Games ScienceAtHome, Department of Management at Aarhus University. 2012. URL: https://www.scienceathome.org/games/quantum-moves-2/.

195. D. S. Quintana. "A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation". *eLife* 9, 2020. DOI: 10.7554/elife.53275.

196. W. S. Ravyse, A. Seugnet Blignaut, V. Leendertz, and A. Woolner. "Success factors for serious games to enhance learning: a systematic review". en. *Virtual Reality* 21:1, 2017, pp. 31–58. ISSN: 1359-4338, 1434-9957. DOI: 10.1007/s10055-016-0298-4.

197. M. Reneses, M. Riberas, N. Bueno, A. Gómez, B. Heylen, E. Andreotti, J. Op den Kelder, L. Verbraeken, and I. Borarosova. *Open Report on Interview Results*. Technical report. H2020 RAYUELA, 2022. URL: https://www.rayuela-h2020.eu/wp-content/uploads/2022/10/Attachment_0.pdf.

198. M. Reneses, M. Riberas, N. Bueno, B. Heylen, E. Andreotti, and J. Op den Kelder. *Open Report on Case Study Results*. Technical report. H2020 RAYUELA, 2022. URL: https://www.rayuela-h2020.eu/wp-content/uploads/2022/10/Attachment_0-1.pdf.

199. M. Reneses, M. Riberas, A. Gómez, N. Bueno, B. Heylen, and J. Ginter. *Open Report on Victim and Offender Profile Description Report*. Technical report. H2020 RAYUELA, 2022. URL: https://www.rayuela-h2020.eu/wp-content/uploads/2022/10/Attachment_0-2.pdf.

200. J. G. Richens, C. M. Lee, and S. Johri. "Improving the accuracy of medical diagnosis with causal machine learning". *Nature Communications* 11:1, 2020, p. 3923. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17419-7.

201. R. Rogers. *How video games impact players: The pitfalls and benefits of a gaming society*. Lexington Books, 2016. ISBN: 978-1-4985-1309-8.

202. J. M. Rohrer. "Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data". *Advances in Methods and Practices in Psychological Science* 1:1, 2018, pp. 27–42. DOI: 10.1177/2515245917745629.

203. J. M. Rohrer. "Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data". *Advances in Methods and Practices in Psychological Science* 1:1, 2018, pp. 27–42. DOI: 10.1177/2515245917745629.

204. M. Rosenberg. "Rosenberg self-esteem scale". 61:52, 1965. Acceptance and commitment therapy. Measures package, p. 18. DOI: 10.1037/t01038-000.

205. S. Ruíz. *Darfur is dying*. Take Action Games. 2006. URL: https://susanaruiz.org/takeactiongames.

206. M. J. Salganik. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019. ISBN: 9780691196107.

207. M. J. Salganik. *Bit by bit: Social research in the digital age*. Ch. 6. Princeton University Press, 2019. ISBN: 9780691196107.

208. M. Salvagno, F. S. Taccone, and A. G. Gerli. "Artificial intelligence hallucinations". *Critical Care* 27:1, 2023, p. 180. ISSN: 1364-8535. DOI: 10.1186/s13054-023-04473-y.

209. J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. "Probabilistic programming in Python using PyMC3". *PeerJ Computer Science* 2, 2016, e55. DOI: 10.7717/peerj-cs.55.

210. A. B. Samčović. "Serious games in military applications". *Vojnotehnički glasnik* 66:3, 2018, pp. 597–613. DOI: 10.5937/VOJTEHG66-16367.

211. F. Samejima. "Estimation of latent ability using a response pattern of graded scores". *Psychometrika* 34:1, 1969, pp. 1–97. ISSN: 1860-0980. DOI: 10.1007/BF03372160.

212. B. Schölkopf. "Causality for Machine Learning". In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 1st ed. Association for Computing Machinery, New York, NY, USA, 2022, pp. 765–804. ISBN: 9781450395861. DOI: 10.1145/3501714.3501755.

213. B. Schouten. "Playful empowerment, the role of game design innovation in participatory citizenship." In: *Serious Games: Second Joint International Conference, JCSG 2016*. Springer. 2016, pp. 1–3. ISBN: 978-3319458403.

214. *Serious Games for Enhancing Law Enforcement Agencies: From Virtual Reality to Augmented Reality*. Springer International Publishing, 2019. ISBN: 9783030299262. DOI: 10.1007/978-3-030-29926-2.

215. C. Shalizi. "Advanced data analysis from an elementary point of view", 2013. Carnegie Mellon University. URL: https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch22.pdf.

216. C. E. Shannon. "A mathematical theory of communication". *The Bell System Technical Journal* 27:3, 1948, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

217. N. Sharifzadeh, H. Kharrazi, E. Nazari, H. Tabesh, M. E. Khodabandeh, S. Heidari, and M. Tara. "Health education serious games targeting health care providers, patients, and public health users: scoping review". *JMIR serious games* 8:1, 2020, e13459. DOI: 10.2196%2F13459.

218. X. Shen et al. "Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology". *Scientific Reports* 10:1, 2020, p. 2975. ISSN: 2045-2322. DOI: 10.1038/s41598-020-59669-x.

219. G. Shi, J. Wang, Y. Qiang, X. Yang, J. Zhao, R. Hao, W. Yang, Q. Du, and N. G.-F. Kazihise. "Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification". *Computer Methods and Programs in Biomedicine* 196, 2020, p. 105611. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2020.105611.

220. C. Shorten and T. M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning". *Journal of Big Data* 6:1, 2019, p. 60. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0197-0.

221. C. Shorten, T. M. Khoshgoftaar, and B. Furht. “Text Data Augmentation for Deep Learning”. *Journal of Big Data* 8:1, 2021, p. 101. ISSN: 2196-1115. DOI: 10.1186/s40537-021-00492-0.

222. I. Shrier and R. W. Platt. “Reducing bias through directed acyclic graphs”. *BMC Medical Research Methodology* 8:1, 2008, p. 70. ISSN: 1471-2288. DOI: 10.1186/1471-2288-8-70.

223. I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson. “The Curse of Recursion: Training on Generated Data Makes Models Forget”. *arXiv preprint arXiv:2305.17493*, 2023. DOI: 10.48550/ARXIV.2305.17493.

224. V. Shute and L. Wang. “Assessing and Supporting Hard-to-Measure Constructs in Video Games”. In: *The Handbook of Cognition and Assessment*. John Wiley & Sons, Inc., 2016, pp. 535–562. DOI: 10.1002/9781118956588.ch22.

225. J. F. Sigurdson, A. M. Undheim, J. L. Wallander, S. Lydersen, and A. M. Sund. “The long-term effects of being bullied or a bully in adolescence on externalizing and internalizing mental health problems in adulthood”. *Child and Adolescent Psychiatry and Mental Health* 9:1, 2015. DOI: 10.1186/s13034-015-0075-2.

226. *Silent Hill Shattered Memories*. Climax Group & KONAMI. 2009.

227. D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play”. *Science* 362:6419, 2018, pp. 1140–1144. DOI: 10.1126/science.aar6404.

228. K. Sitnik-Warchulska, Z. Wajda, B. Wojciechowski, and B. Izydorczyk. “The Risk of Bullying and Probability of Help-Seeking Behaviors in School Children: A Bayesian Network Analysis”. *Frontiers in Psychiatry* 12, 2021. ISSN: 1664-0640. DOI: 10.3389/fpsyt.2021.640927.

229. D. Smahel, H. Machackova, G. Mascheroni, L. Dedkova, E. Staksrud, K. Ólafsson, S. Livingstone, and U. Hasebrink. *EU Kids Online 2020: Survey results from 19 countries*. Technical report. ISSN: 2045-256X. EU Kids Online, 2020. URL: https://www.eukidsonline.ch/files/Eu-kids-online-2020-international-report.pdf.

230. P. K. Smith, L. López-Castro, S. Robinson, and A. Görzig. “Consistency of gender differences in bullying in cross-cultural surveys”. *Aggression and Violent Behavior* 45, 2019. Bullying and cyberbullying: Protective factors and effective interventions, pp. 33–40. ISSN: 1359-1789. DOI: 10.1016/j.avb.2018.04.006.

231. P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000. ISBN: 9780262284158. DOI: 10.7551/mitpress/1754.001.0001.

232. M. Stanitsas, K. Kirytopoulos, and E. Vareilles. “Facilitating sustainability transition through serious games: A systematic literature review”. *Journal of cleaner production* 208, 2019, pp. 924–936. DOI: 10.1016/j.jclepro.2018.10.157.

233. L. A. Stapinski, B. Reda, N. C. Newton, S. Lawler, D. Rodriguez, C. Chapman, and M. Teesson. “Development and evaluation of ‘Pure Rush’: An online serious game for drug education”. *Drug and Alcohol Review* 37:S1, 2018, S420–S428. DOI: 10.1111/dar.12611.

234. *State of AI in the Enterprise, 2nd Edition*. Technical report. Deloitte Insights, 2018. URL: https://www2.deloitte.com/content/dam/insights/us/articles/4780_State-of-AI-in-the-enterprise/DI_State-of-AI-in-the-enterprise-2nd-ed.pdf.

235. P. Stothard and A. van den Hengel. “Development of serious computer game based training module and its integration into working at heights mine site induction—Part I”. *Mining Technology* 119:2, 2010, pp. 68–78. DOI: 10.1179/037178410X12780655704644.

236. A. Streicher and M. Aydinbas. "Bayesian Cognitive State Modeling for Adaptive Serious Games". In: *Adaptive Instructional Systems*. Springer International Publishing, Cham, 2022, pp. 14–25. ISBN: 978-3-031-05887-5. DOI: 10.1007/978-3-031-05887-5_2.

237. A. Streicher and J. D. Smeddinck. "Personalized and Adaptive Serious Games". In: *Entertainment Computing and Serious Games: International GI-Dagstuhl Seminar 15283, Dagstuhl Castle, Germany, Revised Selected Papers*. Springer International Publishing, Cham, 2016, pp. 332–377. ISBN: 978-3-319-46152-6. DOI: 10.1007/978-3-319-46152-6_14.

238. L. E. Sucar. *Probabilistic Graphical Models: Principles and Applications*. Springer International Publishing, 2021. ISBN: 9783030619435. DOI: 10.1007/978-3-030-61943-5.

239. S. Sugasawa. "Grouped Heterogeneous Mixture Modeling for Clustered Data". *Journal of the American Statistical Association* 116:534, 2021, pp. 999–1010. DOI: 10.1080/01621459.2020.1777136.

240. D. Sultan, B. Omarov, Z. Kozhamkulova, G. Kazbekova, L. Alimzhanova, A. Dautbayeva, Y. Zholdassov, and R. Abdrakhmanov. "A Review of Machine Learning Techniques in Cyberbullying Detection." *Computers, Materials & Continua* 74:3, 2023. DOI: 10.32604/cmc.2023.033682.

241. F. Tao, B. Xiao, Q. Qi, J. Cheng, and P. Ji. "Digital twin modeling". *Journal of Manufacturing Systems* 64, 2022, pp. 372–389. ISSN: 0278-6125. DOI: 10.1016/j.jmsy.2022.06.015.

242. J. Textor, B. van der Zander, M. S. Gilthorpe, M. Liśkiewicz, and G. T. Ellison. "Robust causal inference using directed acyclic graphs: the R package 'dagitty'". *International Journal of Epidemiology* 45:6, 2017, pp. 1887–1894. ISSN: 0300-5771. DOI: 10.1093/ije/dyw341.

243. *The State of the World's Children 2017: Children in a Digital World*. Technical report. ISBN: 978-92-806-4930-7. UNICEF Division of Communication, 2017. URL: https://www.unicef.org/media/48581/file.

244. N. Thomas and A. Woodyatt. *Children with ADHD can now be prescribed a video game, FDA says*. CNN Health News. 2020. URL: https://edition.cnn.com/2020/06/16/health/adhd-fda-game-intl-scli-wellness/index.html.

245. J.-N. Tioh, M. Mina, and D. W. Jacobson. "Cyber security training a survey of serious games in cyber security". In: *2017 IEEE Frontiers in Education Conference (FIE)*. 2017, pp. 1–5. DOI: 10.1109/FIE.2017.8190712.

246. T. Tong, M. Chignell, M. C. Tierney, J. Lee, et al. "A serious game for clinical assessment of cognitive status: validation study". *JMIR serious games* 4:1, 2016, e5006. DOI: 10.2196/games.5006.

247. J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and 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. DOI: 10.1109/cvprw.2018.00143.

248. T. T. Um, F. M. J. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and 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. ISBN: 9781450355438. DOI: 10.1145/3136755.3136817.

249. C. N. Vasconcelos and B. N. Vasconcelos. "Increasing Deep Learning Melanoma Classification by Classical And Expert Knowledge Based Image Transforms". *arXiv preprint arXiv:1702.07025*, 2017. URL: https://arxiv.org/abs/1702.07025.

250. S. L. Vincenzi, E. Possan, D. F. d. Andrade, M. M. Pituco, T. d. O. Santos, and E. P. Jasse. "Assessment of environmental sustainability perception through item response theory: A case study in Brazil". *Journal of Cleaner Production* 170, 2018, pp. 1369–1386. ISSN: 0959-6526. DOI: 10.1016/j.jclepro.2017.09.217.

251. L. Von Ahn and L. Dabbish. "Labeling images with a computer game". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2004, pp. 319–326. DOI: 10.1145/985692.985733.

252. M. J. Vowels, N. C. Camgoz, and R. Bowden. "D'ya Like DAGs? A Survey on Structure Learning and Causal Discovery". *ACM Comput. Surv.* 55:4, 2022. ISSN: 0360-0300. DOI: 10.1145/3527154.

253. Z. Wan, Y. Zhang, and H. He. "Variational autoencoder based synthetic data generation for imbalanced learning". In: *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2017, pp. 1–7. DOI: 10.1109/SSCI.2017.8285168.

254. B. Wang, T. Sun, and X. S. Zheng. "Beyond Winning and Losing: Modeling Human Motivations and Behaviors with Vector-Valued Inverse Reinforcement Learning". *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 15:1, 2019, pp. 195–201. DOI: 10.1609/aiide.v15i1.5244.

255. R. Wang, S. J. DeMaria, A. Goldberg, and D. Katz. "A Systematic Review of Serious Games in Training Health Care Professionals". *Simulation in Healthcare* 11:1, 2016. ISSN: 1559-2332. DOI: 10.1097/sih.0000000000000118.

256. Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu. "Time Series Data Augmentation for Deep Learning: A Survey". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. IJCAI-2021. International Joint Conferences on Artificial Intelligence Organization, 2021. DOI: 10.24963/ijcai.2021/631.

257. S. Werning. "Generative AI and the Technological Imaginary of Game Design". In: *Creative Tools and the Softwarization of Cultural Production*. Springer Nature Switzerland, 2024, pp. 67–90. ISBN: 978-3-031-45693-0. DOI: 10.1007/978-3-031-45693-0_4.

258. W. Westera, R. Prada, S. Mascarenhas, P. A. Santos, J. Dias, M. Guimarães, K. Georgiadis, E. Nyamsuren, K. Bahreini, Z. Yumak, C. Christyowidiasmoro, M. Dascalu, G. Gutu-Robu, and S. Ruseti. "Artificial intelligence moving serious gaming: Presenting reusable game AI components". en. *Education and Information Technologies* 25:1, 2020, pp. 351–380. ISSN: 1360-2357, 1573-7608. DOI: 10.1007/s10639-019-09968-2.

259. E. J. Williamson, Z. Aitken, J. Lawrie, S. C. Dharmage, J. A. Burgess, and A. B. Forbes. "Introduction to causal diagrams for confounder selection". *Respirology* 19:3, 2014, pp. 303–311. DOI: 10.1111/resp.12238.

260. M. Willson and T. Leaver. "Zynga's farmville, social games, and the ethics of big data mining". *Communication Research and Practice* 1:2, 2015, pp. 147–158. DOI: 10.1080/22041451.2015.1048039.

261. A. G. C. Wright, M. N. Hallquist, J. Q. Morse, L. N. Scott, S. D. Stepp, K. A. Nolf, and P. A. Pilkonis. "Clarifying Interpersonal Heterogeneity in Borderline Personality Disorder Using Latent Mixture Modeling". *Journal of Personality Disorders* 27:2, 2013, pp. 125–143. ISSN: 0885-579X. DOI: 10.1521/pedi.2013.27.2.125.

262. Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng, Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huang, and T. Gui. "The Rise and Potential of Large Language Model Based Agents: A Survey". *arXiv preprint arXiv:2309.07864*, 2023. DOI: 10.48550/arXiv.2309.07864.

263. B. Yang, Z. Liu, G. Duan, and J. Tan. “Mask2Defect: A Prior Knowledge-Based Data Augmentation Method for Metal Surface Defect Inspection”. *IEEE Transactions on Industrial Informatics* 18:10, 2022, pp. 6743–6755. DOI: 10.1109/TII.2021.3126098.

264. L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. “Diffusion Models: A Comprehensive Survey of Methods and Applications”. *ACM Comput. Surv.* 56:4, 2023. ISSN: 0360-0300. DOI: 10.1145/3626235.

265. G. N. Yannakakis and J. Togelius. *Artificial intelligence and games*. Vol. 2. Springer, 2018. ISBN: 978-3-319-63518-7. DOI: 10.1007/978-3-319-63519-4.

266. M. Yao, C. Chelmis, and D.-S. Zois. “Cyberbullying Ends Here: Towards Robust Detection of Cyberbullying in Social Media”. In: *The World Wide Web Conference*. WWW ’19. Association for Computing Machinery, New York, NY, USA, 2019, pp. 3427–3433. ISBN: 9781450366748. DOI: 10.1145/3308558.3313462.

267. N. Yee, N. Ducheneaut, L. Nelson, and P. Likarish. “Introverted Elves & Conscientious Gnomes: The Expression of Personality in World of Warcraft”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’11. Association for Computing Machinery, New York, NY, USA, 2011, pp. 753–762. ISBN: 9781450302289. DOI: 10.1145/1978942.1979052.

268. A. Zanga, E. Ozkirimli, and F. Stella. “A Survey on Causal Discovery: Theory and Practice”. *International Journal of Approximate Reasoning* 151, 2022, pp. 101–129. ISSN: 0888-613X. DOI: 10.1016/j.ijar.2022.09.004.

269. Q. Zhao, E. Adeli, and K. Pohl. “Training confounder-free deep learning models for medical applications”. *Nat Commun* 11, 2020, p. 6010. DOI: 10.1038/s41467-020-19784-9.

270. W. Zhao, J. P. Queralta, and 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. DOI: 10.1109/SSCI47803.2020.9308468.

271. Y. Zhonggen. “A Meta-Analysis of Use of Serious Games in Education over a Decade”. *International Journal of Computer Games Technology* 2019, 2019, p. 4797032. ISSN: 1687-7047. DOI: 10.1155/2019/4797032.

272. G. D. Zimet, N. W. Dahlem, S. G. Zimet, and G. K. Farley. “The Multidimensional Scale of Perceived Social Support”. *Journal of Personality Assessment* 52:1, 1988, pp. 30–41. DOI: 10.1207/s15327752jpa5201_2.

273. Zoomsim. *Zoom Business Simulation*. 2022. URL: https://www.zoomsim.net/.

274. L. Zujovic, V. Kecojevic, and D. Bogunovic. “Interactive mobile equipment safety task-training in surface mining”. *International Journal of Mining Science and Technology* 31:4, 2021, pp. 743–751. ISSN: 2095-2686. DOI: 10.1016/j.ijmst.2021.05.011.