



OPEN Enhancing social science research on cyberbullying through human machine collaboration

Andrea Baños-Ramos¹, María Reneses^{1,2}, Jaime Pérez³, Gabriel Valverde⁴, Edmond Awad^{5,6,7}, Gregorio López López¹ & Mario Castro^{1,8}✉

Cyberbullying (CB) has emerged as a growing concern among adolescents, with nearly 10% of European children affected monthly and almost half experiencing it at least once. Unlike traditional bullying, CB thrives in digital environments where anonymity and impunity are prevalent. Despite its increasing prevalence, understanding the causal mechanisms behind CB remains challenging due to the limitations of conventional statistical methods, which often rely on correlations and are prone to spurious associations. In this paper, we introduce a novel human–machine consensus framework for causal discovery, aimed at supporting social scientists in unraveling the complex dynamics of CB. We leverage recent advances in data-driven causal inference, particularly the use of Directed Acyclic Graphs (DAGs), to identify and interpret causal relationships from observational data. Our approach integrates automatic causal discovery algorithms with expert knowledge, addressing the limitations of both purely algorithmic and purely expert-driven methods, and allows for the creation of a model ensemble estimation of the causal effects. To enhance interpretability and usability, we advocate for the use of Probabilistic Graphical Causal Models (PGCMs), or Bayesian Networks, which combine probabilistic reasoning with graphical representation. This hybrid methodology not only mitigates cognitive biases and inconsistencies in expert input but also fosters transparency and critical reflection in model construction. Cyberbullying serves as a compelling case study where ethical constraints preclude experimental designs, highlighting the value of interpretable, expert-informed causal models for guiding policy and intervention strategies.

Cyberbullying (CB) represents a relevant issue among adolescents nowadays, necessitating effective prevention and intervention strategies. Approximately 10% of European children are cyberbullied every month¹, and nearly 50% have experienced a CB incident at least once². New technologies facilitate CB and often occur on social media platforms, where the aggressor can perceive impunity or easily hide under a false identity³. A recent study showed that, while traditional forms of adolescent violence declined over time, cyberbullying increased since 2018⁴.

Conventional statistical techniques based on correlations or hypothesis testing are common in social science research but can be problematic due to spurious correlations that obscure causal relationships and fail to address data biases⁵. This is especially an issue in contexts where randomized controlled trials are not feasible or ethical, leaving observational studies as the primary option. Thus, establishing causal mechanisms and estimating the effects of interventions are crucial for research and policy development, particularly when traditional methods provide inconsistent or weak evidence.

In recent years, data-driven causal inference research has flourished, focusing on identifying causal relationships between variables using Directed Acyclic Graphs (DAGs)^{6,7}. This approach enhances our understanding of complex issues by unraveling the intricate relationships between variables and illuminating the mechanisms at play. It also encourages researchers to maintain a critical perspective on data collection and manipulation while clearly stating their assumptions and hypotheses, fostering productive discussions⁸.

¹Institute for Research in Technology (IIT), ICAI Engineering School, Universidad Pontificia Comillas, 28015 Madrid, Spain. ²Faculty of Human and Social Sciences, Universidad Pontificia Comillas, 28049 Madrid, Spain. ³AZTI, Food Research, Basque Research and Technology Alliance (BRTA), Derio, 48160 Bizkaia, Spain. ⁴Celering, IT Consultant and Services, 28020 Madrid, Spain. ⁵Department of Economics, University of Exeter, Exeter EX4 4PU, UK. ⁶The Oxford Uehiro Centre for Practical Ethics, University of Oxford, Oxford OX1 1PT, UK. ⁷Center for Humans and Machines, Max Planck Institute for Human Development, 14195 Berlin, Germany. ⁸Grupo Interdisciplinar de Sistemas Complejos (GISC), 28015 Madrid, Spain. ✉email: marioc@comillas.edu

Automatic causal discovery algorithms aim to infer causal Directed Acyclic Graphs (DAGs) from observational data by identifying statistical patterns based on key assumptions. These include causal sufficiency, which assumes all relevant variables are observed with no hidden confounders; the faithfulness condition, where the conditional independencies in the data accurately reflect the causal structure; and the Markov condition, stating that each variable is conditionally independent of its non-effects given its direct causes⁹. Additionally, these methods assume acyclicity to rule out feedback loops and require sufficient data to reliably estimate statistical dependencies. These assumptions enable the use of constraint-based, score-based, and functional model-based approaches to recover causal structures. The validity and utility of these assumptions have been widely discussed in the literature, notably in recent research on deep generative modeling for causal inference, which highlights the significance of unconfoundedness and structural assumptions in high-dimensional contexts¹⁰.

On the human side, constructing Directed Acyclic Graphs (DAGs) for causal inference often relies on expert knowledge, particularly in areas with limited or noisy data. This reliance can introduce methodological challenges, as experts might unintentionally embed cognitive biases like confirmation bias or overconfidence, distorting the causal structure and leading to misleading conclusions. Additionally, expert knowledge is often incomplete or inconsistent, especially when multiple experts provide divergent views on variable relationships. These inconsistencies can complicate the development of a coherent DAG, resulting in models that may be oversimplified or overly complex, thus limiting their effectiveness in empirical analysis¹¹.

To tackle these issues, we aim to create a human-supervised methodology that systematizes the construction of interpretable, expert-informed causal models using accessible tools. Rather than adopting a purely algorithmic approach to DAG building, we train experts to articulate causal connections and constrain algorithmic discovery, while also leveraging these algorithms to refine or expand their hypotheses. We also advocate for the use of Probabilistic Graphical Causal Models (PGCM)¹², commonly referred to as causal Bayesian Networks (BN). This intuitive framework effectively combines probabilistic, structural, and graphical elements of causal inference. 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 important¹³.

This research is part of the European project H2020 RAYUELA (<https://www.rayuela-h2020.eu/>). This multidisciplinary project aims to better understand the factors influencing risky online behavior (related to cybercrime) in a friendly, safe, and non-invasive manner. The RAYUELA consortium is composed, among others, of researchers in psychology, anthropology, ethics, criminology, computer science, and engineering.

State of the art and scope of the work

CB has been approached recently using Machine Learning techniques to detect CB^{14,15} or based on sentiment analysis^{16,17}. While these methods seem to be a promising alternative to traditional statistical methods, they lack two key ingredients that we consider mandatory for such a sensitive subject: There is a lack of a causal-oriented perspective aimed at deploying interventions or informing policymakers, and there is an 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. As discussed in the Introduction, recent advances in causal inference have introduced a wide range of algorithms for automated structure learning from data. However, these methods often require large datasets and strong assumptions, and their outputs can be implausible or uninterpretable in complex social domains like CB^{10,18,19}. However, they leave expert knowledge behind or present it as a false dichotomy between automatic vs human DAG creation.

DAG creation by experts presents important challenges

It is important to emphasize that expert knowledge plays a foundational role in constructing DAGs for causal inference, particularly in complex domains where purely data-driven approaches may struggle to capture nuanced relationships. However, a growing body of literature highlights that expert input is far from infallible and may introduce significant biases into the modeling process. One common issue is the omission of important confounding pathways: experts may inadvertently leave out relevant variables or causal links due to limited scope or disciplinary blind spots, leading to structurally incomplete DAGs that misrepresent the true data-generating process²⁰. This can introduce serious confounding bias, affecting downstream inferences and decisions²¹.

Another critical issue is the difficulty of validating expert-derived DAGs. Unlike data-driven models, expert-based DAGs often lack a clear ground truth, making assessing their accuracy or robustness challenging. This is compounded by the fact that causal relationships are frequently context-dependent and temporally ambiguous, which static DAGs may fail to capture adequately. Additionally, eliciting expert knowledge is inherently subjective and may not scale well to high-dimensional settings, where the number of variables and potential interactions grows rapidly.

Equally problematic is the directional misspecification of causal arrows. Since DAGs hinge on the correct orientation of edges, reversing a causal link—say, inferring that social behavior causes anxiety rather than vice versa—can dramatically alter the estimated effects^{22,23}. In addition, many experts conflate correlation with causation, leading to conceptual errors in DAG design that reflect statistical associations rather than true causal mechanisms. Disagreements among experts can also yield divergent graphs for the same problem, reflecting subjective perspectives and implicit assumptions rather than objective truths^{24,25}.

Hybrid methodologies balance the best of both approaches

Compounding these issues is the difficulty experts face in circumscribing relevant factors, especially in distinguishing between exogenous and endogenous variables, and in selecting the correct level of abstraction²⁶. Moreover, experts often exhibit overconfidence in their judgments, expressing undue certainty in empirically ambiguous or contested relationships. These challenges are further exacerbated by the reliance on untestable

assumptions, particularly in observational studies where the true data-generating process is unknowable and identification depends on strong, unverifiable premises²⁴.

Some proposals in the literature exist for constructing DAGs, combining expert knowledge and data-driven algorithms. In particular, Knowledge Engineering Bayesian networks^{27–29} also contemplates interaction but is restricted to the hierarchy of variables (in Medicine, conditions are on one level and symptoms on the other). Other approaches also integrate expert knowledge and data but aim to find the most likely DAG compatible with the expert constraints and the data³⁰. However, this method does not use causal ingredients (confounders, colliders, and mediators) in the design. Finally, other methods filter out automatically discovered DAGs but, again, these methods rely exclusively on measurements of goodness of fit or likelihood of the DAG and not necessarily on constraints arising from conducted or the existing uncontroversial body of knowledge in the field^{31,32} and do not take into account the implications of the characterization of the interactions (again, the distinction between direct effects, and the presence of confounders, colliders, or mediators).

Here, we aim to simplify and systematize a bidirectional information flow between these two actors. For this purpose, we selected three classical algorithms—PC, Bayesian Search (BS), and Greedy Thick Thinning (GTT)—all implemented in the GeNIe software, which facilitates expert-guided model construction, and compared them with a state-of-the-art algorithm: DAGs with NO TEARS³³.

Our approach emphasizes the integration of expert perspectives and data-driven tools to build a shared understanding of causal relationships. Although other approaches to integration of expert knowledge and data exist³⁴ and more recently³⁵, our emphasis is on involving end-to-end non-engineering experts in the process of building a causal model, which is a key aspect of our work. Also, to move beyond the mere construction of a Bayesian network and shift the emphasis towards concepts such as confounder, mediator, or collider that can enrich their understanding of the problem. 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.

Methods

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)^{36,37}. Given a DAG, namely G , and a joint probability distribution P over a set of discrete variables $X = \{X_1, \dots, X_n\}$, 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. (1) is satisfied. Where pa_i are the direct parent nodes of x_i in G , and $P(x_i | pa_i)$ is the conditional probability distribution.

$$P(x_1, \dots, x_n) = \prod_i P(x_i | pa_i) \quad (1)$$

Conditional probabilities play a crucial role in establishing causality, as they allow us to compute the likelihood that one event will occur, given that another has already happened. 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 chronological³⁸. As discussed above, determining causal relationships requires additional information or assumptions beyond the data, such as experimental manipulation or adjustment for confounding variables in observational studies³⁹.

What do we mean by “expert knowledge”?

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 study⁴⁰. The construction of this DAG is not a trivial task as causality is a *loaded* word with strong connotations (especially in the Social Sciences where it is sometimes interpreted as *deterministic* association). For the sake of consistency, in this section, we expand on previously open-access documents produced during the development of the RAYUELA project⁴⁰.

In the mentioned project, a team formed by psychologists, criminologists, social workers, and sociologists was given a seminar to introduce them to the relevant terminology related to causal inference³⁷, including concepts such as collider, confounder, and mediator, as well as their implications for causal connections (<https://eventos.comillas.edu/101590/detail/si-la-correlacion-no-implica-causalidad-entonces-que.html>). After the exposition, we asked the participants the following questions to help them to identify the main building blocks of the DAG:

- (i) **Confounder:** Can you think of any factor that might influence both your main independent variable and your outcome variable, even if it's not part of your main hypothesis?
- (ii) **Mediator:** Is there a step or process through which your independent variable affects your outcome? What happens because of your independent variable that then leads to the outcome?
- (iii) **Collider:** Is there a variable that is influenced by both your independent variable and your outcome? Something that might be a result of both, rather than a cause?

Before that seminar, the team members had already assembled (and conducted some data gathering) information from different sources (see Table 1): Focus groups⁴⁰; Interviews⁴¹; Court sentences⁴²; and Literature review (this work, see the following paragraphs and section “Literature-informed causal connections”). Here, we summarize the main findings of our team (led by co-author M. Reneses) (see section “Literature-informed causal connections” for a more detailed account). Previous research⁴³ shows that in countries where young

Risk factor	Bibliography	Rayuela research	Confounder	Mediator	Collider
Age	60,61	–	–	Hours of internet ⁶²	–
				CB awareness ⁶³	
Gender	56,57,64	FG ⁴⁰ ; IV ⁴¹ ; CS ⁴²	–	–	Age ⁵⁶
Sexual orientation	51–53	FG ⁴⁰ ; IV ⁴¹	–	–	–
Migratory background	54,55	FG ⁴⁰ ; IV ⁴¹	–	–	–
Hours of internet	43,65	–	Age ⁶²	–	–
Family communication	48,66	FG ⁴⁰	Age, Gender ^{61,67}	–	–
CB awareness	68	FG ⁴⁰	Age, Gender ⁶³	–	–

Table 1. Risk factors for cyberbullying (CB) victimization. *FG* focus groups (see our work in Ref.⁴⁰), *IV* interviews (see Ref.⁴¹), *CS* court sentences (see Ref.⁴²).

people spend less time online, they reported lower rates of CB. This trend includes CB victimization (e.g.,^{44,45}) as well as CB perpetration (e.g.,^{46,47}). Besides, victimization and perpetration are also related to less open and more avoidant communication with parents⁴⁸ and family conflicts⁴⁸. Aizencot⁴⁹ showed the connection between Internet and social media activity, online self-disclosure, and the education institution phase and CB victimization. Previous cyber victimization also increases the likelihood of later CB activities⁵⁰.

The prevalence of cyber-victimization is greater among LGBTQ students^{51–53}, as well as young migrants and ethnic minorities^{54,55}, and girls⁵⁶. Although girls' and boys' time spent online is similar, their use is not: boys engage in more CB⁵⁷ and report less to an adult when it occurs. Boys are also more aggressive in their interactions⁵⁸. Shohoudi and colleagues⁵⁹ showed that, regardless of background, girls rated abusive behaviors more negatively. Regarding Age, most studies point out that older teenagers are more likely to be both victims and offenders of CB⁶⁰.

Once the information is collected and aggregated from those four sources, the mentioned seminar on causal thinking included hands-on work to clarify the differences between different concepts, in particular, colliders vs confounders, and the perils of controlling (or not) them—including also a discussion on Simpson's effect.

Consensus-driven model building

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, we created Table 1 summarizing our integrated knowledge and as a guide for the first step in the following methodology (a graphical representation of these steps is shown in Fig. 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. The experts are informed on relevant terminology related to Bayesian networks, including concepts such as collider, confounder, and mediator and their implications for causal connections.
- Consistency causal restrictions:** Together with their proposals, the experts forbid a set of causal connections to the algorithms that will be included as initial conditions. Naturally, one must be completely confident when prohibiting these connections. These restrictions come from common sense or widely accepted knowledge in the literature. Finally, in those cases where the collected evidence is not conclusive or even absent, the potential connection is left without restriction.
- 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.
- Quantitative comparison and consensus:** Once the new algorithmic proposals (with restrictions) and the second proposal of the experts have been obtained, a quantitative comparison 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.

Metrics

Each causal DAG candidate is evaluated using the metrics described below to perform the last step of the proposed methodology, which is 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 $\log(\mathcal{L}(\Theta | x))$. 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 overfitting⁶⁹. Formally, it is expressed as shown in Eq. (2) where k represents the number of parameters estimated, n is the number of data points, and \hat{L} is the maximized value of the likelihood function of the model. Lower BIC scores indicate a better trade-off between fit and complexity.

$$\text{BIC} = k \ln(n) - 2 \ln(\hat{L}) \quad (2)$$

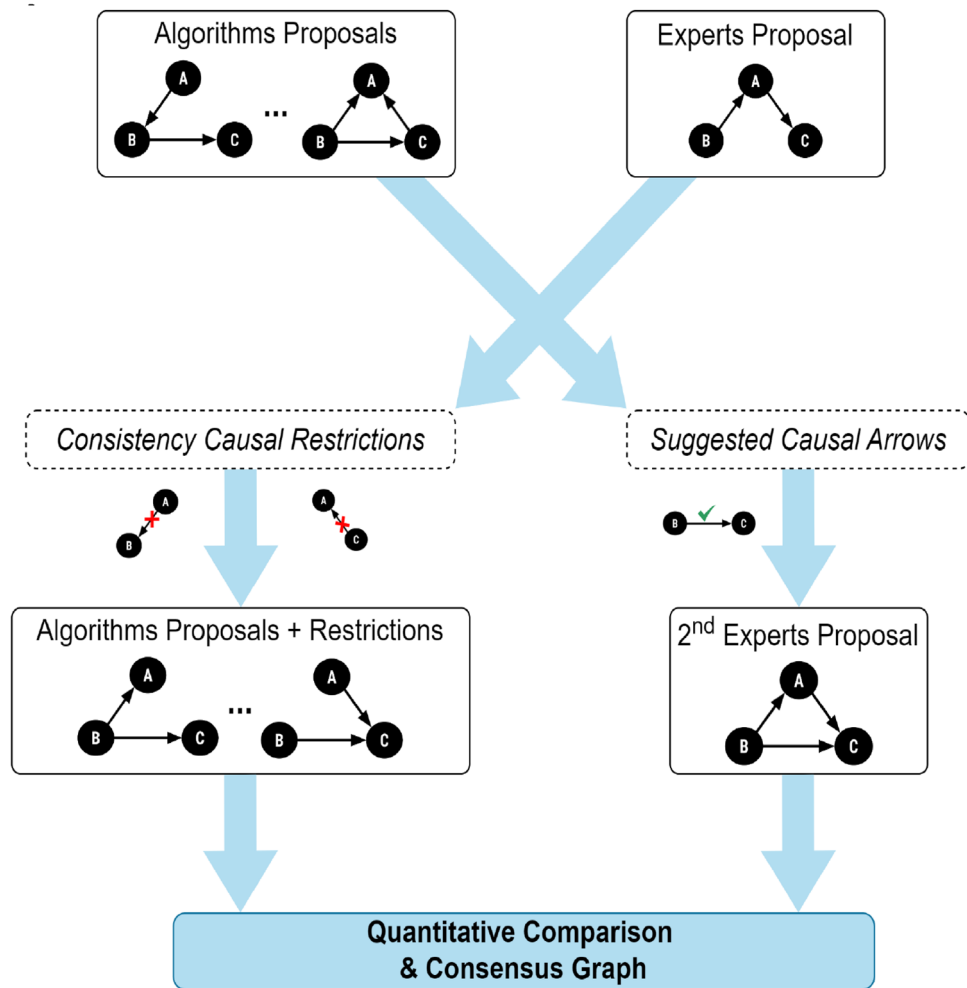


Fig. 1. Conceptual graphical representation of the steps involved in constructing a causal Directed Acyclic Graph (DAG). First, experts and structure learning algorithms propose models based on existing literature and field research. Next, experts impose certain hard restrictions—while remaining open to others—and revise their initial proposals based on the relationships identified by the algorithms. Finally, a quantitative comparison is performed, and consensus is reached, as there are no qualitative differences between semi-supervised automatic DAGs and refined DAGs; both lead to the same conclusions.

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 parameters⁷⁰. 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 evaluates how well the Directed Acyclic Graph (DAG) captures correlations in the data using the d-separation property⁷¹. For each variable pair, a correlation test, typically chi-square, is conducted. We then check if these variables are d-connected in the DAG, leading to the calculation of a classification metric (e.g., F1 score) based on the correlation test as the true value and the d-connections as predicted values. Higher correlation scores reflect better alignment between the data correlations and the DAG.

The choice of metric is influenced by the application goals, model complexity, prior knowledge, and sample size. Importantly, the objective of creating the *best* DAG is to accurately represent hypothetical causal relationships rather than merely achieving predictive accuracy. Each metric has its pros and cons. The LL score may favor overly complex models, risking overfitting. The K2 score serves as a baseline but depends on specific hyperparameters. The BIC is helpful for finding a balance between model fit and complexity, following Occam's razor. The correlation score focuses on testing conditional independence without assessing data fit.

For the LL score, we used a k-fold cross-validation method to compare candidate models, a common approach in machine learning to mitigate overfitting⁷². This technique involves dividing the dataset into k segments and training the model k times with different test sets. Notably, the LL score uniquely requires model parameters for evaluation, making it the only metric justifying k-fold cross-validation.

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?*)³⁷.

In this paper, 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)⁷³, 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 notation⁷⁴, the ACE can be written as in Equation (3). The do operator symbolizes an intervention and can be defined as in Equation (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 criterion⁷⁵. 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.

$$\text{ACE} = \text{ATE} = \mathbb{E}[Y \mid \text{do}(T := A)] - \mathbb{E}[Y \mid \text{do}(T := B)] \quad (3)$$

$$P(Y \mid \text{do}(T := t)) = \sum_{s \in S} P(Y \mid S = s, T = t)P(S = s) \quad (4)$$

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 ($\text{ACE} = P_A - P_B$) and, equivalently, as an Odds Ratio (OR) following Equation (5). Having obtained the results for each variable, we can rank them. Larger ACE values indicate a greater causal influence on the outcome.

$$\text{Odds Ratio (OR)} = \frac{P_A(1 - P_B)}{P_B(1 - P_A)} \quad (5)$$

Ensemble causal analysis

One benefit of the LL and BIC metrics is that they can be interpreted probabilistically. In particular, given a model M_i , with parameters θ_i over data D :

$$P(D|M_i) = \exp(\text{Log-Likelihood}).$$

Similarly, BIC arises from a Laplace approximation to the marginal likelihood (*e.g.*, the evidence) of a model⁷⁶:

$$P(D | M_i) = \int P(D | \theta_i, M_i) \cdot P(\theta_i | M_i) d\theta_i$$

This integral is often intractable, but under certain assumptions (large sample size, regular priors, etc.), it can be approximated as:

$$\log P(D | M_i) \approx \log L_i - \frac{k_i}{2} \log n.$$

So, from (2)

$$P(D | M_i) \propto \exp\left(-\frac{1}{2} \cdot \text{BIC}_i\right) p(M_i).$$

If we do not have any preferred model, then all the prior probabilities for each model, $p(M_i)$, are equal. Using both definitions, in Table 5 we compute an *Ensemble Average Causal Effect* (EACE), in which we weigh ACE for each model according to its plausibility. Namely, for each model i ,

$$w_{M_i} = \frac{P(D | M_i)}{\sum_j P(D | M_j)} \implies \text{EACE} \equiv \sum_i w_{M_i} \text{ACE}_{M_i}. \quad (6)$$

We think that borrowing the concept of ensemble prediction and adapting it to the estimation of the causal effect is an interesting addition to our methodology.

Ablation analysis

Ablation analysis is a geometrically-inspired methodology to evaluate the robustness and reliability of networks after link removal, mostly used in the context of deep learning⁷⁷, but it can be used to analyze BN structures, particularly when comparing expert-constructed and automatically learned DAGs against a ground truth. It involves systematically removing individual edges from the DAGs to assess their impact on key performance metrics. This approach reveals which edges are most influential to the model's fit and accuracy, providing insights into potential overfitting, misspecified relationships, or the loss of critical causal links. By comparing

these metrics before and after edge removal, discrepancies between the expert knowledge embedded in the DAG and the data-driven inferences of automated algorithms can be quantified, thereby validating or refining the network structure.

Here, for illustration, we will limit the analysis to variations after edge removal of LL and Structural Hamming Distance (SHD). These parameters measure how different two graphs are by counting the minimum number of edge insertions, deletions, or direction flips needed to transform one graph into the other. In our case, we do not have the ground truth, but as the final DAGs are relatively similar—they are indeed for the 2nd experts' proposal and DAGS with NO TEARS with restrictions—we will use these ones as an effective ground truth in the analysis.

Data collection

This dataset was collected as part of a previous study⁷⁸ through a representative survey of children in Madrid (Spain) schools, freely available at the Zenodo repository (<https://doi.org/10.1016/j.childyouth.2025.108285>). Informed consent has been obtained from a parent and/or legal guardian. All methods were performed in accordance with the relevant guidelines and regulations.

The survey collected responses from 682 students aged 13 to 17, where 46.6% identified themselves as males, 45.2% as females, 3.1% as non-binary, and 5.1% preferred not to say. The data includes demographic information (*Age, gender, sexual orientation, migratory background, and family communication*), participants' relationship with technology and the Internet, and 4 inquiries about Cyberbullying-related situations and 1 (question 13) related to physical isolation.

Table 2 provides the variable values and their percentage of occurrences (i.e., marginal probability). To analyze the data, we defined a binary variable called *Cyberbullying Victimization Risk* that takes the value 1 if the respondent has answered at least 2 out of the 5 mentioned questions. We use this threshold in 2 answers rather than in 1 to improve the specificity of the survey regarding CB Victimization.

The Universidad Pontificia Comillas Ethics Committee approved the data collection and experimental procedures. In addition, RAYUELA's legal experts also took the necessary measures to ensure that data collection,

Variable	Response values	Marginal probability (%)
Gender	Prefer not to say	5.1
	Male	46.6
	Female	45.2
	Non Binary	3.1
Age	13	0.6
	14	45.3
	15	36.4
	16	14.4
	17	3.4
Sexual orientation	Prefer not to say	5.1
	Heterosexual	75.5
	Non heterosexual	19.4
Migratory background	No	78.2
	Yes	21.8
Family communication	NS/NC	3.1
	1 (never)	5.6
	2	19.9
	3	39.6
	4 (very frequently)	31.8
Daily hours of internet	NS/NC	6.2
	Less than 1 h	4.8
	1–2 h	33.3
	2–4 h	32.6
	More than 4 h	23.2
CB awareness	NS/NC	3.5
	1 (not worried)	5.7
	2	20.8
	3	30.9
	4 (very worried)	39
CB victimization risk	No	86.5
	Yes	13.5

Table 2. Survey of Spanish minors dataset ($N = 682$): the table shows the possible values of each variable and its marginal probability (i.e., the percentage of observations).

storage, and disclosure comply with the European GDPR. In each session, the researchers and teachers explained the project, its main objective, and the data to be collected.

Collaboration with cyberbullying experts from RAYUELA

The expert knowledge comes from members of the RAYUELA project consortium. Work package 1 in that 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 the pathology and physiology of online behaviors, characterizing the victims and offenders of cybercrime and 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)⁴¹ and analyzed 46 court sentences⁴². As a result, the team acquired a profound understanding of the issue, which has been used on several occasions throughout the project and 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, who were the leaders of this work package.

Results

This section presents the results from the case study where we apply the proposed methodology for generating robust causal DAGs. Firstly, we will show the proposals from the experts and the structure learning algorithms. Then, we will apply the method to modify the proposals and compare them quantitatively. We assess the validity using the metrics described in the “Methods” section and a systematic ablation study. Finally, we will perform causal tasks on the winning models to extract relevant information and verify the robustness of the proposed causal DAGs while providing an integrated human–machine ensemble estimation of the causal metrics.

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. 2). The experts do not have any information about the dataset (not even a basic exploratory analysis), and the algorithms are trained without restrictions or forced connections from expert knowledge. As discussed in Table S1, many algorithms in the literature aim to refine different aspects of causal discovery. As shown in Fig. 2c–f, these methods lead to the same unrealistic problems as those used in the main text, pointing to unrealistic causal connections. For instance, the proposition that experiences related to cyberbullying and migratory background influence an individual’s sexual orientation or that sexual orientation determines gender and family communication.

The problem here is twofold: first, although the sample size is considerable for a Social Science study ($n = 682$), not all the classes are equally balanced (as shown in Table 2). This can be responsible for the differences in the role of, for instance, *Cyberbullying Awareness* across different methods. As shown in Sect. S5 (Supporting Information), the minimum sample size for this problem is about 2600 entries in conservative cases, potentially reaching 10^5 based on standard Machine Learning rules (which suggest at least 10 data points per parameter). Consequently, automatic algorithms tend to be biased towards more frequently represented parameter combinations.

Besides, in the case of DAGs with NO TEARS, the regularization method is imposed to forbid automatic loops that (by definition) should be absent in a DAG, and this, probably behind the reversed arrows in Gender → Family Communication or Sexual Orientation → Gender. Thus, as emphasized throughout this work, expert input is mandatory to avoid nonsensical causal implications. We illustrate the dramatic effect of not including one mediator on the discovered DAG in Sect. S3 (Supporting Information).

Another problem is that, while experts can successfully merge different causal routes between two variables (and different mediators) into a single causal arrow (collectively aggregating the total interventional effect), this lack of information on variables missing from the data is not guaranteed in the case of automatic discovery (Fig. S2 in Supporting Information). It is worth noting that this arrow reversal is displayed in the connection between the outcome variable and Hours of internet in the case of BS, PC, and GITT. Also, except in the case of DAGs with NO TEARS, they also seem to fail the relevance of Age on Victimization, something that even a black-box algorithm can capture (see Sect. S4 in the Supporting Information).

For reference, Fig. 2a shows the structure that would resemble a traditional statistical pairwise association between factors and the target variable. We have called this structure “naive” as it does not incorporate any causal assumption and presupposes that all the variables are potentially explanatory. Next, Fig. 2b shows the experts’ first proposal where causal relationships were extracted from the group discussions and the systematic analyses summarized in Table 1. Fig. 2c–f correspond to those obtained through the structure learning algorithms. We can observe that the algorithms’ proposals often go against common sense. For example, in Fig. 2c, *sexual orientation* causes *gender*, and *gender* causes *age*; in Fig. 2d, numerous variables cause *gender*, etc. Moreover, in these algorithms’ proposals, some nodes are left unconnected in the network, such as *migratory background*. Overall, this nonsensical connection is related to two issues with the data: scarcity and class imbalance.

To overcome the limitations of automatic algorithms while extracting useful information from the data, 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). 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 *migratory background*. Based on the research, they force the

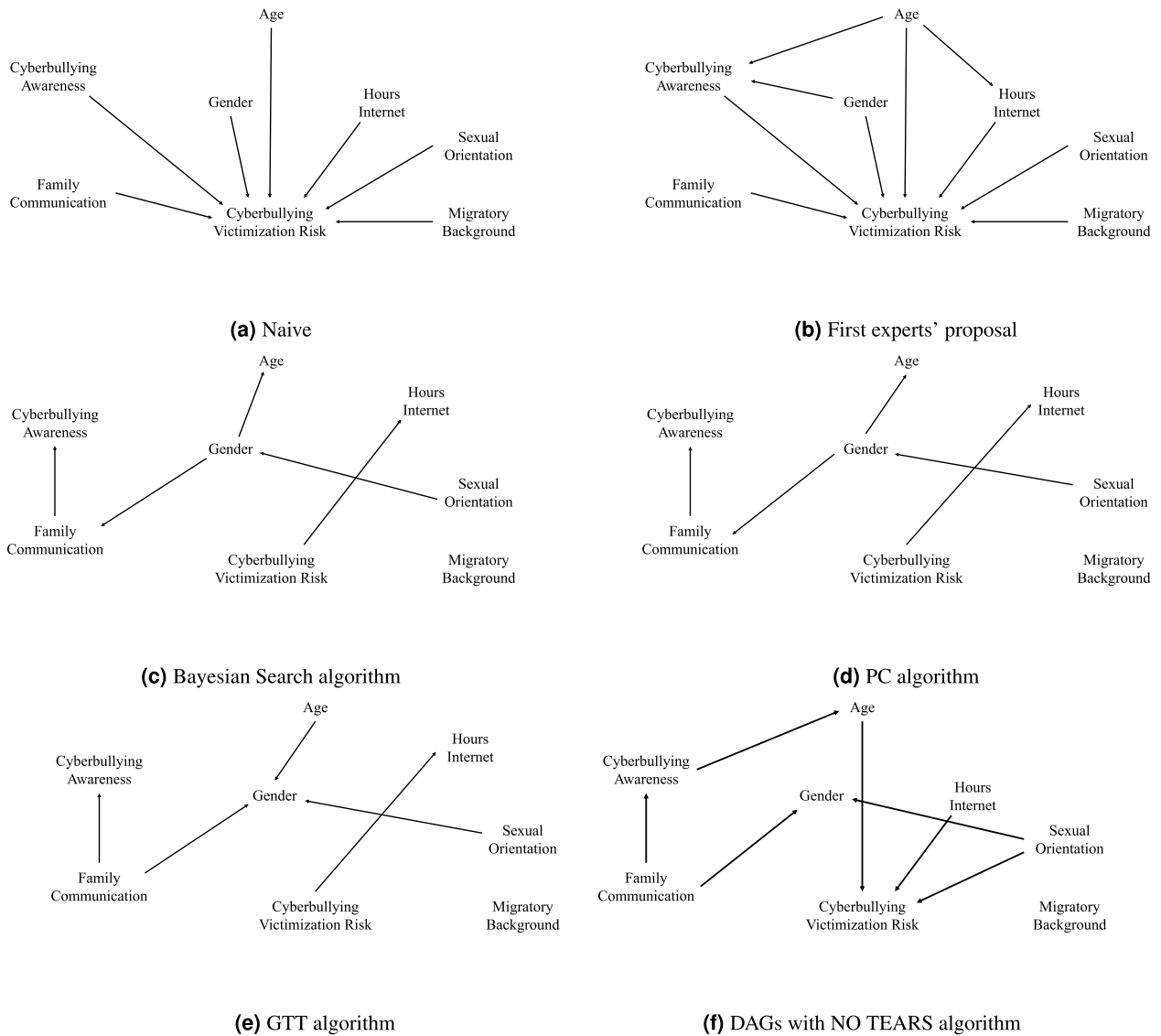


Fig. 2. Step 1: Initial proposed causal DAGs. Structure (a) is a baseline where all variables are connected to the target variable. Structure (b) is the first experts' proposal. Structures (c)–(e) were derived by the algorithms without using restrictions or forced connections. For the sake of completeness, the popular DAGs with NO TEARS algorithm is shown in panel (f).

link between *gender* and *Cyberbullying Victimization Risk*. Similarly, the experts reconsider some of the causal arrows (both presence and direction).

As a result, we obtain refined versions of the initial proposals (Fig. 3). For ease of visualization, the connections that differ between the proposals have been highlighted in blue. Figure 3a–c correspond to those obtained from the structure learning algorithms, using restrictions and forced connections from the experts. The structure shown in Fig. 3e corresponds to the 2nd experts' proposal after analyzing the initial results from the algorithms in detail. 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. Notably, after correcting for expert knowledge, DAGs with NO TEARS (Fig. 3d) and the 2nd experts' proposal are equal.

Model comparison and evaluation

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 “Metrics”. While in the context of the Social Sciences, the existing body of knowledge must be reflected in the model, and the concept of *goodness of fit* is not as relevant as in other fields, here 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 Table 4 are the mean of the LL in the test segments. Figure 4 shows the results, including the standard deviations of the LL score in the test sets. Table 4 presents the findings from this quantitative comparison phase. It is important to note that all



Variable	 Forbidden connection	 Forced connection
Gender	[Age, sexual orientation, and migratory background]	[CB Victimization Risk]
Age	[Gender, sexual orientation, and migratory background]	[CB Victimization Risk]
Sexual orientation	[Age, gender, and migratory background]	[CB Victimization Risk]
Migratory background	[Age, gender, and sexual orientation]	-
Family communication	[Age, gender, sexual orientation, migratory background, and CB awareness]	-
Daily hours of internet	[Age, gender, sexual orientation, and migratory background]	[CB Victimization Risk]
CB awareness	[Age, gender, sexual orientation, migratory background, and family communication]	-
CB-related situations	[Age, gender, sexual orientation, migratory background, family communication, hours of Internet, and CB awareness]	-

Table 3. Step 2 and 3: list of restrictions and forced connections imposed by the experts in training the structure learning algorithms. Other potential restrictions are omitted when the Experts do not have strong evidence and are unconstrained.

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. 3a), GTT (Fig. 3c), and the 2nd experts' proposal/DAGS with NO TEARS (Fig. 3e). Thus, these structures will be used for the following causal analysis.

Average causal effect estimation

Having determined the winning structures, we computed the ACE explained in section “Performing causal tasks”. This analysis allows us to estimate the causal influence that each network node has on the target, which in this case is *Cyberbullying related situations*, see Table 5. This provides insights into the most relevant variables influencing the model output, considering them individually. As outlined in the methodology, we express the ACE as a percentage difference and the equivalent OR. Combining them with the concept of ensemble prediction—encapsulated in the metric EACE—we find that the variable with the greatest causal influence is *hours of Internet* followed by *Age* and *Sexual Orientation*. These findings can inform specific awareness campaigns focused on younger ages or specific sectors of the population and on the abuse of hours spent online. The remaining variables have a highly similar causal influence on the output of the models, except for *family communication* and *CB awareness*, which have the most negligible influence in all the proposals.

Front-door adjustment and the role of age

Using PGCMs provides key benefits for our study, such as enabling intervention simulations and reducing spurious correlations. It also requires explicit hypotheses, promoting transparency and critical discussion in line with open science principles. Notably, we observe that while *Age* is a risk factor, its effect diminishes over time. In contrast to the possibility that internet hours are driven by other variables like gender, our analysis shows a direct effect. Our causal network quantifies this using the front-door adjustment³⁷. The unadjusted EACE of Age on victimization risk is 0.06 (Table 5), and after adjustment.

$$P_{\text{front-door}}(\text{CB}|\text{Age}) = \sum_{\text{Hours}} P(\text{Hours}|\text{Age}) \sum_{\text{Age}'} P(\text{CB}|\text{Hours}, \text{Age}') P(\text{Age}'). \quad (7)$$

In this case, we find that the value of ACE, if we correct for mediators and unmeasured confounders, is EACE = 0.004. Note that this is almost 0, implying that Age is not, by itself, an important cause of CB victimization, but rather it is solely mediated by the number of hours spent online by adolescents (in particular, EACE(Hours|Age) = 0.035). This finding has important implications for prevention efforts and the ongoing debate on connectivity among young people. Reducing the number of hours online may serve as a protective measure by lowering exposure, which would be a critical message to convey in both educational and family settings. This analysis also diminishes the potential impact of a latent variable⁷⁹ as the front-door adjustment corrects for those potential hidden confounders³⁹.

Of course, this analysis does not discard other potential latent explanatory variables not taken into account in this study, such as the socioeconomic status of the participants (which was minimized by sampling selection) or other personal factors, such as personality. *Personality*, moreover, is not one of the most frequently mentioned risk factors when talking about adolescents and CB. In addition to the volatility of the developmental stage, personality may affect other CB risk factors, such as personal information disclosure, but not so much the ones mentioned. Previous research⁸⁰ has found, indeed, that gender and age moderate the relationships between personality traits and CB. However, the direct effect of gender and age can be explained, on the one hand,

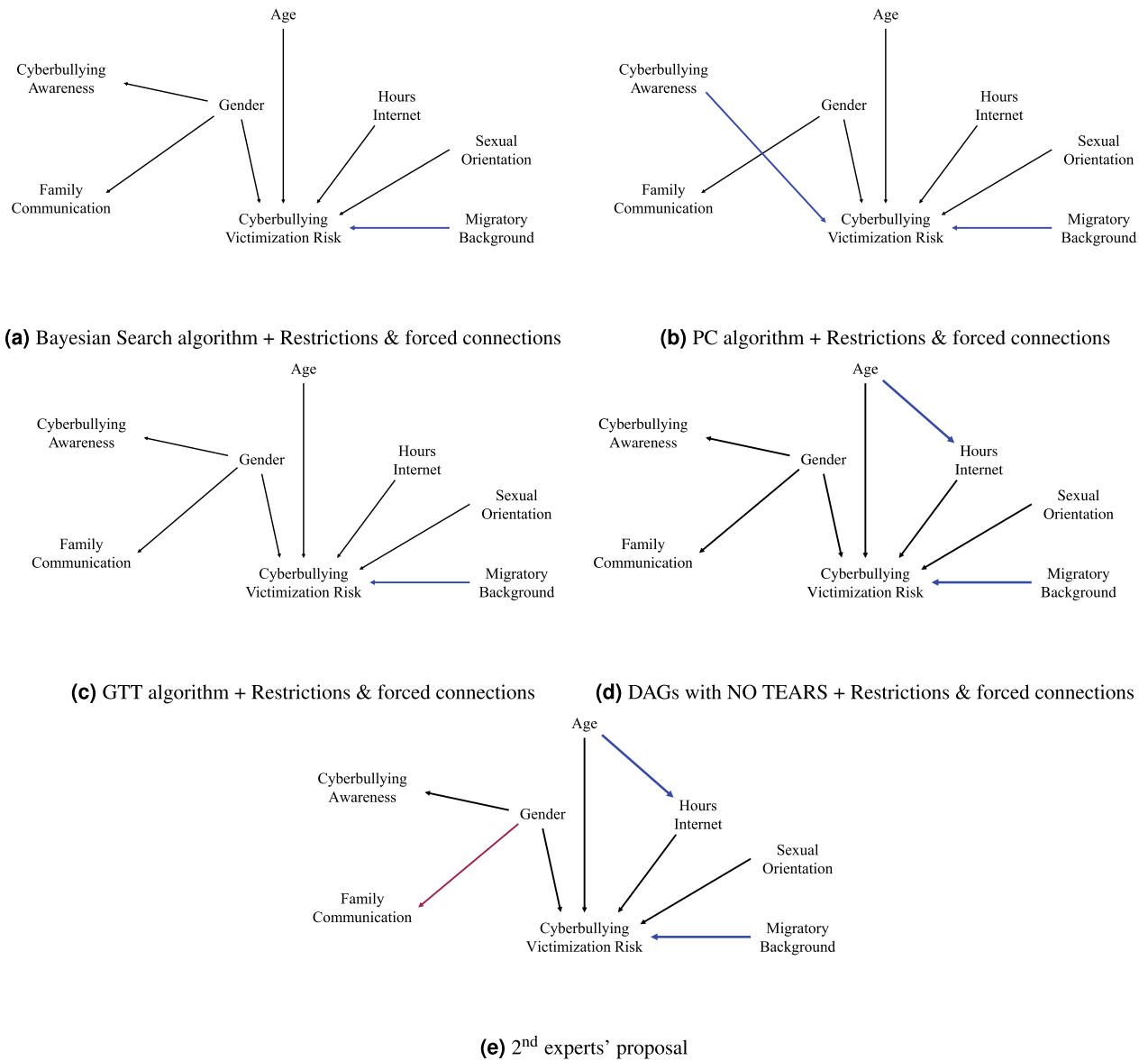


Fig. 3. Steps 2 and 3: Refined causal DAGs obtained through the methodology described in this chapter. Structures (a)–(d) were derived by the algorithms using the restrictions and forced connections from expert knowledge (Table 3). Structure (e) is the 2nd experts' proposal. For ease of visualization, the connections that differ between the proposals have been highlighted in blue, except the one in red that represents a case not taken into account in the first expert proposal but that is included after inspection by the experts of the first round of automatic causal discovery (Fig. 2a–e).

Metrics	Naive	1st experts' proposal	Bayesian search	PC	GTT	2nd experts' and NO TEARS
Mean Log likelihood [test sets]	-1248.54	-1234.92	-1200.72	-1218.82	-1200.72	-1197.39
BIC	54,121.09	54,362.15	7274.62	15,027.07	7274.62	7315.01
K2 score	-5521	-5546.13	-5421.98	-5467.67	-5421.98	-5432.44
Correlation score	0.61	0.63	0.67	0.63	0.67	0.63

Table 4. Step 4: Quantitative comparison of causal DAGs. Cells highlighted in bold indicate the best result in each of the metrics. BS and GTT win in those metrics that penalize complexity, but the 2nd Experts'/NO TEARS proposals are the ones that explain best the dataset.

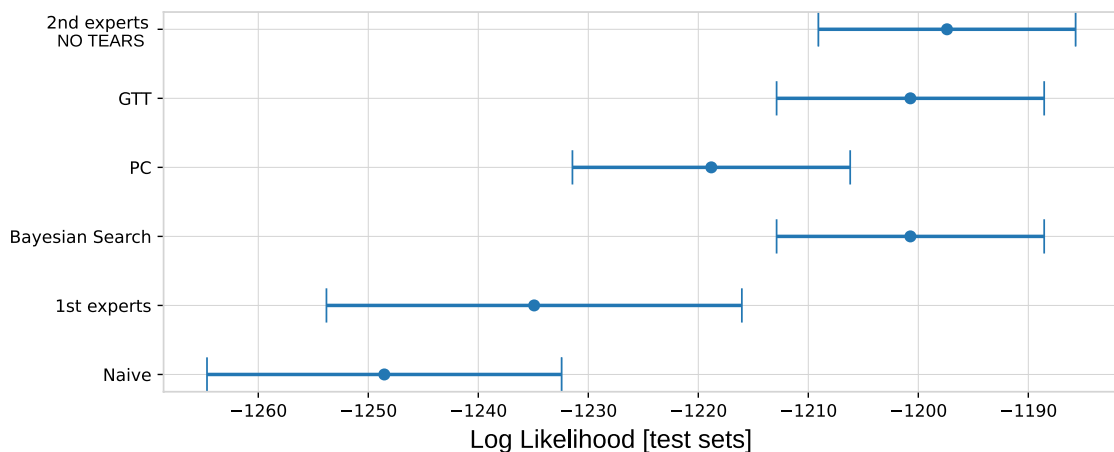


Fig. 4. Results of the cross-validation ($k = 5$) evaluation using the Log-Likelihood (LL) metric on the test sets. Standard deviations are included. The 2nd experts’/NO TEARS proposals are the winner, although it should be noted that the mean of the Bayesian Search and GTT proposals are within their standard deviation.

Variable	Bayesian search		GTT		2nd experts’/NO TEARS		Ensemble ACE	
	ACE (%)	Odds ratio	ACE (%)	Odds ratio	ACE (%)	Odds ratio	EACE _{LL} (%)	EACE _{BIC} (%)
Daily hours internet	6	1.69	6	1.69	6	1.69	6	6
Age	5	1.55	5	1.55	6	1.58	6	5
Sexual orientation	4	1.36	4	1.36	4	1.40	4	4
Gender	2	1.18	2	1.18	3	1.27	3	2
Migratory background	1	1.05	1	1.05	0.1	1.02	0.1	1
CB awareness	0	1.00	0	1.00	0	1.00	0	0
Family communication	0	1.00	0	1.00	0	1.00	0	0

Table 5. Average causal effect estimation across three causal structures: BS (Fig. 3a), GTT (Fig. 3c), and 2nd experts’ Proposal/DAGs with NO TEARS (Fig. 3d, e). Larger values indicate a greater causal influence on the outcome. ACE 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 the ACE (Eq. 5). The last two columns stand for the ensemble prediction (EACE) using weights computed from Eq. (6) using LL and BIC to compute the weights.

through the reproduction of gender roles and social structure and, on the other hand, through maturity and evolutionary characteristics.

Ablation analysis results

Figure S6 in the Supporting Information presents the results of an ablation analysis conducted on two directed acyclic graphs (DAGs): the 1st experts’ proposal DAG and all the automatically generated ones in Step 1 (before including expert restrictions on the DAGs). The analysis evaluates the impact of removing individual edges on the change of SHD (Δ SHD) and LL (Δ LL), shown in each subplot. As a rule of thumb, the more negative the values of Δ SHD, the worse, and the more negative Δ LL, the more important is that arrow with respect to the ground truth. Also, models with more green bars are better from the beginning. The ablation analysis across the models highlights the superior performance of the 1st experts’ proposal and the DAGs with NO TEARS models. Both include arrows that are determinant in comparison with the final DAGs (reflected in the values of Δ SHD). Positive values of Δ SHD mean that the arrow is important in comparison with the ground truth. That, added to the fact that the model maintains structural integrity and data fit.

In Table 6 we filter for those arrows that already have correct edges in all the initial proposals and display the value of the Δ LL. As shown, none of the automatically discovered DAGs without restrictions was as plausible as the first experts’ proposal. This justifies quantitatively the idea that algorithms are not necessarily reliable out of the box.

Model	Removed arrow	ΔLL
1st experts'	Age \rightarrow Hours_Internet	-14.19
	Gender \rightarrow CB	-32.21
	Age \rightarrow CB	-48.73
	Hours_Internet \rightarrow CB	-51.94
	Sexual_Orientation \rightarrow CB	-26.31
BS	Gender \rightarrow Family_Communication	-19.05
GTT	Gender \rightarrow Family_Communication	-19.05
NO TEARS	Age \rightarrow CB	-12.08
	Hours_Internet \rightarrow CB	-15.53
	Sexual_Orientation \rightarrow CB	-10.86

Table 6. Ablation analysis results across models for originally correct causal arrows. Note that the 1st experts' proposal has the most negative values of all, meaning that removing those arrows individually produces less likely models.

Discussion

To demonstrate the validity of our proposal, we present a case study applying a consensus-building framework rather than performing causal discovery in the strict sense. Our emphasis is on assisting domain experts in refining and validating causal models that are theoretically grounded. While our method uses structure learning algorithms, these serve a secondary role: to inform or challenge expert beliefs, not to replace them. We deliberately rely on well-established methods (PC, GTT, BS) because they offer interpretability and ease of constraint integration, which is essential in sensitive and data-scarce settings. Remarkably, DAGs with NO TEARS performs particularly well and converge to the same DAG as the experts. However, these algorithms alone do not resolve the core limitations: sample scarcity and the existence of non-negotiable causal connections defined by theory or policy relevance. However, overall, the value of using causality-based approaches like PGCM (Probabilistic Graphical Causal Models) prioritizes understanding the underlying mechanisms of phenomena, promoting discussion and open science, which makes them particularly suitable for sensitive issues like bullying. In the context of causality, there is no such thing as an *optimal* DAG. Therefore, we can use LL and BIC obtained during training to probabilistically combine the best-scoring models into a single ensemble prediction. While the results may not differ significantly from those of any individual model, there may be other problems or fields where greater discrepancies arise. In such cases, this approach could serve as a useful tool for balancing different solutions.

Although our work has produced positive outcomes, we must acknowledge the existence of unmeasured factors that influence bullying, which limits the scope of our results. The complex nature of this social phenomenon suggests that additional variables may play a significant role in its dynamics, highlighting the need for future research to capture a more comprehensive understanding. The accuracy of our findings is closely linked to the quality and diversity of the data we used. Our study focused on a group of minors from Madrid (Spain), which may limit the generalizability of our methods. To strengthen our conclusions, we must validate our approach with a larger and more diverse dataset encompassing various demographic and cultural contexts.

The finding that *Gender* and *Sexual Orientation* exert a direct effect (independent, for instance, of hours spent online) is highly significant for prevention efforts. CB prevention programs often focus on social skills and empathy, yet they make limited reference to structural issues that are also instrumental in these behaviors. For example, the reinforcement of gender roles (and heteronormativity) and the tendency to punish defiance of these norms are influential factors that deserve attention in prevention strategies.

Another implication of our causal diagram is related to the spurious correlations induced by stratifying a collider (in this case, the target variable, *Cyberbullying Victimization Risk*). In particular, we have not found any evidence on the role of gender in the number of hours spent online. However, if we stratify on the victims (in our case, meaning that we condition on the target value being 1), we find a strong association between them: the mode of the conditional distribution for females is 2-4 per day, while for males just 1-2 (and, as Pearl stated³⁷, is one of the benefits of causal reasoning to spot the so-called Simpson effect).

Also relevant for prevention is that the slight effect of receiving information about internet risks at home disappears when the gender of the participants is taken into account. This brings us back to the need to address prevention from a gender perspective, also when carrying out interventions with families: the type of communication that is established at home and our perception of girls as more vulnerable could mean that we do not sufficiently inform boys about the risks of the Internet, of which they are also potential victims as well as future aggressors. As can be seen in that part of the DAG, *Communication*, and *Awareness* are mediated by *Gender*, leading to a negligible causal effect in this case.

It is worth emphasizing that this work aims not to introduce new causal discovery techniques or to compare the accuracy of algorithms. Rather, we aim to create a practical, transparent methodology for guiding social scientists in constructing causal diagrams using a balance of domain knowledge and data. The use of GeNIe's PC, Bs, and GTT is a deliberate choice that favors interpretability, ease of constraint specification, and widespread availability over algorithmic novelty. We reviewed newer alternatives, such as NOTEARS, that provide better results but might be intimidating to program for non-technical social scientists. Also, as already emphasized,

they do not address the structural constraints imposed by expert knowledge, nor the sample size limitations that characterize much of social science research. The key contribution of our work lies in operationalizing a consensus-based modeling process that is reproducible, explainable, and grounded in both theory and data.

Finally, it is important to recognize the inherent limitations of PGCM. The accuracy of our results depends on the restrictions and connections imposed by experts. Additionally, as the number of variables in the PGCM increases, the complexity and computational demands can grow exponentially, potentially making the model difficult to interpret. This underscores the necessity of thoroughly validating our results and assumptions, especially as the model's complexity increases.

Conclusions

This article contributes to the study of CB, 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 using data from a representative survey of children in Spanish (Madrid) schools.

The combination of expert training, expert causal graph-proposal, and PGCM in CB is a novel approach incorporating a causal analytical perspective, allowing us to delve deeper into its complex dynamics and reach a consensus between expert knowledge and data-driven algorithms. This approach distinguishes our work from other methodologies based on machine learning, focused on maximizing predictive power or correlation search techniques commonly employed in classical social sciences, which can sometimes be problematic due to their tendency to rely on spurious correlations and not take into account data biases.

Integrating expert knowledge with external data using structure learning algorithms overcomes the inherent limitations of each approach. 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 developmental stages, such as promoting acceptance of diverse sexual orientations and gender identities among children and implementing awareness campaigns. However, it is worth acknowledging the unexplored factors influencing victimization that were not captured in our survey. Thus, the conclusions drawn from the results may be limited. Despite these limitations, adopting PGCM and incorporating expert knowledge is invaluable for elucidating complex phenomena, especially in scenarios with limited sample sizes, as is often the case in the social sciences.

In summary, this research proposes a general methodology to obtain robust causal models, improves our knowledge of CB, and demonstrates the effectiveness of causality-based methods in addressing similar issues. By doing so, we enable the creation of better preventive measures to reduce cybercrime among children and ensure their safety and well-being.

Finally, some potential improvements in our methodology should be mentioned. The trickiest part is codifying expert knowledge regarding causal relationships between variables. Here, we have constrained only those arrows sustained by strong evidence while leaving the rest unconstrained. An improvement would be to leave all unconstrained but use weights stating the confidence (as odds ratios or probabilities) that could serve as priors for those automatic methods. Another improvement could be the use of *signed* biases (the direction of the effect) as prior information for the algorithms. These two suggestions could simplify DAG building and reduce the biases that the experts forced.

Data availability

The datasets generated and/or analysed during the current study are available in the zenodo repository <https://zenodo.org/records/15245216>.

Received: 11 April 2025; Accepted: 13 August 2025

Published online: 26 September 2025

References

- Smahel, D. et al. EU Kids Online 2020: Survey results from 19 countries. Tech. Rep., EU Kids Online (2020). ISSN: 2045-256X.
- European Commission et al. *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).
- Englander, E., Donnerstein, E., Kowalski, R., Lin, C. A. & Parti, K. Defining cyberbullying. *Pediatrics* **140**, S148–S151. <https://doi.org/10.1542/peds.2016-1758U> (2017).
- Molcho, M. et al. Trends in indicators of violence among adolescents in Europe and North America 1994–2022. *Int. J. Public Health* **70**, 1607654 (2025).
- Rohrer, J. M. Thinking clearly about correlations and causation: Graphical causal models for observational data. *Adv. Methods Pract. Psychol. Sci.* **1**, 27–42. <https://doi.org/10.1177/2515245917745629> (2018).
- Nogueira, A. R., Pugnana, A., Ruggieri, S., Pedreschi, D. & Gama, J. Methods and tools for causal discovery and causal inference. *WIREs Data Min. Knowl. Discov.* **12**, e1449. <https://doi.org/10.1002/widm.1449> (2022).
- Guo, R., Cheng, L., Li, J., Hahn, P. R. & Liu, H. A survey of learning causality with data: Problems and methods. *ACM Comput. Surv.* **53**, 1–37. <https://doi.org/10.1145/3397269> (2020).
- Grosz, M. P., Rohrer, J. M. & Thoemmes, F. The taboo against explicit causal inference in nonexperimental psychology. *Perspect. Psychol. Sci.* **15**, 1243–1255. <https://doi.org/10.1177/1745691620921521> (2020).
- Bareinboim, E. & Pearl, J. Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci.* **113**, 7345–7352 (2016).
- Liu, Q., Chen, Z. & Wong, W. H. An encoding generative modeling approach to dimension reduction and covariate adjustment in causal inference with observational studies. *Proc. Natl. Acad. Sci.* **121**, e2322376121 (2024).
- Gangl, M. Causal inference in sociological research. *Annu. Rev. Sociol.* **36**, 21–47 (2010).

12. Sucar, L. E. *Probabilistic Graphical Models: Principles and Applications* (Springer International Publishing, 2021).
13. Hernan, M. & Robins, J. *Causal Inference: What If* (Chapman & Hall/CRC, 2020).
14. Sultan, D. et al. A review of machine learning techniques in cyberbullying detection. *Comput. Mater. Contin.* **74**, 5625–5640. <https://doi.org/10.32604/cmc.2023.033682> (2023).
15. Yao, M., Chelms, C. & Zois, D.-S. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *The World Wide Web Conference, WWW '19*, 3427–3433 (Association for Computing Machinery, New York, NY, USA, 2019). <https://doi.org/10.1145/3308558.3313462>
16. Bozyiğit, A., Utku, S. & Nasibov, E. Cyberbullying detection: Utilizing social media features. *Expert Syst. Appl.* **179**, 115001 (2021).
17. Rodríguez-Ibáñez, M., Casánez-Ventura, A., Castejón-Mateos, F. & Cuenca-Jiménez, P.-M. A review on sentiment analysis from social media platforms. *Expert Syst. Appl.* **223**, 119862 (2023).
18. Luma-Osmari, S., Ismaili, F., Pathak, P. & Zenuni, X. Identifying causal structures from cyberstalking: Behaviors severity and association. *J. Commun. Softw. Syst.* **18**, 1–8. <https://doi.org/10.24138/jcomss-2021-0139> (2022).
19. Wendong, L., Xiang, X., Su, R. & Li, B. Algorithmic causal structure emerging through compression. [arXiv:2502.04210](https://arxiv.org/abs/2502.04210) (2025).
20. Imbens, G. W., Kallus, M. M., Mao, X.-Y. & Wang, W.-C. Long-term causal inference under persistent confounding via data combination. *ResearchGate* (2024).
21. Ananth, C. V. & Schisterman, E. F. Confounding, causality and confusion: The role of intermediate variables in interpreting observational studies in obstetrics. *Paediatr. Perinat. Epidemiol.* **31**, 319–325 (2017).
22. Haber, N. et al. DAG with omitted objects displayed (DAGWOOD): A framework for revealing causal assumptions in DAGs. *PLoS One* **17**, e0271545 (2022).
23. Eulig, J. et al. How to develop causal directed acyclic graphs for observational health research: a scoping review. *Int. J. Epidemiol.* (2025).
24. Morgan, M. G., Henrion, M. & Clark, R. G. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proc. Natl. Acad. Sci.* **111**, 7166–7171 (2014).
25. Binkytė, A. et al. Causality is key to understand and balance multiple goals in trustworthy ml and foundation models. [arXiv:2502.21123](https://arxiv.org/abs/2502.21123) (2025).
26. Eulig, E., Mastakouri, A. A., Blöbaum, P., Hardt, M. W. & Janzing, D. Toward falsifying causal graphs using a permutation-based test. In *AAAI Conference on Artificial Intelligence* (2023).
27. Julia Flores, M., Nicholson, A. E., Brunskill, A., Korb, K. B. & Mascaro, S. Incorporating expert knowledge when learning Bayesian network structure: A medical case study. *Artif. Intell. Med.* **53**, 181–204. <https://doi.org/10.1016/j.artmed.2011.08.004> (2011).
28. Lachapelle, S., Brouillard, P., Deleu, T. & Lacoste-Julien, S. Gradient-based neural DAG learning. [arXiv:1906.02226](https://arxiv.org/abs/1906.02226) (2019).
29. Luo, Y., Peng, J. & Ma, J. When causal inference meets deep learning. *Nat. Mach. Intell.* **2**, 426–427 (2020).
30. Masegosa, A. R. & Moral, S. An interactive approach for Bayesian network learning using domain/expert knowledge. *Int. J. Approx. Reason.* **54**, 1168–1181. <https://doi.org/10.1016/j.ijar.2013.03.009> (2013).
31. Cano, A., Masegosa, A. R. & Moral, S. A method for integrating expert knowledge when learning bayesian networks from data. *IEEE Trans. Syst. Man Cybern. Part B (Cybernetics)* **41**, 1382–1394. <https://doi.org/10.1109/TSMCB.2011.2148197> (2011).
32. Amirkhani, H., Rahmati, M., Lucas, P. J. F. & Hommersom, A. Exploiting experts' knowledge for structure learning of Bayesian networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2154–2170. <https://doi.org/10.1109/TPAMI.2016.2636828> (2017).
33. Zheng, X., Aragam, B., Ravikumar, P. & Xing, E. P. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems* (2018).
34. Sousa, H. S., Prieto-Castrillo, F., Matos, J. C., Branco, J. M. & Lourenço, P. B. Combination of expert decision and learned based Bayesian networks for multi-scale mechanical analysis of timber elements. *Expert Syst. Appl.* **93**, 156–168 (2018).
35. Ankan, A. & Textor, J. Expert-in-the-loop causal discovery: Iterative model refinement using expert knowledge. In *The 41st Conference on Uncertainty in Artificial Intelligence* (2025).
36. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, 2009).
37. Pearl, J. *Causal Inference in Statistics: A Primer* (Wiley, 2016).
38. Pearl, J. *Causality* (Cambridge University Press, 2009).
39. Cinelli, C., Forney, A. & Pearl, J. A crash course in good and bad controls. *Sociol. Methods Res.* **53**, 1071–1104 (2024).
40. Reneses, M. et al. Deliverable 1.7: Open report on victim and offender profile description report. Tech. Rep., H2020 RAYUELA (2022).
41. Reneses, M. et al. Deliverable 1.3: Open report on interview results. Tech. Rep., H2020 RAYUELA (2022).
42. Reneses, M. et al. Deliverable 1.5: Open report on case study results. Tech. Rep., H2020 RAYUELA (2022).
43. Sorrentino, A., Baldry, A., Farrington, D. & Blaya, C. Epidemiology of cyberbullying across Europe: Differences between countries and genders. *Educ. Sci. Theory Pract.* **19**, 74–91 (2019).
44. Hinduja, S. & Patchin, J. W. Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant Behav.* **29**, 129–156. <https://doi.org/10.1080/01639620701457816> (2008).
45. Juvonen, J. & Gross, E. F. Extending the school grounds? Bullying experiences in cyberspace. *J. School Health* **78**, 496–505. <https://doi.org/10.1111/j.1746-1561.2008.00335.x> (2008).
46. Barlett, C. P., Madison, C. S., Heath, J. B. & DeWitt, C. C. Please browse responsibly: A correlational examination of technology access and time spent online in the Barlett gentile cyberbullying model. *Comput. Hum. Behav.* **92**, 250–255. <https://doi.org/10.1016/j.chb.2018.11.013> (2019).
47. Handono, S. G., Laeheim, K. & Sittichai, R. Factors related with cyberbullying among the youth of Jakarta, Indonesia. *Child. Youth Serv. Rev.* **99**, 235–239. <https://doi.org/10.1016/j.childyouth.2019.02.012> (2019).
48. Ortega Barón, J., Postigo, J., Iranzo, B., Buelga, S. & Carrascosa, L. Parental communication and feelings of affiliation in adolescent aggressors and victims of cyberbullying. *Soc. Sci.* **8**, 3. <https://doi.org/10.3390/socsci8010003> (2019).
49. Aizenkot, D. Social networking and online self-disclosure as predictors of cyberbullying victimization among children and youth. *Child. Youth Serv. Rev.* **119**, 105695 (2020).
50. Akgül, G. & Artar, M. Cyberbullying: Relationship with developmental variables and cyber victimization. *Scand. J. Child Adolesc. Psychiatry Psychol.* **8**, 25. <https://doi.org/10.21307/sjcap-2020-004> (2020).
51. Elipse, P., de la Oliva Muñoz, M. & Del Rey, R. Homophobic bullying and cyberbullying: Study of a silenced problem. *J. Homosex.* **65**, 672–686. <https://doi.org/10.1080/00918369.2017.1333809> (2018).
52. Minton, S. J. Homophobic bullying: Evidence-based suggestions for intervention programmes. *J. Aggress. Confl. Peace Res.* **6**, 164–173. <https://doi.org/10.1108/JACPR-10-2013-0027> (2014).
53. Toomey, R. B. & Russell, S. T. The role of sexual orientation in school-based victimization: A meta-analysis. *Youth Soc.* **48**, 176–201. <https://doi.org/10.1177/0044118X13483778> (2016).
54. Hong, J. S., Zhang, S., Wright, M. F. & Wachs, S. Racial and ethnic differences in the antecedents of cyberbullying victimization in early adolescence: An ecological systems framework. *J. Early Adolesc.* **43**, 59–89. <https://doi.org/10.1177/02724316211042939> (2023).
55. Mazzone, A., Thornberg, R., Stefanelli, S., Cadei, L. & Caravita, S. C. “Judging by the cover”: A grounded theory study of bullying towards same-country and immigrant peers. *Child. Youth Serv. Rev.* **91**, 403–412. <https://doi.org/10.1016/j.childyouth.2018.06.029> (2018).
56. Barlett, C. & Coyne, S. M. A meta-analysis of sex differences in cyber-bullying behavior: The moderating role of age. *Aggress. Behav.* **41**, 513–513. <https://doi.org/10.1002/ab.21555> (2015).

57. Li, Q. Bullying in the new playground: Research into cyberbullying and cyber victimisation. *Australas. J. Educ. Technol.* **23**, 435–454. <https://doi.org/10.14742/ajet.1245> (2007).
58. Maher, D. Cyberbullying: An ethnographic case study of one Australian upper primary school class. *Youth Stud. Aust.* **27**, 50–57 (2008).
59. Shohoudi, A., Leduc, K., Shohoudi, A. & Talwar, V. Examining cross-cultural differences in youth's moral perceptions of cyberbullying. *Cyberpsychol. Behav. Soc. Netw.* **22**, 243–248. <https://doi.org/10.1089/cyber.2018.0339> (2019).
60. Zhu, C., Huang, S., Evans, R. & Zhang, W. Cyberbullying among adolescents and children: A comprehensive review of the global situation, risk factors, and preventive measures. *Front. Public Health* **9**, 634909 (2021).
61. Moreno-Ruiz, D., Martínez-Ferrer, B. & García-Bacete, F. Parenting styles, cyberaggression, and cybervictimization among adolescents. *Comput. Hum. Behav.* **93**, 252–259 (2019).
62. Tsitsika, A. K. et al. Online social networking in adolescence: Patterns of use in six European countries and links with psychosocial functioning. *J. Adolesc. Health* **55**, 141–147 (2014).
63. Thumronglaohapun, S. et al. Awareness, perception and perpetration of cyberbullying by high school students and undergraduates in Thailand. *PLoS One* **17**, e0267702 (2022).
64. Lozano-Blasco, R., Robres, A. Q. & Cosculluela, C. L. Sex, age and cyber-victimization: A meta-analysis. *Comput. Hum. Behav.* **139**, 107491. <https://doi.org/10.1016/j.chb.2022.107491> (2023).
65. Kircaburun, K. et al. Problematic online behaviors among adolescents and emerging adults: Associations between cyberbullying perpetration, problematic social media use, and psychosocial factors. *Int. J. Mental Health Addict.* **17**, 891–908 (2019).
66. Buelga, S., Martínez-Ferrer, B. & Cava, M. J. Differences in family climate and family communication among cyberbullies, cybervictims, and cyber bully–victims in adolescents. *Comput. Hum. Behav.* **76**, 164–173 (2017).
67. Keijsers, L. & Poulin, F. Developmental changes in parent-child communication throughout adolescence. *Dev. Psychol.* **49**, 2301–2308 (2013).
68. Yurdakul, Y. & Ayhan, A. B. The effect of the cyberbullying awareness program on adolescents' awareness of cyberbullying and their coping skills. *Curr. Psychol.* **42**, 24208–24222 (2023).
69. Carvalho, A. M. Scoring functions for learning Bayesian networks. *Inesc-id Tec. Rep.* **12**, 1–48 (2009).
70. Cooper, G. F. & Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* **9**, 309–347. <https://doi.org/10.1007/BF00994110> (1992).
71. Geiger, D., Verma, T. & Pearl, J. D-separation: From theorems to algorithms. In *Uncertainty in Artificial Intelligence, Machine Intelligence and Pattern Recognition*, Vol. 10, 139–148 (North-Holland, 1990). <https://doi.org/10.1016/B978-0-444-88738-2.50018-X>
72. Jabbar, H. K. & Khan, R. Z. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). In *Computer Science, Communication and Instrumentation Devices, AET*, 978–981 (Research Publishing Services, 2014). https://doi.org/10.3850/978-981-09-5247-1_017
73. Maldonado, G. & Greenland, S. Estimating causal effects. *Int. J. Epidemiol.* **31**, 422–429 (2002). <https://doi.org/10.1093/ije/31.2.422>
74. Pearl, J. The do-calculus revisited. [arXiv:1210.4852](https://arxiv.org/abs/1210.4852) (2012).
75. Pearl, J. *Direct and Indirect Effects* 1st edn, 373–392 (Association for Computing Machinery, 2022).
76. MacKay, D. J. *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, 2003).
77. Vishnusai, Y., Kulakarni, T. R. & Sowmya Nag, K. Ablation of artificial neural networks. In *International Conference on Innovative Data Communication Technologies and Application*, 453–460 (Springer, 2019).
78. Reneses, M., Parder, M.-L., Riberas-Gutiérrez, M. & Bueno-Guerra, N. Cyberbullying and cyberhate as an overlapping phenomenon among adolescents in Estonia and Spain: Cross-cultural differences and common risk factors. *Child. Youth Serv. Rev.* **2025**, 108285. <https://doi.org/10.1016/j.childyouth.2025.108285> (2025).
79. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (Chapman and Hall/CRC, 2020).
80. Xu, W., Zhao, B. & Jin, C. A meta-analysis of the relationship between personality traits and cyberbullying. *Aggress. Violent Behav.* **79**, 101992 (2024).

Acknowledgements

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

Author contributions

A.B.-R. Investigation, data curation, software, writing the original draft, writing, review and editing. M.R. Investigation, methodology, data curation, writing, review and editing. J.P. Investigation, data curation, software, writing—original draft, writing, review and editing. G.V. Investigation, data curation, software, writing—original draft. E.A. Supervision, validation, methodology, writing, review and editing. G.L.L. Conceptualization, funding acquisition, investigation, supervision, validation, writing—review and editing. M.C. Conceptualization, methodology, formal analysis, investigation, software, validation, writing, review and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-16149-4>.

Correspondence and requests for materials should be addressed to M.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025