Usando un juego serio e IA causal para estudiar el ciberbullying

Jaime Pérez
Institute for Research in Technology (IIT)
ICAI Engineering School
Universidad Pontificia Comillas
jperezs@comillas.edu

Mario Castro
Institute for Research in Technology (IIT)
ICAI Engineering School
Universidad Pontificia Comillas

Gregorio López
Institute for Research in Technology (IIT)
ICAI Engineering School
Universidad Pontificia Comillas

Edmond Awad
The Oxford Uehiro Centre for Practical Ethics
University of Oxford

María Reneses
Faculty of Human and Social Sciences
Universidad Pontificia Comillas

***Resumen*—El ciberbullying entre los menores de edad es una preocupación creciente en nuestra sociedad digital, que requiere estrategias eficaces de prevención e intervención. Las metodologías tradicionales de recolección de datos en este ámbito pueden ser intrusivas y proporcionar resultados limitados. En este artículo se explora un enfoque innovador que utiliza un juego serio —diseñado con fines que van más allá del entretenimiento— como una herramienta de investigación atractiva y no intrusiva para estudiar problemáticas sociales delicadas. Además, proponemos el uso de técnicas de IA causal (modelos causales gráficos probabilísticos) para analizar los datos obtenidos. Este marco analítico proporciona resultados interpretables e intuitivos, aumenta la transparencia y fomenta un discurso científico abierto. Los resultados obtenidos indican que el uso de juegos serios puede desempeñar un papel más relevante en el estudio del ciberbullying que estudios basados en datos demográficos o de perfilado, lo que evidencia su potencial como metodología de investigación alternativa. Finalmente, demostramos cómo nuestro enfoque nos permite analizar perfiles de riesgo e identificar de estrategias de intervención para mitigar este ciberdelito. Mediante la integración de la IA causal con los juegos serios, proporcionamos un marco analítico completo para explorar y mitigar los retos que plantea el ciberbullying, contribuyendo a crear entornos en línea más seguros para los menores y demostrando el potencial de este enfoque para investigar problemáticas sociales complejas.**

***Index Terms*—Juegos Serios, Ciberbullying, Redes Bayesianas, Ciencias Sociales computacionales**

Tipo de contribución: *Investigación original*

I. Introducción

El auge de la era digital ha provocado un aumento del uso de Internet por parte de los niños. En 2017, los menores de 18 años representaban casi un tercio de todos los usuarios de Internet del mundo, según el estudio de UNICEF "Los niños en un mundo digital"[1]. Por desgracia, el acceso incontrolado a Internet también expone a los niños a nuevos peligros. Aproximadamente el 10 % de los niños europeos sufre ciberbullying (CB) cada mes [2], y casi el 50 % ha experimentado un incidente relacionado con el CB al menos una vez [3].

La investigación sobre CB en las ciencias sociales tradicionales suele involucrar metodologías difíciles de escalar e invasivas como los cuestionarios, lo que dificulta aún más el acercamiento a sectores de la población como los menores. Esta barrera exige enfoques novedosos que se adecuen más a su bagaje cultural. Los Juegos Serios (JS) son una alternativa atractiva para estudiar el CB, pudiendo utilizarse como herramienta educativa y de investigación, y promoviendo un enfoque preventivo [4]. Los JS se diseñan explícitamente con un propósito principal más allá del entretenimiento (por ejemplo, la formación o el aprendizaje de nuevas habilidades, la transmisión de valores, la concienciación social) [5].

Los juegos tratan de lograr la inmersión del jugador mediante el diseño, lo que facilita la apertura y deshinibición de los participantes y presenta una oportunidad única para acceder a grupos demográficos más amplios y diversos. Los JS sirven como herramienta de investigación no invasiva, ofreciendo a los participantes un entorno agradable e interactivo, lo que es fundamental cuando se abordan cuestiones sociales delicadas e intrincadas como el ciberbullying. Aunque los JS son ampliamente reconocidos por su valor educativo, su potencial como herramientas de investigación social permanece en gran medida inexplorado [6]. Sin embargo, ha habido notables ejemplos recientes en este contexto [7], [8], [9]. Este enfoque de investigación basado en juegos tiene un potencial significativo para mejorar la amplitud y profundidad de la investigación social y conductual [10].

La investigación con JS dependen en gran medida del análisis de datos para modelar al jugador o a la población. Sin embargo, cuando se investigan temas delicados como el CB, dichos análisis suponen un resto técnico . Las metodologías tradicionales suelen emplear técnicas estadísticas de búsqueda de correlaciones, pasan por alto los sesgos potenciales introducidos por los factores de confusión y las variables de colisión, o se centran únicamente en los test de hipótesis de valor p, que a menudo resultan problemáticos [11]. Alternativamente, algunos investigadores recurren a las impresionantes capacidades predictivas de los algoritmos de aprendizaje automático [12], que, aunque potentes, pueden basar sus predicciones en patrones irreales o relaciones espurias que no son fácilmente interpretables ni explicables.

En este artículo, proponemos un enfoque diferente para analizar los datos provenientes de investigación con JS, una perspectiva basada en la causalidad, más adecuada para tratar

ISBN:978-84-09-62140-8

la delicada naturaleza de la cuestión. En concreto, utilizamos redes bayesianas causales [13], también conocidas como modelos gráficos causales probabilísticos (MGCP) [14], que proporcionan un marco sofisticado pero explicable e intuitivo para modelar relaciones causales complejas. Estos modelos permiten a los investigadores incorporar conocimiento experto (definiendo la estructura de la red) o conocimiento cuantitativo bien conocido (definiendo distribuciones de probabilidad a priori).

Los MGCP se han aplicado con éxito en diversos campos como la Biología [15], la Psicología [16], las Ciencias Sociales [17], la Econometría [18], o la Epidemiología [19]. También se han utilizado para analizar datos de CB en algunos estudios [20], [21]. La adopción de los MGCP en el análisis de los datos derivados de los JS promete una comprensión más precisa y matizada de las relaciones causales. Este enfoque también es beneficioso cuando se trabaja con datos limitados o ruidosos, ya que trata de reducir los posibles sesgos al tiempo que ayuda a diseñar intervenciones eficaces [22].

Este trabajo forma parte del proyecto de investigación europeo H2020 RAYUELA[1], cuyo objetivo es aprovechar el atractivo natural de un juego serio para recopilar datos y educar a los menores acerca de algunos ciberdelitos. Los datos utilizados en este artículo se han recopilado mediante pilotos experimentales en los que han participado menores de edad europeos. Los jugadores se sumergen en una aventura visual interactiva, en la que toman decisiones relacionadas con la ciberdelincuencia. El objetivo principal es comprender mejor los factores que influyen en los comportamientos de riesgo en línea de una forma amigable, segura y no invasiva.

En concreto, este artículo se centra en el estudio del fenómeno del acoso en CB. En este contexto, surgen de forma natural dos preguntas de investigación (PI):

PI1: ¿Qué factores humanos o tecnológicos están más asociados con el riesgo de ser un agresor de ciberbullying?

PI2: ¿Qué combinación de factores conforman los perfiles de riesgo identificados?

Los principales objetivos de este artículo son (i) demostrar el potencial de los JS como una valiosa herramienta de investigación en ciencias sociales, (ii) ampliar las metodologías existentes en este campo integrando técnicas de IA causal, y (iii) contribuir a una comprensión más exhaustiva del fenómeno del CB. De este modo, nuestro trabajo puede ayudar a fundamentar estrategias de prevención e intervención más eficaces para abordar el CB.

II. Metodología

II-A. Descripción del Juego Serio

Como se ha descrito en la sección anterior, el objetivo principal del proyecto H2020 RAYUELA es estudiar en profundidad los factores humanos y tecnológicos que contribuyen a determinados tipos de ciberdelincuencia que afectan a los menores [23]. Este objetivo se logra a través de un enfoque único que aprovecha el lenguaje de los juegos, proporcionando una plataforma para el aprendizaje y el modelado de comportamientos de una manera atractiva y no invasiva. El proyecto fue realizado por un equipo interdisciplinar, que incluía psicólogos y antropólogos expertos en CB y ciberdelincuencia.

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

El JS es un videojuego para PC, desarrollado por el equipo Tecnalia (parte del consorcio RAYUELA). Se trata de una aventura gráfica point-and-click de narrativa interactiva, en la que los jugadores toman decisiones que condicionan la progresión y el desenlace de la narración. El juego está ambientado en un instituto y presenta escenarios relacionados con algunos delitos cibernéticos que afectan los menores. Consta de 6 aventuras, cada una de las cuales aborda un tipo de cibercrimen o una faceta específica. La duración total de las sesiones de juego es de aproximadamente 1,5 horas. La figura 1 muestra una captura de pantalla de una decisión que los jugadores tienen que tomar en el juego. La figura 2 muestra una pregunta final relacionada con el nivel de ”honestidad”durante el juego.

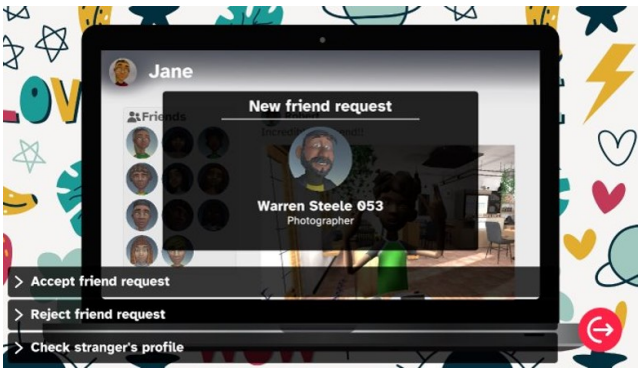


Figura 1: Captura de pantalla de una decisión que los jugadores deben tomar en el juego serio de RAYUELA.

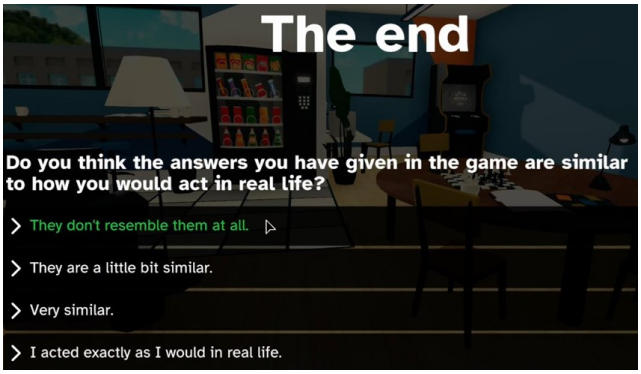


Figura 2: Captura de pantalla de la pregunta sobre ”honestidad” en el juego serio de RAYUELA.

Las preguntas/decisiones del juego fueron cuidadosamente diseñadas y discutidas por el equipo de RAYUELA para analizar los patrones de los jugadores e identificar a los más propensos a cometer o experimentar los ciberdelitos estudiados, y se perfeccionaron tras las pruebas realizadas en los primeros pilotos experimentales. Los ciberdelitos considerados son CB, online grooming, fake news y ciberseguridad. Aunque en este artículo sólo abordamos el CB. Además, también se recogieron variables demográficas y psicológicas durante la sesión de juego para validar el resultado. De este modo, pudimos medir ciertos factores potenciales de CB tanto *in-game* como *out-of-game* (Tabla I), lo que nos permitió verificar la validez de las

situaciones presentadas en el videojuego. La figura 1 es un ejemplo de medición de una variable *in-game*.

Tabla I: Factores/variables de CB identificados, ilustrando si cada variable se midió *in-game* o *out-of-game*.

Tipo	CB Factor/Variable	Medido in-game	Medido out-of-game
Ambiental	Aislamiento/falta de apoyo social	✓	✓
	Comunicación Familiar	✓	✓
Personal	Agresión previa CB	✓	✓
	Victimización previa CB	✓	✓
	Baja autoestima	✓	✓
	Dificultad en hacer amigos cara a cara	✓	✗
	Edad	✗	✓
	Género	✗	✓
	Orientación sexual	✗	✓
	Antecedentes migratorios	✗	✓
Tecnológico	Perfil público en redes sociales y publicar excesiva información	✓	✗
	Tiempo en Internet	✓	✓
	Contraseña débil o compartir contraseña	✓	✗

II-B. Recogida de datos

Todos los procedimientos experimentales y de recopilación de datos fueron aprobados por los comités éticos del consorcio RAYUELA en cada uno de los países implicados en la investigación. Además, los expertos legales de RAYUELA también tomaron las medidas necesarias para garantizar que la recogida, el almacenamiento y la difusión de los datos se ajustaran al GDPR europeo. En cada sesión, los investigadores y profesores explicaron a los participantes el proyecto y su objetivo principal, así como los datos que se iban a recopilar. Dependiendo de la edad, los participantes o sus padres firmaron un consentimiento informado para poder participar.

Este dataset se creó con las entradas de los pilotos experimentales que utilizaron el JS desarrollado en escuelas e institutos europeos. En concreto, se recogieron datos de Grecia, Bélgica y España. Recogimos 224 respuestas de estudiantes de entre 12 y 16 años (Media=13,6, DT=1,37), de los que el 53 % se identificaban como varones, el 44 % como mujeres y el 1 % como no binarios. La tabla II muestra estadísticas resumidas de los datos recogidos, los posibles valores de cada variable y su porcentaje de ocurrencia (es decir, la probabilidad marginal).

Datos previos al juego

Antes de comenzar la sesión de juego, todos los participantes se registraron y rellenaron unos cuestionarios demográficos y psicológicos. Los expertos en CB de RAYUELA eligieron meticulosamente cada una de las variables demográficas y psicológicas que debían recogerse, en consonancia con las investigaciones previas y propias, que sugieren que estos factores pueden influir sustancialmente en la propensión y la respuesta al CB [24], [25], [26]. Las siguientes variables se obtuvieron durante la primera fase de cada piloto, antes de jugar al JS:

- Variables demográficas: *Edad*, *género*, *orientación sexual*, *antecedentes migratorios*, y *horas diarias en Internet*. Estas variables se tuvieron en cuenta para comprender los diversos contextos demográficos de los menores participantes, junto con una medida de su relación con la tecnología.
- Variables psicológicas y ambientales: *Apoyo social* (aislamiento), *apoyo familiar*, y *autoestima*. Estas variables se eligieron para proporcionar una referencia que permitiera comprender el estado emocional, social y psicológico de los menores antes de jugar al juego. Se obtuvieron de los siguientes cuestionarios validados: la *Escala de Autoestima de Rosenberg* [27] y la *Escala multidimensional de apoyo social percibido* [28].

Datos del juego

Los datos recogidos exclusivamente a través del JS abarcan dos elementos esenciales: (i) las decisiones tomadas por el jugador en cada pregunta del juego y (ii) una pregunta de ”honestidad.ª”l finalizar la partida (”*¿Crees que las respuestas que has dado en el juego se parecen a cómo actuarías en la vida real?*”) que se muestra en la figura 2. Aunque se les pidió a los participantes que jugaran tal y como actuarían en la vida real al principio de las sesiones, esta pregunta al finalizar actúa como calibración y control sobre su ”honestidad”mientras juegan, ya que el propio formato del juego puede animar a algunos jugadores a tomar decisiones más aventuradas para explorar opciones narrativas.

Datos posteriores al juego

Después de cada sesión de juego, los menores rellenaron un cuestionario sobre sus experiencias pasadas relacionadas con el CB, que sirvió como ”verdad de base”para el análisis de los datos. El cuestionario validado fue el *Cuestionario del Proyecto Europeo de Intervención contra el Ciberacoso*. [29]. El cuestionario cuantifica si la persona ha sufrido o cometido (o ambos) actos relacionados con el CB.

II-C. Colaboración con expertos en ciberbullying de RAYUELA

El conocimiento experto procede de los miembros del consorcio del proyecto RAYUELA. El paquete de trabajo 1 dentro del proyecto se ocupó de crear una base de conocimientos sobre los factores que impulsan la ciberdelincuencia entre los jóvenes. Para ello, este equipo llevó a cabo una investigación que pretendía comprender tanto la patología como la fisiología de los comportamientos en línea, caracterizando a las víctimas y a los delincuentes de las formas de ciberdelincuencia consideradas, así como el modus operandi.

Este equipo estaba formado por miembros de la Universidad Pontificia Comillas (España), la Universidad de Gante (Bélgica), la Universidad de Tartu (Estonia), el Colegio Universitario de Limburgo (Bélgica), el Instituto de Política de Bratislava (Eslovaquia), Ellinogermaniki Agogi (Grecia), la Polícia Judiciária (Portugal), la Policía Local Valenciana (España), el Servicio de Policía de Irlanda del Norte (Reino Unido), la Policía de Estonia y la Junta de la Guardia de Fronteras (Estonia).

En relación con el delito de CB este equipo realizó un total de 33 entrevistas (8 delincuentes, 12 víctimas y 13 expertos) [24] y analizó 46 sentencias judiciales [25]. Como resultado de esta investigación, el equipo adquirió un profundo conocimiento de la cuestión del CB, que se ha utilizado en varias ocasiones a lo largo del resto del proyecto y se ha

Tabla II: Variables demográficas y procedentes de cuestionarios psicológicos. Aquí se muestran los posibles valores de cada variable y su probabilidad marginal (es decir, el porcentaje de observación). La etiqueta «missing» agrega los valores incorrectos y los ítems que los encuestados han decidido no responder.

Variable	Valores de respuesta	Probabilidad marginal
Género	Hombre	62.9 %
	Mujer	35.1 %
	No binario	1 %
	«missing»	1 %
Edad	12	18.8 %
	13	4.4 %
	14	26.8 %
	15	33 %
	16	17 %
Orientación sexual	Heterosexual	55.3 %
	No Heterosexual	5.4 %
	«missing»	39.3 %
Antecedentes migratorios	No	71.4 %
	Mis padres nacieron en otro país	8.6 %
	Yo nací en otro país	20 %
Autoestima	Bajo	37.5 %
	Medio	41.5 %
	Alto	21 %
Apoyo social	Bajo	3.6 %
	Medio	33.5 %
	Alto	62.9 %
Apoyo familiar	Bajo	7.6 %
	Medio	24.5 %
	Alto	67.9 %
Horas diarias de Internet	Menos de 1 h	8.9 %
	1-2 h	18.7 %
	2-3 h	21.4 %
	3-4 h	33 %
	Más de 4 h	15.6 %
	«missing»	2.4 %

documentado en los citados informes técnicos. Para interactuar con los expertos a la hora de construir el modelo causal, mantuvimos sesiones de discusión con algunos miembros del equipo de la Universidad Pontificia Comillas, ya que eran los líderes de este paquete de trabajo.

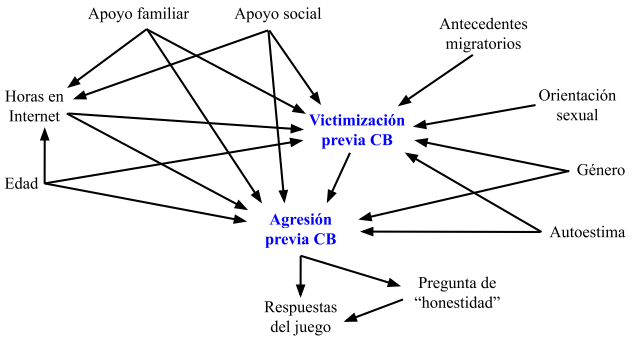
II-D. Análisis basado en la causalidad

Para comprender mejor el CB y la interconexión de sus variables, hemos aplicado una metodología basada en la causalidad para realizar un análisis exhaustivo.

Los enfoques basados en la causalidad son especialmente beneficiosos cuando se estudian temáticas sociales delicadas como el CB, ya que pueden revelar los procesos causales subyacentes en lugar de limitarse a hacer predicciones o buscar correlaciones. Comprender las relaciones causales es fundamental para identificar los puntos de intervención que permitan optimizar las estrategias de prevención y detección y, en última instancia, mitigar las consecuencias dañinas del CB.

Hemos utilizado los MGCP, también conocidos como redes Bayesianas causales, un marco que permite codificar relaciones estadísticas y causales entre variables utilizando un grafo acíclico directo (DAG). Los nodos presentes en un MGCP pueden representar variables observables o latentes unidas por flechas que indican cuándo existe una relación causal. Los MGCP permiten mejorar la interpretabilidad y la transparencia de los presupuestos adoptados por los investigadores. Además, estas redes pueden detectar eficazmente factores de confusión y colisión, para mitigar posibles sesgos. También pueden manejar análisis contrafactuales y simular intervenciones basadas en los datos disponibles. La construcción de una estructura MGCP (DAG) coherente que se ajuste a la realidad es crucial. Aunque existen algoritmos basados en datos para inferir estas estructuras de red a partir de datos, la incorporación de conocimientos expertos es esencial, especialmente en escenarios con datos limitados. Por lo tanto, en este artículo hemos utilizamos una estructura MGCP proporcionada por expertos en CB de RAYUELA (Figura 3).

Figura 3: Estructura de los MGCP propuestos por los expertos en CB del proyecto RAYUELA. La variable de interés (*Agresión previa CB*) está resaltada en azul.



Una vez seleccionada la estructura del MGCP, entrenamos sus parámetros (es decir, estimamos las tablas de probabilidad condicional) utilizando los datos disponibles. A continuación, realizamos dos análisis causales para *interrogar* al modelo:

Cuantificación de fuerza de la flecha

Este análisis pretende responder a la PI1: *¿Qué factores humanos o tecnológicos están más asociados con el riesgo de ser un agresor de ciberbullying?* Para ello, analizamos el poder de influencia que cada variable del MGCP tiene sobre la variable objetivo, que en este caso es *Agresión previa CB*, obtenida del cuestionario posterior al juego [29]. Este análisis permite conocer las variables más relevantes que influyen en el resultado del modelo, considerándolas individualmente. Es decir, sin tener en cuenta las combinaciones de variables. Este método se basa en el trabajo de Koiter [30]. Consiste principalmente en medir la similitud entre múltiples distribuciones de probabilidad de la variable de interés en función de los estados de los nodos padre. La métrica elegida para calcular esta similitud entre distribuciones es la distancia Jensen-Shannon [31] normalizada entre 0 y 1 (para mejorar la interpretabilidad).

Análisis multifactorial de perfiles

Este análisis pretende responder a PI2: *¿Qué combinación de factores conforman los perfiles de riesgo identificados?* Para ello, examinamos cómo simular múltiples observaciones de evidencias simultáneas pueden influir en el resultado. En otras palabras, pretendemos identificar qué combinaciones de características son las más comunes en los perfiles de riesgo de cometer delitos de CB. Además, utilizaremos este análisis para comparar la relevancia de las variables procedentes del juego frente a las procedentes de variables demográficas o cuestionarios psicológicos. Las variables que no proceden del juego se engloban en el término ”perfilado”. La elaboración de perfiles de jugadores consiste en el análisis

o la categorización utilizando únicamente variables estáticas que no están necesariamente relacionadas de forma directa con la jugabilidad [32].

El método consiste en simular observaciones de todas las combinaciones posibles de características por fuerza bruta y analizar las características comunes de los perfiles de riesgo identificados. Para filtrar qué perfiles consideramos de riesgo, analizamos las diferencias entre las distribuciones de probabilidad a priori y a posteriori utilizando el factor de Bayes (FB) en conjunción con la escala de Jeffreys [33]. El FB puede expresarse formalmente como en la ecuación (1), donde $P(Y)$ es la probabilidad resultante, y E representa la evidencia observada. La escala de Jeffreys traduce el orden de magnitud del FB en un juicio cualitativo que nos permite decidir la cantidad de evidencia necesaria para apoyar una hipótesis/modelo y no otro. Para esta investigación, consideraremos que los valores de $BF > 10^{1/2}$ ya constituyen pruebas suficientes para considerarlo un perfil de riesgo. Hemos utilizado una probabilidad previa del 10 % de observar *Victimización previa CB*. Por lo tanto, una evidencia *sustancial* y *fuerte* se producirá con probabilidades posteriores de $\sim 26\,\%$ y $\sim 53\,\%$, respectivamente (Ecuación 1).

$$
\begin{aligned}
\text{Prior odds } &= \frac{P(Y)}{1-P(Y)} \\
\text{Posterior odds } &= \frac{P(Y \mid E)}{1-P(Y \mid E)} \\
\text{Factor Bayes } (FB) &= \frac{\text{Posterior odds}}{\text{Prior odds}} \\
&= \frac{P(Y \mid E)(1-P(Y))}{P(Y)(1-P(Y \mid E))} \\
P(Y) &= 0{,}1 \text{ (probabilidad prior)} \\
FB &= 10^{1/2} \text{ (Evidencia } \textit{Sustancial}) \Rightarrow P(Y \mid E) \approx 0{,}26 \\
FB &= 10 \text{ (Evidencia } \textit{Fuerte}) \Rightarrow P(Y \mid E) \approx 0{,}53
\end{aligned}
\tag{1}
$$

III. Resultados

Esta sección presenta los resultados analíticos obtenidos aplicando la metodología propuesta a los datos de los pilotos experimentales del proyecto RAYUELA.

III-A. Cuantificación de fuerza de la flecha

La figura 4 presenta los resultados del análisis relativos a la variable de interés (*Agresión previa CB*). La pregunta del juego *Aventura 1 - Pregunta 1 - Compartir foto* actúa como variable de control, generando respuestas aleatorias de los jugadores. Por lo tanto, las variables con métricas similares o inferiores a esta primera pregunta se consideran irrelevantes a efectos prácticos. Los resultados indican que las variables relevantes se derivan únicamente de las respuestas del JS, asumiendo la veracidad del MGCP construido y analizando cada variable de forma individual. Este resultado sugiere que el JS podría ser una herramienta de investigación fiable para la variable de interés (*Agresión previa CB*). Por otro lado, las variables demográficas y provenientes de test psicológicos muestran una significación mínima, cuando se examinan individualmente.

La figura 5 aclara la importancia de la variable *ganadora* presentando las probabilidades condicionales de *Agresión previa CB* basadas en todos los valores posibles de *Aventura 3 Pregunta 7: Ayuda a Pol*. Esta figura demuestra que los jugadores que eligen la respuesta 4 tienen un 26,5 % de probabilidad estimada de haber cometido ciberagresiones en el pasado, independientemente de otras variables. Por el contrario, los que seleccionan la respuesta 1 tienen apenas un 7 % de probabilidades. Esta pregunta del juego en concreto trata de una situación en la que un personaje está siendo víctima de ciberacoso. El jugador tiene que dar su opinión sobre el mejor curso de acción, que podría ser contárselo a los padres, denunciarlo a la red social, no denunciarlo para no ser acusado de chivato, o no denunciarlo porque cree que será inútil.

III-B. Análisis multifactorial de perfiles

La figura 6 ilustra el resultado de este análisis mostrando la máxima probabilidad posterior de observar *Agresión previa CB* para cada número de evidencias fijadas. En otras palabras, estamos representando la probabilidad de que el perfil más arriesgado fijando una variable, luego 2, luego 3, y así sucesivamente. Esta metodología se aplica tanto a las variables procedentes de las preguntas del juego como a las variables de perfilado. La fig. 6 también muestra los dos umbrales obtenidos con la escala de Jeffreys (Ecuación 1). Podemos observar que a medida que aumenta el número de evidencias fijadas, la probabilidad posterior del perfil más arriesgado aumenta progresivamente. Sin embargo, también hay que señalar que a medida que aumenta el número de evidencias fijadas, el número de jugadores que cumplen estos criterios también disminuye progresivamente y, por lo tanto, tenemos menos evidencias sobre estos perfiles.

Estos resultados confirman la conclusión del primer análisis (*Cuantificación de fuerza de la flecha*) de que las variables derivadas del juego son más eficaces para distinguir a los jugadores que tienen un historial de agresiones relacionadas con el CB de los que no. Esta conclusión sugiere que el apetito de riesgo de una persona está mejor definido por sus acciones en escenarios específicos, que por sus características personales o psicológicas. Por lo tanto, los JS podrían ser una alternativa eficaz para estudiar cómo se comportan los jugadores en situaciones de la vida real, siempre que se respeten las limitaciones éticas y que el juego esté bien diseñado para la función deseada.

Para analizar las características compartidas más comunes en los perfiles de riesgo, necesitamos elegir el número de evidencias fijadas a partir del cual empezamos a considerar los perfiles de riesgo. Es decir, cuando las probabilidades de observar *Agresión previa CB* empiezan a superar el umbral del criterio de Jeffreys ($\sim 26\,\%$). En nuestro caso, esto sucede para cinco evidencias fijadas (Fig. 6). Podemos observar en la figura 7 que la característica más común entre los perfiles de riesgo detectados es *Victimización previa CB*, con una prevalencia superior al 80 %. Esto sugiere que los individuos que han sido víctimas de CB en el pasado muestran una propensión relevante de cometer CB en el futuro. Esta relación es conocida desde hace tiempo en la literatura científica [34], [35]. También podemos afirmar que *alta autoestima*, *hombre*, *14 años*, *alto apoyo familiar*, y *alto número de horas diarias en Internet* son también características bastante prevalentes en los perfiles de riesgo identificados.

Figura 4: Cuantificación de fuerza de la flecha: Analizamos la influencia de cada variable en la salida del MGCP (*Agresión previa CB*). Esto consiste en medir la distancia entre las distribuciones de probabilidad de cada variable marginalizando *Agresión previa CB* en sus posibles valores. La distancia Jensen-Shannon normalizada (0-1) sirve de métrica. "*Aventura 1 - Pregunta 1 - Compartir foto*.ªctúa como una variable de control que genera respuestas aleatorias, haciendo que las variables con métricas similares o inferiores sean irrelevantes a efectos prácticos.

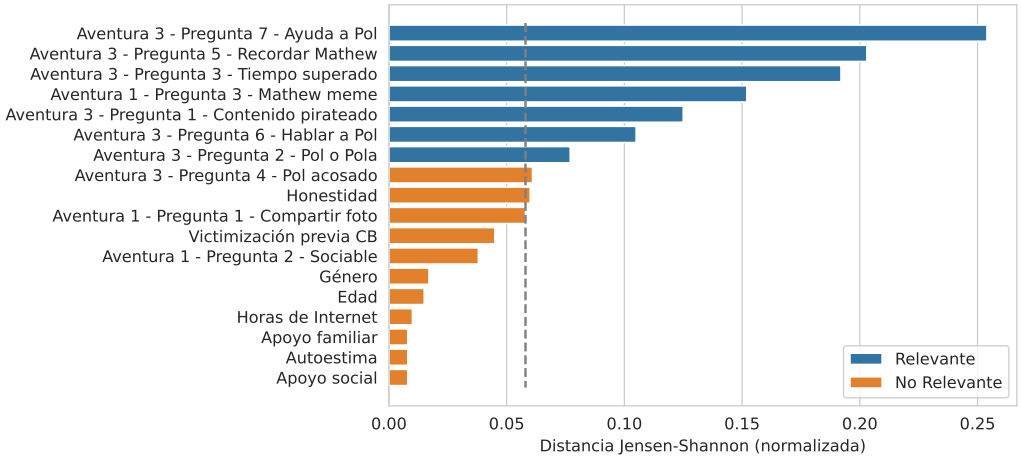
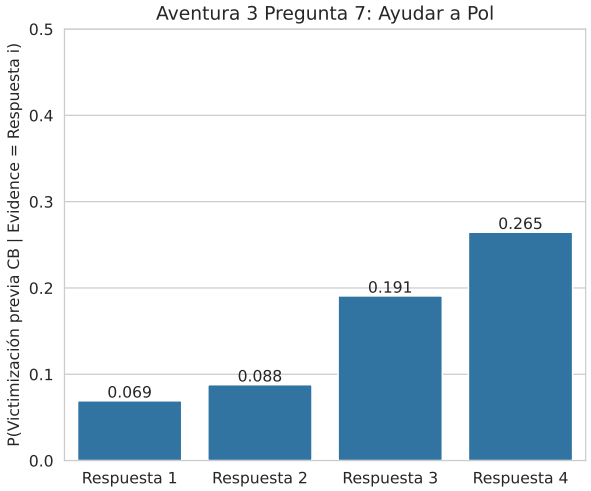
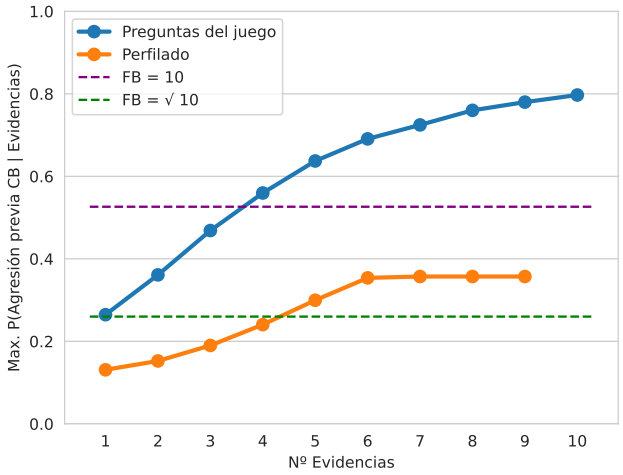


Figura 5: Probabilidades condicionales de *Agresión previa CB*, dadas las posibles respuestas a la pregunta *Aventura 3 - Pregunta 7 - Ayuda a Pol*. Según los resultados del análisis de fuerza de la flecha, esta variable es la más influyente en el resultado del MGCP. Esto se ilustra marginando los posibles valores de la pregunta y actualizando las probabilidades en el MGCP, ya que la probabilidad de observar *Agresión previa CB* cambia significativamente.



Respuesta 1: 'Deberíamos decírselo al profesor, él sabrá qué hacer'
Respuesta 2: 'Deberíamos denunciar los comentarios en la red social, para que no vuelva a ocurrir'
Respuesta 3: 'No debemos denunciarlo, porque no quiero que se metan conmigo por ser un chivato...'
Respuesta 4: 'No debemos denunciarlo ya que suele ser inútil'

Figura 6: Análisis multifactorial de perfiles: Se muestra la máxima probabilidad posterior de observar *Agresión previa CB* para cada número de evidencias fijadas fijadas. Esto se hace, por una parte, para las variables obtenidas a través de las preguntas del juego y, por otra, para las variables de perfilado. La figura también muestra las probabilidades correspondientes a los umbrales pertinentes según el criterio de Jeffreys (Ecuación 1).

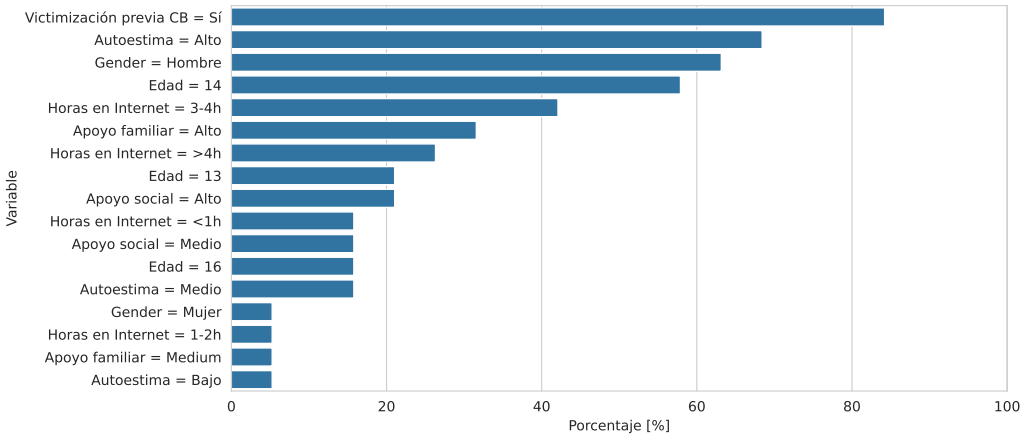


IV. Discusión y limitaciones

Nuestra investigación aporta emocionantes evidencias de que los datos derivados de los JS pueden aportar una visión más profunda de los comportamientos alrededor del CB que los enfoques tradicionales de elaboración de perfiles, como los datos demográficos y los cuestionarios psicológicos. Estos resultados ejemplifican la eficacia de los JS como una herramienta novedosa para la investigación en ciencias sociales y que puede utilizarse como una forma amena y no invasiva de recopilar información de poblaciones de difícil acceso, como los menores de edad.

A pesar de estos prometedores hallazgos, debemos ser cautos a la hora de interpretar nuestros resultados y reconocer las posibles limitaciones del trabajo. En cuanto a las limitaciones

Figura 7: Análisis multifactorial de perfiles: Se muestra la prevalencia de las características compartidas por los perfiles de riesgo para cinco evidencias fijadas. En particular, la característica *Victimización previa CB* aparece en más del 80 % de los perfiles de riesgo considerados.



experimentales, los datos se recogieron únicamente en tres países europeos, por lo que serían necesarios más experimentos en poblaciones diferentes para confirmar la fiabilidad de los resultados. También hay que tener en cuenta que los datos de los cuestionarios y los videojuegos suelen ser ruidosos y heterogéneos, especialmente cuando se trata de menores (por ejemplo, algunos participantes pueden haber jugado al azar o haber respondido deliberadamente a las preguntas de forma incorrecta).

Los MGCP permiten mitigar los sesgos de los datos (por ejemplo, seleccionando variables colisionadores y de confusión) [36], [37]. Este enfoque también nos obliga a hacer explícitas nuestras presuposiciones e hipótesis, lo que da lugar a debates y planteamientos críticos sobre las cuestiones que intentamos abordar y contribuye a una ciencia más accesible y comprensible. Sin embargo, al mismo tiempo, los modelos causales dependen en gran medida de supuestos como la ausencia de factores de confusión no medidos, la estabilidad de las relaciones causales a lo largo del tiempo y de las poblaciones, y la correcta especificación del modelo. Las estimaciones causales pueden estar sesgadas o ser engañosas si estos supuestos se incumplen o son incorrectos.

Los resultados muestran que las variables obtenidas a través del JS son relevantes para explicar y predecir el riesgo de cometer CB. Las variables demográficas y las obtenidas mediante cuestionarios psicológicos no tienen una relevancia significativa cuando se analizan individualmente. Sin embargo, la variable de *Victimización previa CB* aparece como muy prevalente a través del análisis multifactorial de perfiles. Indicando que las personas que previamente habían sido víctimas de CB muestran una propensión significativamente mayor a cometerlo.

Nuestros resultados indican que es menos fructífero intentar establecer perfiles basados en datos demográficos y cuestionarios psicológicos, lo que sugiere que la propensión al riesgo de las personas se define mejor por sus acciones (es decir, las respuestas a situaciones específicas) que por sus rasgos personales. Sin embargo, en la red definida por los expertos de RAYUELA (Fig. 3) hay un gran número de interdependencias entre las variables personales y demográficas, por lo que cabría esperar que, a medida que aumente la cantidad de datos de la muestra, estas variables adquieran más importancia de la que hemos encontrado en nuestros experimentos.

V. Conclusiones

Este artículo propone una metodología para analizar los datos de una JS sobre CB mediante un análisis basado en causalidad. Para ello, proponemos utilizar MGCP, que proporciona un marco sofisticado para modelar gráficamente las relaciones causales. Además, permite introducir intuitivamente el conocimiento experto en la estructura de la red y las probabilidades a priori. Esta investigación pretende demostrar la validez y el potencial del uso de los JS como herramientas de investigación prácticas, atractivas y no invasivas en las ciencias sociales. Aplicamos las técnicas propuestas sobre datos recogidos a través de pilotos experimentales dentro del proyecto H2020 RAYUELA. Por lo tanto, esta investigación también pretende aportar nuevos conocimientos sobre los mecanismos causales subyacentes del CB.

Los resultados del análisis causal confirman la importancia de las preguntas del juego para identificar y comprender el fenómeno del CB, lo que podría utilizarse para diseñar estrategias eficaces de prevención y entrenamiento del CB entre los menores. Cabe destacar que la variable demográfica con mayor relevancia es *Victimización previa CB*, lo que significa que los menores que habían sido víctimas de CB tenían más probabilidades de convertirse ellos mismos en agresores. Este ciclo de victimización y perpetración subraya la importancia de abordar las experiencias de victimización pasadas para comprender y prevenir futuros comportamientos de CB.

Los JS se postulan como una alternativa eficaz para medir cómo actuarían los jugadores en determinadas situaciones de la vida real, siempre que existan limitaciones éticas y el juego esté bien diseñado para la función deseada. El uso de un JS contribuye a una mejor inmersión y desinhibición de los jugadores, lo que resulta crucial para estudiar temas delicados o segmentos de población difíciles de alcanzar en investigación, como los menores de edad. Además, el análisis causal resulta especialmente útil para comprender en profundidad cuestiones

delicadas, contribuyendo a desarrollar medidas preventivas más eficaces para reducir la ciberdelincuencia entre los menores y garantizar su bienestar en línea.

AGRADECIMIENTOS

Este trabajo ha recibido financiación del programa de investigación e innovación Horizonte 2020 de la Unión Europea en virtud del acuerdo de subvención nº 882828. El contenido de este documento es responsabilidad exclusiva de los autores y no refleja en modo alguno la opinión de la Unión Europea. Este trabajo ha contado con el apoyo parcial de la subvención PID2022-140217NB-I00 financiada por MCIN/AEI/ 10.13039/501100011033 y, por .ERDF A way of making Europe". Agradecemos a todos los participantes del proyecto RAYUELA su implicación y desarrollo del juego, la recogida de datos y la realización de estudios piloto.

REFERENCIAS

[1] UNICEF, "The State of the World's Children 2017: Children in a Digital World," UNICEF Division of Communication, Tech. Rep., 2017, ISBN: 978-92-806-4930-7. [Online]. Available: https://www.unicef.org/media/48581/file

[2] 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," EU Kids Online, Tech. Rep., 2020, ISSN: 2045-256X. [Online]. Available: https://www.eukidsonline.ch/files/Eu-kids-online-2020-international-report.pdf

[3] 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.

[4] 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*, vol. 157, p. 103958, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360131520301561

[5] C. C. Abt, *Serious games*. University press of America, 1987.

[6] J. Pérez, M. Castro, and G. López, "Serious games and ai: Challenges and opportunities for computational social science," *IEEE Access*, vol. 11, pp. 62 051–62 061, 2023.

[7] 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*, vol. 604, no. 7904, pp. 104–110, Mar. 2022. [Online]. Available: https://doi.org/10.1038/s41586-022-04486-7

[8] J. K. Hartshorne, J. B. Tenenbaum, and S. Pinker, "A critical period for second language acquisition: Evidence from 2/3 million english speakers," *Cognition*, vol. 177, pp. 263–277, Aug. 2018. [Online]. Available: https://doi.org/10.1016/j.cognition.2018.04.007

[9] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The Moral Machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, Nov. 2018. [Online]. Available: http://www.nature.com/articles/s41586-018-0637-6

[10] B. Long, J. Simson, A. Buxó-Lugo, D. G. Watson, and S. A. Mehr, "How games can make behavioural science better," *Nature*, vol. 613, no. 7944, pp. 433–436, Jan. 2023. [Online]. Available: https://doi.org/10.1038/d41586-023-00065-6

[11] A. L. S. Ronald L. Wasserstein and N. A. Lazar, "Moving to a world beyond "p¡0.05"," *The American Statistician*, vol. 73, no. sup1, pp. 1–19, 2019. [Online]. Available: https://doi.org/10.1080/00031305.2019.1583913

[12] C. Rahal, M. Verhagen, and D. Kirk, "The rise of machine learning in the academic social sciences," *AI & SOCIETY*, pp. 1–3, 2022.

[13] J. Pearl, *Causality*. Cambridge university press, 2009.

[14] L. E. Sucar, *Probabilistic Graphical Models: Principles and Applications*. Springer International Publishing, 2021. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-61943-5

[15] E. M. Airoldi, "Getting started in probabilistic graphical models," *PLoS Comput Biol*, vol. 3, no. 12, p. e252, 12 2007. [Online]. Available: https://doi.org/10.1371/journal.pcbi.0030252

[16] J. M. Rohrer, "Thinking clearly about correlations and causation: Graphical causal models for observational data," *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 1, pp. 27–42, 2018. [Online]. Available: https://doi.org/10.1177/2515245917745629

[17] F. Elwert, *Graphical Causal Models*. Dordrecht: Springer Netherlands, 2013, pp. 245–273. [Online]. Available: https://doi.org/10.1007/978-94-007-6094-3_13

[18] P. Hünermund and E. Bareinboim, "Causal inference and data fusion in econometrics," *arXiv preprint arXiv:1912.09104*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.1912.09104

[19] S. Greenland, J. Pearl, and J. M. Robins, "Causal diagrams for epidemiologic research," *Epidemiology*, vol. 10, no. 1, pp. 37–48, 1999. [Online]. Available: http://www.jstor.org/stable/3702180

[20] 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*, vol. 12, 2021. [Online]. Available: https://doi.org/10.3389/fpsyt.2021.640927

[21] L. Cheng, R. Guo, and H. Liu, "Robust cyberbullying detection with causal interpretation," in *Companion Proceedings of The 2019 World Wide Web Conference*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 169–175. [Online]. Available: https://doi.org/10.1145/3308560.3316503

[22] J. PEARL, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 12 1995. [Online]. Available: https://doi.org/10.1093/biomet/82.4.669

[23] G. López, N. Bueno, M. Castro, M. Reneses, J. Pérez, M. Riberas, M. Álvarez-Campana, M. Vega-Barbas, S. Solera-Cotanilla, L. Bastida *et al.*, "The H2020 project RAYUELA: A fun way to fight cybercrime," *Jornadas Nacionales de Investigación en Ciberseguridad - JNIC*, 2021.

[24] 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," H2020 RAYUELA, Tech. Rep., 2022. [Online]. Available: https://www.rayuela-h2020.eu/wp-content/uploads/2022/10/Attachment_0.pdf

[25] M. Reneses, M. Riberas, N. Bueno, B. Heylen, E. Andreotti, and J. Op den Kelder, "Open report on case study results," H2020 RAYUELA, Tech. Rep., 2022. [Online]. Available: https://www.rayuela-h2020.eu/wp-content/uploads/2022/10/Attachment_0-1.pdf

[26] M. Reneses, M. Riberas, A. Gómez, N. Bueno, B. Heylen, and J. Ginter, "Open report on victim and offender profile description report," H2020 RAYUELA, Tech. Rep., 2022. [Online]. Available: https://www.rayuela-h2020.eu/wp-content/uploads/2022/10/Attachment_0-2.pdf

[27] M. Rosenberg, "Rosenberg self-esteem scale (rse)," *Acceptance and commitment therapy. Measures package*, vol. 61, no. 52, p. 18, 1965.

[28] G. D. Zimet, N. W. Dahlem, S. G. Zimet, and G. K. Farley, "The multidimensional scale of perceived social support," *Journal of Personality Assessment*, vol. 52, no. 1, pp. 30–41, 1988. [Online]. Available: https://doi.org/10.1207/s15327752jpa5201_2

[29] A. Brighi, R. Ortega, J. Pyzalski, H. Scheithauer, P. K. Smith, H. Tsormpatzoudis, H. Tsorbatzoudis, and et al., "European cyberbullying intervention project questionnaire," 2012. [Online]. Available: https://doi.org/10.1037/t66195-000

[30] J. R. Koiter, "Visualizing inference in bayesian networks," M.Sc. thesis, Faculty of Electrical Engineering, Mathematics, and Computer Science, Department of Man-Machine Interaction, Delft University of Technology, 2006.

[31] D. Endres and J. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1858–1860, 2003.

[32] G. N. Yannakakis and J. Togelius, Artificial intelligence and games. Springer, 2018, vol. 2.

[33] H. Jeffreys, *The theory of probability*. OuP Oxford, 1998.

[34] 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*, vol. 123, p. 105946, 2021. [Online]. Available: https://doi.org/10.1016/j.childyouth.2021.105946

[35] G. Livazović and E. Ham, "Cyberbullying and emotional distress in adolescents: the importance of family, peers and school," *Heliyon*, vol. 5, no. 6, p. e01992, 6 2019. [Online]. Available: https://doi.org/10.1016/j.heliyon.2019.e01992

[36] I. Shrier and R. W. Platt, "Reducing bias through directed acyclic graphs," *BMC Medical Research Methodology*, vol. 8, no. 1, p. 70, 10 2008. [Online]. Available: https://doi.org/10.1186/1471-2288-8-70

[37] 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*, vol. 19, no. 3, pp. 303–311, 2014. [Online]. Available: https://doi.org/10.1111/resp.12238