Ex Ante Evaluation of Safety by Design in Social Gaming Platforms: Risk Vectors and Age-Assurance Trade-offs in Brazil

Filipe Pessôa Sousa[1], Augusto Magalhães Pinto de Mendonça[2]
Igor Machado Coelho[2], Cristiane Oliveira de Faria[1]
Rosa Maria Esteves Moreira da Costa[1], Vera Maria Benjamim Werneck[1]

[1] Instituto de Matemática e Estatística –
Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro – RJ – Brasil

[2]Instituto de Computação –
Universidade Federal Fluminense (UFF)
Niterói – RJ – Brasil

filipe.sousa@pos.ime.uerj.br, {cofaria,rcosta,vera}@ime.uerj.br
augustompm@id.uff.br, imcoelho@ic.uff.br

Abstract. ***Introduction:*** *Children and adolescents in social gaming ecosystems face multiple online risks, while interventions such as age assurance may also impose costs in privacy, access friction, and digital inclusion.* ***Objective:*** *We propose an ex ante evaluability framework to assess Safety by Design measures for online child protection in Brazil.* ***Methodology:*** *Using TIC Kids Online Brasil 2024 and complementary parameters from TIC Domicílios, we construct five group-level risk vectors, represent mechanisms through two qualitative matrices (mechanism-to-risk and mechanism-to-cost), and run an exploratory net-benefit simulation under alternative cost scenarios.* ***Results:*** *Risk exposure is heterogeneous across groups, with higher concentration in contact, informational exposure, and harassment. The evaluated package is more coherent for reducing contact-related risks than for directly addressing content or scam-related harms, while the most protective mechanisms also tend to be the most costly in privacy and access friction. The framework makes these trade-offs explicit and supports evaluation before post-implementation causal evidence is available.*
Keywords *Safety by Design, Age Assurance, Online Child Safety, Digital Risk, Platform Governance.*

1. Introduction

In Brazil, expanded internet access among children and adolescents has increased participation in digital platforms and social gaming ecosystems [CGI.br e NIC.br 2025]. While these environments support socialization and learning, they also expose minors to risks such as unwanted contact with strangers, scams, harassment, and exposure to inappropriate or extremist content [UNICEF EAPRO and YRRC 2022]. In response, the *Safety by Design* agenda advocates embedding safeguards directly into platform architecture [Molnar e Robinson 2024].

Yet interventions centered on *age assurance* and age verification are not neutral. They may raise privacy concerns and deepen digital exclusion, and they can increase *access friction*, defined here as added steps, time, and effort to access or use a platform, often increasing drop-off rates [Livingstone et al. 2024, Jarvie e Renaud 2024, CGI.br e NIC.br 2021]. Moreover, the literature highlights implementation loopholes and questions whether mandatory verification delivers proportional safety gains without introducing new data-protection vulnerabilities [Eltaher et al. 2025, Stardust et al. 2024].

Against this background, this paper proposes an *ex ante* framework to evaluate *Safety by Design* measures in social gaming platforms in Brazil. We separate what the intervention aims to reduce (risk vectors) from what it may impose on users (costs in privacy, access friction, and digital inclusion). Using TIC Kids Online Brasil 2024 microdata aggregated by sex, age group, social class, and race/color [CGI.br e NIC.br 2025], and complementary parameters from TIC Domicílios [CGI.br e NIC.br 2018], we build composite indicators, encode protection mechanisms in two qualitative matrices (mechanism-to-risk and mechanism-to-cost), and present an exploratory net-benefit simulation ($Net(g)$). The goal is not to claim realized causal effects, but to provide an evaluability-oriented account of the intervention logic and the trade-offs involved in online child protection.

2. Literature Review

2.1. Safety by Design and Age Assurance

The growing use of digital platforms by children and adolescents has intensified the debate on *Safety by Design*, i.e., embedding safeguards directly into the architecture of digital services [Molnar e Robinson 2024]. Within this agenda, *age assurance* is often presented as a prerequisite for tailoring interactions and content by age, yet it raises persistent concerns about effectiveness, circumvention, privacy, and surveillance, especially under document-based or biometric regimes [Livingstone et al. 2024, Stardust et al. 2024]. Evidence also suggests that these measures can increase user burden and perceived friction, affecting uptake and compliance [Woodley et al. 2025].

2.2. Risk Vectors and Intervention Costs

Online harms faced by children and adolescents are commonly framed as multidimensional, spanning content, contact, conduct, and contract [Livingstone et al. 2024, CGI.br e NIC.br 2025]. In social gaming ecosystems, this includes unwanted contact with strangers, harassment, scams, and exposure to inappropriate content [CGI.br e NIC.br 2025]. At the same time, interventions can impose non-trivial costs in privacy, access friction, and inclusion, particularly in unequal settings where connectivity constraints amplify barriers [CGI.br e NIC.br 2021, Stardust et al. 2024]. Recent regulatory debates formalize these trade-offs across assurance, data protection, and user convenience, reinforcing the need to treat costs as first-class outcomes in evaluation [Liu e Scheffler 2026].

2.3. Ex Ante Evaluation, Composite Indicators, and Sensitivity

When post-implementation evidence is limited, *ex ante* evaluation supports scrutinizing causal coherence and anticipated impacts before deployment [European Commission 2014, Livingstone e Pothong 2025]. Composite indicators offer a practical way to aggregate multiple risk dimensions into interpretable measures, and sensitivity analysis helps assess how weighting schemes and cost scenarios affect conclusions [OECD e JRC 2008, Libório et al. 2024, Morgan 2014, Tarantola et al. 2006, Kelemen et al. 2024].

3. Methodology

This study adopts an *ex ante* evaluability approach to examine the coherence of *Safety by Design* mechanisms aimed at protecting children and adolescents in social gaming platforms. Rather than estimating realized causal impact, the goal is to organize, analytically, the risk vectors the intervention intends to reduce and the costs it may impose on users [European Commission 2014, UNICEF EAPRO and YRRC 2022].

Our main empirical basis is TIC Kids Online Brasil 2024 [CGI.br e NIC.br 2025]. We use published tables (including margins of error) and individual-level microdata, summarized by survey group cuts (sex, age group, social class, and race/color). As a complement, we draw on TIC Domicílios to inform parameters related to digital inclusion [CGI.br e NIC.br 2021]. The pipeline follows successive steps of extraction, harmonization, and indicator aggregation.

Table 1. Risk and cost components, operational indicators, and data treatment

Component	Operational indicators	Source and treatment
$R_{contact}$	Contact with strangers, interaction exposure, and offline meetings with people first contacted online.	TIC Kids 2024; normalized to $[0, 1]$; grouped with survey weights.
$R_{informational\_exposure}$	Disclosure or sharing of personal information and other identifying data in online interactions.	TIC Kids 2024; normalized to $[0, 1]$; grouped with survey weights.
R_{scams}	Identity or credential misuse, fraud-related events, and financial-loss situations reported in online use.	TIC Kids 2024; normalized to $[0, 1]$; grouped with survey weights.
$R_{harassment}$	Offensive treatment, unwanted requests, and harmful or abusive interactions in online environments.	TIC Kids 2024; normalized to $[0, 1]$; grouped with survey weights.
$R_{content}$	Exposure to sensitive content, including violence, drugs, self-harm or suicide, and extreme body-image content.	TIC Kids 2024; normalized to $[0, 1]$; grouped with survey weights.
$C_{inclusion}$	Connectivity constraints, mobile-only access, and access conditions that may amplify exclusion under high-friction mechanisms.	TIC Domicilios; used as inclusion-cost parameter.

As summarized in Table 1, the model is structured around five risk components: $R_{contact}$, $R_{informational\_exposure}$, R_{scams}, $R_{harassment}$, and $R_{content}$. Percentage indicators were rescaled to the $[0, 1]$ interval, and component scores were computed by aggregating normalized indicators within each risk dimension. In addition to central scores, we consider uncertainty bounds derived from the margins of error reported in the original tables.

As a methodological complement, TIC Kids microdata are used to capture empirical co-occurrence of vulnerabilities. Each behavior is binarized, aggregated at the respondent level, and then summarized by group using survey weights, mitigating limitations of marginal-only analyses [OECD e JRC 2008].

The intervention is decomposed into five mechanisms: a communication gate via *age check*, age-group segmentation, *Trusted Connections*, parental consent, and periodic age re-checks. From these mechanisms, we construct two qualitative matrices. The first maps each mechanism to the risk vectors, classifying the expected effect as direct reduction, neutral, or uncertain. The second maps the same mechanisms to the cost dimensions considered in the study, $C_{privacy}$, $C_{friction}$, and $C_{inclusion}$, classified in qualitative levels.

Matrix entries were coded according to explicit mechanism-based rules. In Matrix A, direct reduction was assigned when a mechanism directly restricts the channel through which a risk occurs; neutral effect was assigned when no direct mechanism–risk pathway was identifiable; and uncertain effect was assigned when the effect was plausible but indirect or implementation-dependent. In Matrix B, cost levels were assigned according to the expected burden imposed on privacy, access friction, and digital inclusion. The coding rationale for each matrix entry is provided in the replication package.

Here, $C_{friction}$ refers to *access friction*, namely additional steps, time, effort, and higher drop-off risk introduced by technical barriers in access and use. This separation prevents conflating expected risk reduction with the burden imposed by the intervention itself [Livingstone et al. 2024, Stardust et al. 2024, Molnar e Robinson 2024]. We adopt vectors and matrices as a direct operationalization of *ex ante* evaluability: vectors summarize latent harm dimensions as composite indicators, while matrices make the intervention logic auditable before post-implementation evidence is available [European Commission 2014, OECD e JRC 2008].

Based on the risk vectors, we compute an expected benefit score for each demographic group as a weighted combination of the five risk components, and contrast it with an aggregated intervention cost composed of privacy, access friction, and digital inclusion. The exploratory net-benefit simulation is defined as:

$$Benefit(g) = \sum_{k=1}^{5} w_k R_k(g)$$

$$Net(g) = Benefit(g) - (\alpha C_{privacy} + \beta C_{friction} + \gamma C_{inclusion})$$

Table 2. Cost scenarios used in the net-benefit simulation

Scenario	$C_{privacy}$	$C_{friction}$	$C_{inclusion}$
Conservative	0.80	0.70	0.60
Moderate	0.50	0.40	0.35
Optimistic	0.20	0.15	0.10

Table 2 reports the scenario-level cost values used in the exploratory simulation. Risk-vector weights are reported in the order $(R_{contact}, R_{informational\_exposure}, R_{scams}, R_{harassment}, R_{content})$. The security-prioritized regime uses $w = (0.35, 0.10, 0.20, 0.25, 0.10)$ and $(\alpha, \beta, \gamma) = (0.20, 0.40, 0.40)$. The exposure/inclusion regime uses $w = (0.15, 0.10, 0.10, 0.15, 0.10)$, with input weights normalized internally before aggregation, and $(\alpha, \beta, \gamma) = (0.50, 0.25, 0.25)$. The balanced regime uses $w = (0.25, 0.20, 0.15, 0.20, 0.20)$ and $(\alpha, \beta, \gamma) = (0.33, 0.33, 0.34)$. Matrix B remains qualitative; $Net(g)$ uses the scenario-level cost values reported in Table 2. As a robustness check, we sample 100 random weight vectors from a Dirichlet distribution under the moderate scenario, using random seed 42, and assess output stability under alternative specifications [Tarantola et al. 2006, OECD e JRC 2008, Kelemen et al. 2024, Gelman et al. 2013]. Accordingly, $Net(g)$ is treated as a secondary and exploratory output; the central results remain the group-level risk vectors and the two qualitative matrices of mechanisms and costs.

The replication package includes preprocessing scripts, variable dictionaries, normalization routines, risk-vector construction scripts, qualitative matrix coding rules, cost-scenario parameters, Dirichlet random seed, processed outputs, figures, and step-by-step reproduction instructions. It is publicly available at `https://github.com/FilipePessoa30/Avaliacao-ex-ante-seguranca`.

4. Results

4.1. Risk vectors in the Brazilian context

Results indicate that risk vectors are not homogeneously distributed across the analyzed groups. At the aggregate level, the highest component is $R_{contact}$, followed by $R_{harassment}$, whereas R_{scams} presents the lowest mean value. This pattern suggests that, in the Brazilian child online context captured by the data sources used, the most salient risks are associated with expanded interactions and relational exposure rather than direct financial events.

Across demographic cuts, the clearest differences emerge by age group. Adolescents aged 13–14 and 15–17 show the highest values in $R_{contact}$ and $R_{informational\_exposure}$, indicating greater openness to interaction and more frequent participation in digital environments with intensified communication. For $R_{harassment}$, the 15–17 group also stands out, suggesting that increased autonomy online is accompanied by higher vulnerability to offensive or unwanted interactions.

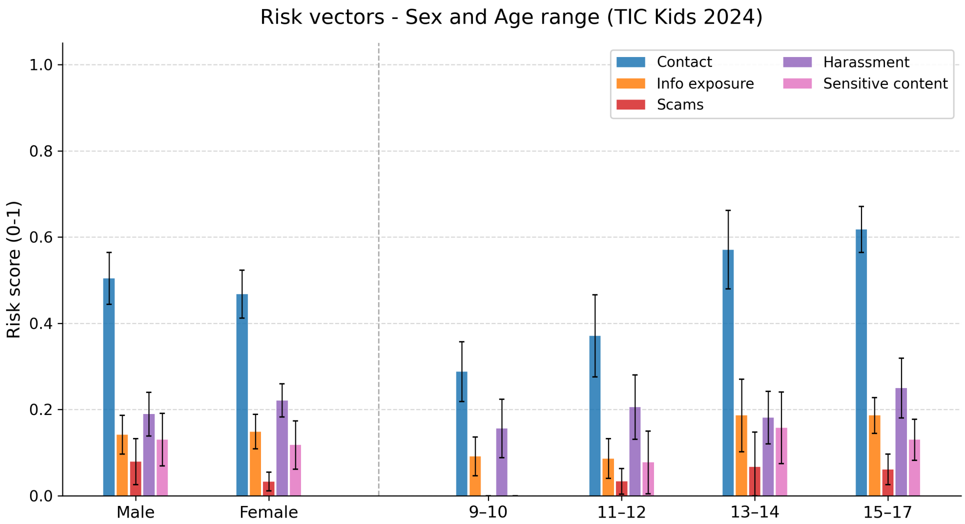

Figure 1. Risk vectors by sex and age group, with uncertainty intervals.

Figure 1 reinforces this finding by showing that age progression does not increase all vectors uniformly, but consistently raises components related to contact, informational exposure, and harassment. By sex, differences are more specific: boys present higher values in $R_{contact}$ and R_{scams}, whereas girls exhibit higher values in $R_{harassment}$. This pattern suggests that online exposure is multidimensional and cannot be summarized by a single synthetic risk indicator.

By social class and race/color, the results also suggest heterogeneity. Class AB shows higher observed values in components related to informational exposure, scams, and content, whereas class DE tends to show lower values in these components, potentially under distinct access and infrastructure conditions. Across race/color groups, some profiles present higher observed values in specific components, especially harassment, content, and scams. These differences should be interpreted as exploratory, since smaller subgroup sizes and wider uncertainty intervals may affect estimate stability. Thus, the results indicate possible heterogeneity in digital exposure, but should not be read as definitive evidence of stable differences between racial/color groups.

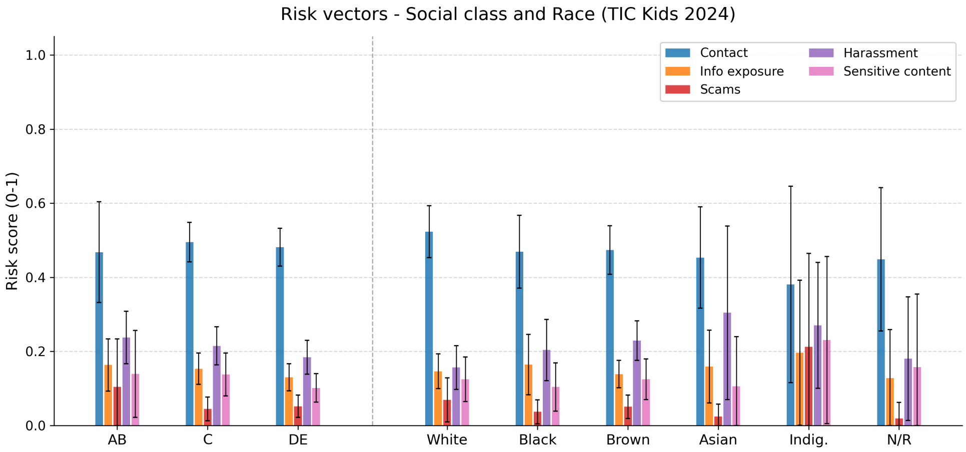

Figure 2. Risk vectors by social class and race/color, with uncertainty intervals.

Table 3 synthesizes the overall pattern of vectors and highlights, for each dimension, the group with the highest observed score in the analyzed set.

Table 3. Summary of risk vectors: aggregate total, highest observed score, uncertainty interval, and subgroup size

Component	Total	Highest score	Interval / n
$R_{contact}$	0.486	0.618 (15–17)	[0.565, 0.671] / 861
$R_{informational\_exposure}$	0.145	0.196 (Indigenous)	[0.001, 0.392] / 95
R_{scams}	0.055	0.213 (Indigenous)	[0.000, 0.465] / 95
$R_{harassment}$	0.205	0.304 (Asian)	[0.070, 0.538] / 87
$R_{content}$	0.124	0.230 (Indigenous)	[0.005, 0.456] / 95

The highest subgroup scores should be read as exploratory maxima under uncertainty, particularly for Indigenous and Asian respondents, whose smaller subgroup sizes and wider intervals may affect estimate stability. As a complementary check, composite indicators built from microdata show that empirical co-occurrence of vulnerabilities is non-negligible. In the full sample, 72.3% exhibit at least two behaviors associated with $R_{contact}$, and 29.4% exhibit at least two behaviors associated with $R_{content}$. This result supports the value of modeling co-occurrence rather than relying only on marginal means.

4.2. Coherence of the proposed mechanisms

The second stage of results examines coherence between the mechanisms of the analyzed package and the estimated risk vectors. Matrix A does not measure realized impact; instead, it organizes the *ex ante* expected effects of each mechanism on each risk dimension.

Figure 3 highlights an important result: mechanisms do not reduce all vectors simultaneously. In particular, the communication gate via *age check* (M1), age-group segmentation (M2), parental consent (M4), and periodic re-checks (M5) present expected reduction mainly in $R_{contact}$. By contrast, effects on $R_{harassment}$, $R_{content}$, and part of R_{scams} are predominantly neutral or uncertain. This supports the interpretation that the package is more coherent for restricting contact and relational openness than for addressing sensitive content or harms among peers who are already connected.

This finding is relevant because it avoids a maximalist reading of the intervention.

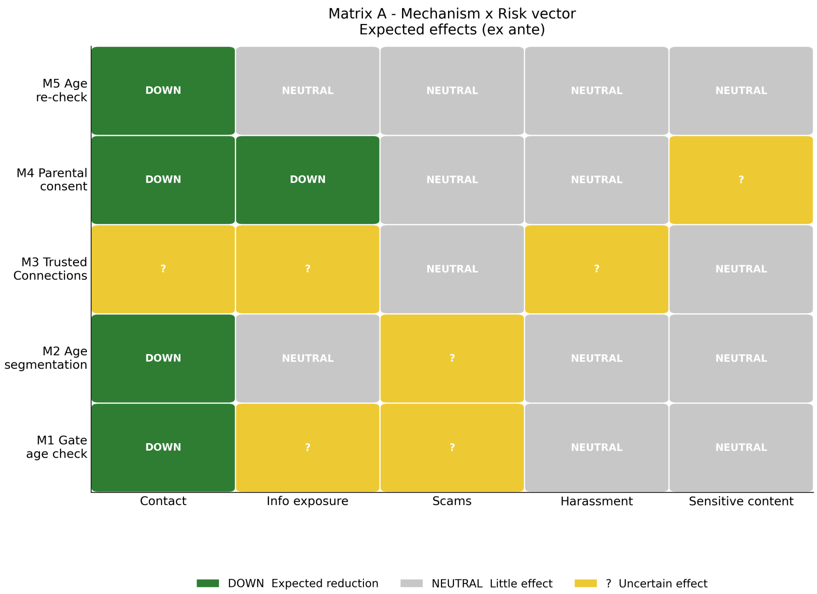

Figure 3. Matrix A: expected effects of mechanisms on the risk vectors.

The model suggests the policy may be coherent for certain vectors but insufficient for others, reinforcing the need for layered protection and complementary instruments such as content governance, moderation, and digital literacy.

4.3. Expected intervention costs

Matrix B organizes expected costs by mechanism across privacy, access friction, and digital inclusion. Unlike Matrix A, the focus here is not risk reduction, but the potential burden imposed on users.

Figure 4 shows that the most costly mechanism is the initial gate via *age check* (M1), with high cost in privacy and friction and moderate cost in inclusion. Periodic re-checks (M5) also concentrate high costs in privacy and friction. By contrast, *Trusted Connections* (M3) appears relatively less burdensome, with low cost in friction and inclusion while maintaining moderate privacy cost. Parental consent (M4) stands out for its high inclusion cost, suggesting a higher risk of exclusion for families with lower digital mediation capacity.

Taken together, Figures 3 and 4 show that the package's protective coherence concentrates on specific vectors, while costs are unevenly distributed across mechanisms. In other words, the strongest instruments for restricting contact also tend to be the most costly in terms of privacy and access friction.

4.4. Exploratory trade-off simulation

As a secondary step, we perform an exploratory simulation of $Net(g)$ under different cost scenarios. Results confirm that net-benefit readings are highly sensitive to normative assumptions and cost calibration. Under the moderate scenario, some groups appear relatively more sensitive to the intervention, particularly older adolescents and groups with higher concentration of risk vectors; however, values remain negative in this regime, making any interpretation of realized net benefit inappropriate.

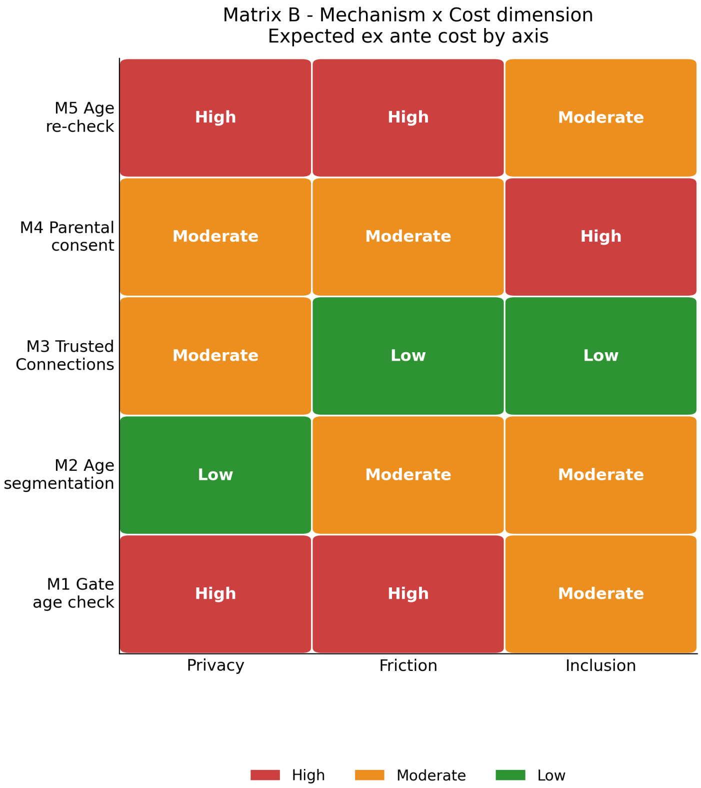

Figure 4. Matrix B: expected costs by mechanism.

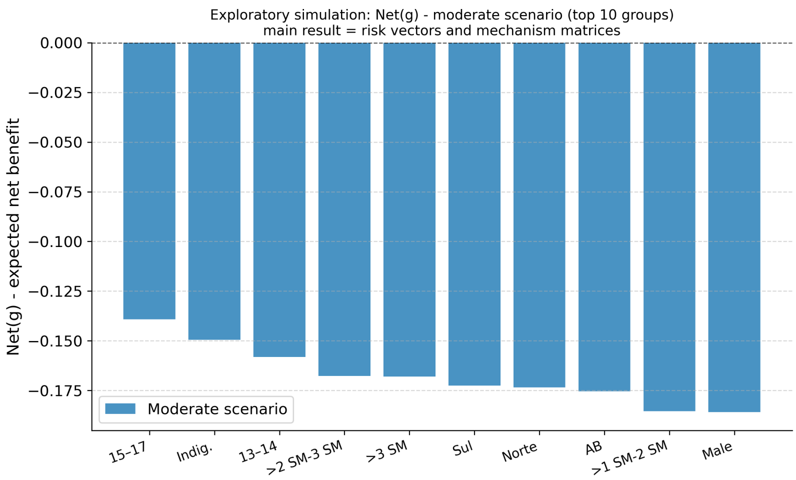

Figure 5. Exploratory $Net(g)$ simulation under the moderate scenario for groups with higher gross benefit.

Figure 5 should therefore be read as a sensitivity exercise on trade-offs. More exposed groups tend to occupy the top of the relative ranking even when net benefit does not become positive under the moderate scenario, indicating that the result depends on the relation between risk intensity and assumed cost magnitude. Additional sensitivity outputs are provided in the replication package. They show that some group rankings remain relatively stable under weight perturbation, while others vary more substantially, supporting the exploratory interpretation of $Net(g)$ rather than a substantive estimate of realized intervention benefit.

5. Conclusion

This paper proposed an *ex ante* framework to evaluate *Safety by Design* measures in social gaming platforms in Brazil, explicitly separating the risk vectors an intervention aims to reduce from the costs it may impose in terms of privacy, access friction, and digital inclusion. Using TIC Kids Online Brasil data and complementary parameters from TIC Domicílios, we constructed group-level risk vectors, organized protection mechanisms into two qualitative matrices, and presented an exploratory trade-off simulation.

Results indicate that children's and adolescents' digital exposure is not homogeneous across groups and that the most salient components concentrate primarily on contact, informational exposure, and harassment. Moreover, the mechanism analysis suggests that the evaluated package is more coherent for restricting contact-related risks than for directly addressing sensitive content, scams, or harmful dynamics among already connected peers. At the same time, the mechanisms that are potentially strongest from a protective standpoint also appear to be the most burdensome, especially in privacy and access friction.

These findings reinforce that online child safety policies should not be assessed only by their protective intent, but also by the costs and inequalities they may generate. In this sense, the main contribution of the paper is to provide an analytical account that makes such trade-offs explicit, avoiding both techno-solutionist readings and purely abstract rejections of platform intervention.

As a limitation, this study does not estimate realized causal effects. Instead, it focuses on coherence and evaluability based on aggregated indicators, microdata-derived measures, and qualitative parameters. Future work can advance through post-implementation validation using longitudinal evidence, independent platform audits, and evaluation designs that test whether the studied mechanisms reduce the intended risks without imposing disproportionate costs on users.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001; by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), under Grants No. 312165/2025-0 and No. 102199/2024-8; and by the Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), through the Jovem Cientista do Nosso Estado Program (JCNE 2025), under Process No. SEI-260003/020882/2025.

References

CGI.br e NIC.br (2018). *Pesquisa sobre o uso das tecnologias de informação e comunicação nos domicílios brasileiros: TIC Domicílios 2017*. Comitê Gestor da Internet no Brasil, São Paulo.

CGI.br e NIC.br (2021). *Pesquisa sobre o uso das tecnologias de informação e comunicação nos domicílios brasileiros: TIC Domicílios 2020 (Edição COVID-19)*. Comitê Gestor da Internet no Brasil, São Paulo.

CGI.br e NIC.br (2025). *Pesquisa sobre o uso da Internet por crianças e adolescentes no Brasil: TIC Kids Online Brasil 2024*. Comitê Gestor da Internet no Brasil, São Paulo.

Eltaher, F., Gajula, R. K., Miralles-Pechuán, L., Thorpe, C., e McKeever, S. (2025). The digital loophole: Evaluating the effectiveness of child age verification methods on social media. Technical report.

European Commission (2014). Guidelines for the ex-ante evaluation of 2014-2020 emff ops. Technical report.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., e Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, 3 edition.

Jarvie, C. e Renaud, K. (2024). Online age verification: Government legislation, supplier responsibilization, and public perceptions. *Children*, 11(9):1068.

Kelemen, A., Szabó, Z. K., Bozóki, S., Szádoczki, Z., e Hartvig, Á. D. (2024). A sensitivity analysis of composite indicators: Min/max thresholds. *Environmental and Sustainability Indicators*, 23:100453.

Libório, M. P., Ekel, P., D'Angelo, M. F. S. V., Martínez, L., e Rabiei, H. (2024). Constructing composite indicators through extreme values reductions-ordered weighted averaging: Human development index. *IEEE Access*.

Liu, S. e Scheffler, S. (2026). Adequately tailoring age verification regulations. arXiv:2601.20241v2.

Livingstone, S., Nair, A., Stoilova, M., van der Hof, S., e Caglar, C. (2024). Children's rights and online age assurance systems: The way forward. *The International Journal of Children's Rights*, 32(3):721–747.

Livingstone, S. e Pothong, K. (2025). Child rights impact assessment: A policy tool for a rights-respecting digital environment. *Policy & Internet*, 17:e70008.

Molnar, A. e Robinson, L. (2024). Towards digital safety by design for children. Technical report, OECD Digital Economy Papers.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, 111(20):7176–7184.

OECD e JRC (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing, Paris.

Stardust, Z., Obeid, A., McKee, A., e Angus, D. (2024). Mandatory age verification for pornography access: Why it can't and won't save the children. *Big Data & Society*.

Tarantola, S., Nardo, M., Saisana, M., e Gatelli, D. (2006). A new estimator for sensitivity analysis of model output: An application to the e-business readiness composite indicator. *Reliability Engineering and System Safety*, 91:1135–1141.

UNICEF EAPRO and YRRC (2022). Evaluating online safety initiatives: How to build the evidence base on what works to keep children safe online. Technical report, UNICEF East Asia and Pacific Regional Office and Young and Resilient Research Centre, Bangkok.

Woodley, G., See, H. W., O'Neill, B., Green, L., Staksrud, E., e Haskell-Dowland, P. (2025). Australian teen voices on age verification and age assurance measures. *Policy & Internet*, 17:e70019.