Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

Protection mechanisms for minors on digital platforms and proposal of the GuardaMirim prototype

Mecanismos de proteção de menores em plataformas digitais e proposta do protótipo GuardaMirim

Mecanismos de protección de menores en plataformas digitales y propuesta del prototipo GuardaMirim

Elis do Nascimento Barbosa – Faculdade Metropolitana de Manaus, elis.nbarbosa22@gmail.com
Henrique da Silva Lima – Faculdade Metropolitana de Manaus, silvalimahenrique14@gmail.com
Laina Santos Almeida – Faculdade Metropolitana de Manaus, lainasantos352@gmail.com
Luiza Natasha Tavares Azevedo – Faculdade Metropolitana de Manaus, luizanatashaxd@gmail.com
Vitor Hugo Guimarães Itou – Faculdade Metropolitana de Manaus, vitoritouguimaraes@gmail.com

Abstract:
Internet access among children and adolescents is growing in Brazil, increasing their exposure to risks such as cyberbullying, manipulation, coercion and contact with strangers. This article analyzes technological mechanisms for protecting this audience in digital environments, considering the Brazilian General Data Protection Law (LGPD) and the Digital Statute of Children and Adolescents (ECA Digital, Law No. 15,211/2025). It is an applied study with a qualitative, exploratory and descriptive approach, based on bibliographic and documentary research and on a comparative analysis of platforms (Discord, Instagram, TikTok and WhatsApp) according to five criteria: age verification, content moderation, parental controls, privacy of minors' content, and incident response and transparency. The analysis indicated that parental controls and reporting tools are described in official sources for all four platforms, whereas age verification and transparency appear unevenly in the analyzed sources. Based on this framework, the interface prototype of the GuardaMirim app is proposed to alert guardians about signs of risky behavior without exposing the content of conversations, in line with the principles of security, privacy and the best interests of children and adolescents. The prototype uses simulated data, does not implement a classifier, and has not been validated with users. The study concludes that parental supervision may be strengthened by solutions that complement the platforms' native tools, which depends on future validation, without replacing the responsibility of platforms, guardians and authorities.

Keywords:
Cyberbullying. Privacy. Supervision. LGPD. Adolescents.

Resumo:
O acesso de crianças e adolescentes à internet cresce no Brasil e amplia a exposição a riscos como cyberbullying, manipulação, coerção e contato com desconhecidos. Este artigo analisa mecanismos tecnológicos de proteção desse público em ambientes digitais, considerando a Lei Geral de Proteção de Dados (LGPD) e o Estatuto Digital da Criança e do Adolescente (ECA Digital, Lei nº 15.211/2025). Trata-se de uma pesquisa aplicada, de abordagem qualitativa e caráter exploratório e descritivo, baseada em pesquisa bibliográfica e documental e em análise comparativa de plataformas (Discord, Instagram, TikTok e WhatsApp), segundo cinco critérios: verificação de idade, moderação de conteúdo, controle parental, privacidade do conteúdo do menor e resposta a incidentes e transparência. A análise indicou que o controle

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

parental e os canais de denúncia estão descritos em fonte oficial nas quatro plataformas, ao passo que a verificação de idade e a transparência aparecem de forma desigual nas fontes analisadas. A partir desse referencial, propõe-se o protótipo de interface do aplicativo GuardaMirim, concebido para alertar os responsáveis sobre sinais de comportamentos de risco sem expor o conteúdo das conversas, em consonância com os princípios de segurança, privacidade e melhor interesse da criança e do adolescente. O protótipo usa dados simulados, não implementa classificador e não foi validado com usuários. Conclui-se que a supervisão parental pode vir a ser fortalecida por soluções complementares às ferramentas nativas das plataformas, o que depende de validação futura, sem substituir a responsabilidade das plataformas, dos responsáveis e das autoridades.

Palavras-chave:
Cyberbullying. Privacidade. Supervisão. LGPD. Adolescentes.

Resumen:
El acceso a Internet entre niños y adolescentes está creciendo en Brasil, aumentando su exposición a riesgos como el ciberacoso, la manipulación, la coerción y el contacto con extraños. Este artículo analiza los mecanismos tecnológicos para proteger a esta audiencia en entornos digitales, considerando la Ley General de Protección de Datos (LGPD) y el Estatuto Digital de la Niñez y la Adolescencia. (ECA Digital, Ley n.º 15.211/2025). Es un estudio aplicado con un enfoque cualitativo, exploratorio y descriptivo, basado en la investigación bibliográfica y documental y en un análisis comparativo de plataformas (Discord, Instagram, TikTok y WhatsApp) según cinco criterios: verificación de edad, moderación de contenido, controles parentales, privacidad de contenido de menores, y respuesta a incidentes y transparencia. El análisis indicó que los controles parentales y las canales de denuncia se describen en fuentes oficiales para las cuatro plataformas, mientras que la verificación de la edad y la transparencia aparecen de manera desigual en las fuentes analizadas. A partir de este marco, se propone el prototipo de interfaz de la aplicación GuardaMirim para alertar a los tutores sobre signos de comportamiento de riesgo sin exponer el contenido de las conversaciones, en línea con los principios de seguridad, privacidad y el interés superior de los niños y adolescentes. El prototipo utiliza datos simulados, no implementa un clasificador y no ha sido validado con los usuarios. El estudio concluye que la supervisión parental puede fortalecerse con soluciones que complementen las herramientas nativas de las plataformas, que dependen de la validación futura, sin reemplazar la responsabilidad de las plataformas, los tutores y las autoridades.

Palabras clave:
Cyberbullying. Privacidad. Supervisión. LGPD. Adolescentes

1 Introduction

In the contemporary scenario, technological advancement has made internet access by children and adolescents increasingly common, modifying the way this public communicates and interacts with the world. According to the research TIC Kids Online Brazil 2025, carried out by the Information and Coordination Center of the BR Point (NIC.br), 92% of Brazilian children and adolescents between 9 and 17 years old were users of the internet, evidencing the

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

presence of digital technologies in their daily lives (NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR, 2026).

Despite the benefits provided by connectivity, the growing presence of this audience in the digital environment also expands its exposure to risks such as cyberbullying, contact with unknowns, manipulation, coercion and inadequate content. Given this scenario, the protection of children and adolescents requires family monitoring and technological mechanisms capable of assisting in the identification and prevention of harmful behaviors.

In this context, child and youth protection should consider the legal provisions related to security and privacy. The General Data Protection Law (LGPD), Law no 13.709/2018, establishes that the treatment of data from children and adolescents should observe their best interest. In addition, the Digital Statute of Children and Adolescents (ECA Digital), established by Law No. 15.211/2025, establishes protection measures and responsibilities for agents involved in the provision of technological services.

Although platforms such as Discord, Instagram, TikTok and WhatsApp provide mechanisms of parental security and supervision, these tools present limitations, especially in the identification of risks in private conversations. Considering these needs, the research proposes the Guarda Mirim application, a complementary protection tool with foreseen mechanisms of detection and alert, seeking to preserve the privacy of minors without exposing the original content of the conversations. In this stage, the proposal is presented through a navigable prototype with simulated data.

In view of this scenario, the following research problem is established: how technological mechanisms for detection and prevention can contribute to the protection of children and adolescents against cyberbullying and other risk behaviors in the digital environment, considering LGPD, ECA Digital and the participation of those responsible for supervision?

The general objective of this research is to analyze technological mechanisms of protection of children and adolescents in digital environments, with emphasis on the prevention of cyberbullying and other risk behaviors, considering Brazilian legislation and proposing the Guarda Mirim application as a complementary tool for parental supervision.

To achieve this goal, the following specific objectives were established: to identify the main risks and forms of digital violence against children and adolescents; to analyze the protection mechanisms of the Discord, Instagram, TikTok and WhatsApp platforms, considering the requirements of LGPD and the ECA Digital; and to present the requirements,

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

the proposed architecture and the interface prototype of Guarda Mirim, respecting the principles of security and privacy.

It is hypothesized that technological mechanisms of detection and prevention, developed according to the principles of safety, privacy and better interest of children and adolescents, can contribute to the early identification of risk situations and strengthen the performance of those responsible, without replacing family monitoring.

As for the methodological procedures, the research is characterized as applied, qualitative, exploratory and descriptive, based on bibliographic and documentary research and on the comparative analysis of the protection mechanisms of four digital platforms. The development of the prototype was guided by the design science research approach, contemplating the definition of requirements, modeling and implementation of the interfaces.

As a result, limitations were identified in the protection mechanisms of the platforms and a navigable prototype of Guarda Mirim, composed of seven screens, was developed. However, automated detection has not yet been implemented, and future evaluations are needed to verify its effectiveness.

Finally, the article is organized into five sections, in addition to this introduction: theoretical framework, methodology, development of the Guarda Mirim prototype, results and discussion, and final considerations.

2 Theoretical framework

2.1 Digital risks and violence against children and adolescents

In 2025, 92% of people aged 9 to 17 years in Brazil were internet users, about 24 million children and adolescents (NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR, 2026). For this audience, the internet is a space of study, leisure and socialization, but also of exposure to risks that depend not only on the content accessed, but on who comes into contact with the minor and how relationships develop online.

Cyberbullying is a form of digital violence. In accordance with Law No. 13,185/2015, "there is systematic intimidation in the world computer network (cyberbullying), when using the instruments that are appropriate to depreciate, incite violence, adulterate photos and personal data in order to create means of psychosocial constraints". (BRASIL, 2015, art. 2º, parágrafo único). Law No. 14,811/2024 began to treat systematic intimidation, including virtual intimidation, as a crime (BRASIL, 2024). In comparison with face-to-face bullying,

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

cyberbullying can reach expanded audiences, maintain accessible records and occur under anonymity, relevant characteristics for its detection and coping (SALAWU; HE; LUMSDEN, 2020).

For the purposes of this research, the risks were organized into four categories, which also guide the GuardaMirim prototype:

- Cyberbullying: repeated aggressions, humiliation and exclusion of a minor through messages, publications or groups.
- Offensive language: curses, hate speech and threats, which can occur in isolation, without the typical repetition of cyberbullying, but with potential for damage.
- Inadequate image: receiving or sharing images unfit for age, such as nudity and violence, including intimate images of the minor himself.
- Unknown contact: approximation of adults or unknowns, which can evolve to grooming.

The manipulation and coercion are not necessarily isolated categories, but transversal strategies that may be present in different manifestations of violence. Studies on the online mediation of children and adolescents and on coercive control point out that these practices can progressively develop, through the construction of trust, emotional control and social isolation of the victim, establishing conditions that favor the domain exercised by the aggressor (WHITTLE et al., 2013; WIENER, 2017; MCLINDON et al., 2025). This pattern appears, for example, in police investigation reported in 2026, in which the suspects, after gaining the confidence of children and adolescents, subject us to coercion, isolation and psychological violence (FANTTI, 2026). The categories are also not exclusive, because the same situation can combine, for example, unknown contact, inadequate image and coercion.

2.2 Legal framework for the protection of children and adolescents in the digital environment

In Brazil, the General Law on Personal Data Protection (LGPD) dedicates art. 14 to the data treatment of children and adolescents, which should be performed in their best interest (BRASIL, 2018). In the case of children, specific consent is required and emphasis is given to at least one parent or legal guardian, and the information about the treatment should be presented in simple and accessible language (BRASIL, 2018, art. 14, §§ 1º e 6º). The principles of art. 6th, in particular, need, transparency, security and prevention. For technological solutions, this means collecting only the strictly necessary data and explaining clearly what is treated.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

The Digital Statute of Children and Adolescents (ECA Digital), established by Law No. 15.211/2025, entered into force on March 17, 2026 (AGÊNCIA BRASIL, 2026). It applies to information technology products and services directed to children and adolescents or likely access by them, which covers social networks, messaging applications and game platforms (BRASIL, 2025; FUNDAÇÃO ABRINQ PELOS DIREITOS DA CRIANÇA E DO ADOLESCENTE, 2026). Among its main requirements are:

- Age verification by reliable mechanisms, being insufficient the mere self-declaration;
- Parental supervision, with accessible tools for parents and guardians to follow and configure the use, including the attachment of the accounts of adolescents up to 16 years old to that of a guardian;
- Transparency, with periodic reports of the platforms (art. 31), whose first publication period was fixed by the National Agency for Data Protection on September 17, 2026 (AGÊNCIA NACIONAL DE PROTEÇÃO DE DADOS, 2026).

The law also establishes that compliance with these obligations cannot result in mechanisms of massive, generic or indiscriminate surveillance (BRASIL, 2025, art. 34, § 1º). This prediction is central to this study, because it delimits what a protection solution can do without violating the privacy of the minor.

In the international scenario, the United States Children's Online Privacy Protection Rule (COPPA) requires verifiable parental consent for data collection of children under 13 years old (FEDERAL TRADE COMMISSION, 2025), and art. 8th of the General Data Protection Regulation of the European Union (GDPR) fixed the age of consent in 16 years, and the Member States can reduce it up to 13 (UNIÃO EUROPEIA, 2016). These milestones focus on consent and treatment of data, while the ECA Digital also deals with the design of services, age verification and parental supervision.

2.3 Mechanisms of protection in digital environments

From the legal framework, five protection mechanisms are identified, which correspond to the criteria of Table 1.

Age verification: seeks to establish the age group of the user to offer an adequate experience and restrict inappropriate content. It differs from the verification of identity, as it identifies whether the user is smaller, and not who he is. Self-declaration is considered insufficient by the Digital ECA (BRASIL, 2025).

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

Content modulation: it includes automated detection, human review and reporting channels used to identify and remove harmful behaviors. The literature on automated cyberbullying detection points to contextual interpretation and classification challenges (SALAWU; HE; LUMSDEN, 2020); these systems can err and, in services with end-to-end cryptography, the analysis of the content by the platform is technically restricted; therefore, isolated moderation is not enough to protect minors in private conversations.

Parental control: it consists of tools that allow the responsible to link accounts, establish time limits of use and contacts, in addition to monitoring certain aspects of the use of services. The ECA Digital establishes the need for availability and accessibility of these tools (BRASIL, 2025). However, studies on parental control identify tensions between protection, autonomy and privacy of adolescents (WISNIEWSKI et al., 2017; GHOSH et al., 2018a).

Privacy of the minor's content: refers to the degree of access that the platform, the responsible or third parties have to the content of the interactions of the minors. This mechanism is related to the principles of best interest and necessity (BRASIL, 2018), as well as the prohibition of massive surveillance provided by the Digital ECA (BRASIL, 2025, art. 34, § 1º). Therefore, supervision should seek the lowest possible level of data exposure, without compromising the necessary protection.

Response to incidents and transparency: it covers the publication of reports, the attendance of complaints and the submission of platforms to inspection and the sanctions provided for in law. These procedures allow society and the government to evaluate, based on public documents, whether the mechanisms of protection are effective in practice (BRASIL, 2025, art. 31).

2.4 Synthesis and gap

The scientific literature and the legislation examined point to two trends (WISNIEWSKI et al., 2017; GHOSH et al., 2018a; BRASIL, 2025). The first is the accountability of the platforms to verify the age, moderate content and offer parental supervision. The second is the concern with the privacy of the minor, which limits unrestricted access to the content of conversations. The native mechanisms, however, operate within each service, which forces the responsible to follow several panels, and not always manage to alert him in time about situations such as cyberbullying, grooming and coercion, which manifest themselves in private conversations.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

This gap indicates the need for complementary solutions that identify signs of risk and inform the responsible without delivering the content of the conversations, in line with the principles of minimization and of best interest. It is in this space that the proposal of GuardaMirim is inserted, detailed in section 4, and whose adequacy to the mark presented here is discussed below.

3 Material and method

3.1 Characterization of the research

This research is of an applied nature, with a qualitative approach and an exploratory and descriptive character (GIL, 2017). It was conducted on two complementary fronts. The first is a comparative documentary analysis of the mechanisms of security and moderation adopted by digital platforms to protect children and adolescents. The second is the construction of a technological artifact, the GuardaMirim prototype, guided by the design science research, in which the development of an artifact is part of the research method (HEVNER et al., 2004; PEFFERS et al., 2007).

The option for document analysis is justified by the recent character of the object: Law No. 15.211/2025 (ECA Digital) entered into force on March 17, 2026 (AGÊNCIA BRASIL, 2026), and the platforms still adjust their resources and documents. As the materials differ in scope and level of detail, the comparison does not presuppose direct equivalence between them. The research did not include interviews or direct data collection with the companies, restricted to the analysis of public documents and the development of the prototype.

3.2 Data sources and selection criteria

Three categories of sources were consulted: (i) legislation and official documents, including Law No. 15.211/2025 (BRASIL, 2025), Law no 13.709/2018 (BRASIL, 2018), the North American COPPA rule (FEDERAL TRADE COMMISSION, 2025), the General Data Protection Regulation of the European Union (UNIÃO EUROPEIA, 2016) and materials of the National Data Protection Agency (ANPD); (ii) institutional documents of each platform, selected from security, help, privacy, parental supervision and transparency of the Discord, Instagram/Meta, TikTok and WhatsApp (DISCORD, [2026?]; META, 2026a, 2026b, 2026c; TIKTOK, [2026?], 2026; WHATSAPP, [2026?]); and (iii) institutional materials, research

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

reports and reports on inspection, litigation, child and youth protection and changes in platform resources.

The primary sources were consulted in the official portals of Planalto, ANPD, Federal Trade Commission and EUR-Lex, as well as in the security and help centers of the platforms. To locate studies and news, Google Search, Google Scholar, SciELO and CAPES Journal Portal were used, with terms such as "ECA Digital", "age verification", "parental control" and "cyberbullying", combined by AND and OR operators; the analysis was based on the recovered and verified documents.

The final corpus consisted of 20 documents, individually accounted for: four normative frameworks (ECA Digital, LGPD, COPPA and GDPR), seven institutional materials of the platforms (one of the Discord, two of the Meta/Instagram, two of the WhatsApp and two of the TikTok), an official page of the ANPD on the ECA Digital, two institutional/research materials (Fundação Abrinq and NIC.br/Cetic.br) and six reports (Agencia Brasil, Migalhas, Piauí Hoje, Showmetech, Canaltech and Jornal de Brasília/Folhapress). The theoretical and methodological references used only to support the research were not included in this count.

The inclusion criteria were: (a) publication or update between 2025 and 2026 (for pages without explicit date, the version available in the collection was considered, with registration of the date of access), except for previous normative frameworks (LGPD, COPPA and GDPR); (b) direct relationship with at least one of the criteria of section 3.4; and (c) identification of authorship or the responsible body. Primary sources were prioritized; the reports were used in a complementary way, such as journalistic reports and not as audited facts. The main collection took place in September 2026, and the final conference of the links was made until November 1, 2026.

3.3 Selection of platforms

Four platforms were analyzed: Discord, Instagram, TikTok and WhatsApp. The cut was based on (i) its relevance in the Brazilian public and regulatory debate in 2026, including the public civil action of the Attorney General of the Union against the Discord (MIGALHAS, 2026), the investigation of the Federal Public Prosecutor's Office on the Security of the Discord (PIAUÍ HOJE, 2026) and the alteration of the indicative classification (SHOWMETECH, 2026); and (ii) its framework in the law, which covers social networks, messaging applications and game platforms (FUNDAÇÃO ABRINQ PELOS DIREITOS DA CRIANÇA E DO ADOLESCENTE, 2026). It is an intentional sampling, without pretension of

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

representativeness, and the differences between the types of service limit the direct comparability.

3.4 Criteria for analysis

The comparison was structured in five criteria, derived from the theoretical framework (Quadro 1).

Table 1 – Analysis criteria and guiding question

Criterion	Guiding question
Age verification	Does the platform use a mechanism beyond self-declared age?
Content moderation	Are automated moderation, human review and a reporting channel available? What limitations appear in the sources?
Parental control	Is native parental supervision available? What can the guardian view and configure?
Privacy of minors' content	Can the guardian or the platform access the content of interactions?
Incident response and transparency	Are reports, sanctions or regulatory actions publicly available?

Source: elaborated by the authors (2026).

3.5 Procedures

The procedure followed four steps: (1) bibliographic and normative survey and definition of the criteria; (2) collection and organization of information in a platform × criterion matrix; (3) comparison of the identified mechanisms with the requirements of the ECA Digital and LGPD; and (4) interpretation by categorical content analysis (BARDIN, 2016), with the five criteria of Table 1 as categories. Each cell was classified as a mechanism identified in official source, identified only in journalistic source, or not identified in the sources consulted, which indicates absence of public information, and not absence of the resource.

3.6 Ethical and legal aspects

The study used only publicly accessible documents and did not involve human participants. Ethics committee approval was therefore not required under Article 1, sole paragraph, item II, of CNS Resolution No. 510/2016 (BRASIL, 2016). No personal data were collected from children or adolescents: the prototype's alerts, names and metrics are fictitious. The artifact's design considered the best interests of the child and data minimization (BRASIL,

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

2018, art. 14), as well as the prohibition of mass surveillance (BRASIL, 2025, art. 34, § 1º). Any future study involving participants will require ethics review before recruitment.

4 Development of the GuardaMirim prototype

4.1 Overview and scope

GuardaMirim is an application to support child and youth protection, designed to detect cyberbullying signals and other digital risks and alert the person responsible without exposing the content of the child's conversations. The development followed three iterative steps, typical of design science research: problem identification, design and development, and evaluation, which summarize the six activities of the original model (PEFFERS et al., 2007).

The scope of the artifact is limited, at this stage, to the interface design and to the navigation flow. The content classification engine was not implemented, and all the data displayed are simulated. The complete architecture, described in section 4.3, is a proposal for future steps.

4.2 Requirements

The requirements were derived from the analysis criteria (Quadro 1) and the legal restrictions discussed in the theoretical framework. Table 2 presents the main ones.

Table 2 - GuardaMirim requirements

Code	Type	Description
RF01	Functional	Allow selection between the minor’s profile and the guardian’s profile
RF02	Functional	Require authentication (a 4-digit PIN, used only as a minimum barrier in the prototype) to access the guardian’s profile
RF03	Functional	Display device protection status and occurrences by category over the last 7 days
RF04	Functional	List recent alerts with severity (low, medium or high), originating application and time
RF05	Functional	Provide alert details, including metadata and a classification summary
RF06	Functional	Offer the guardian the following actions: mark as resolved, consult conversation guidance and pause the originating application for 1 hour (a conceptual requirement, subject to operating system permissions)
RF07	Functional	Inform the minor, in a dedicated dashboard, that protection is active and what is analyzed
RF08	Functional	Allow the guardian to adjust detection sensitivity and enable or disable alert categories

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

Code	Type	Description
RF09	Functional	Notify the guardian of new alerts, including on the lock screen, without displaying raw content
RF10	Functional	Allow the guardian to record agreement or disagreement with the alert classification (feedback)
RNF01	Non-functional	Privacy by default: no screen may display the raw content of messages or images
RNF02	Non-functional	Communicate severity through both color and text, using a non-alarmist tone
RNF03	Non-functional	Accessibility: visible focus, large touch targets and respect for the reduced-motion preference

Source: elaborated by the authors (2026).

The risk categories of the prototype (cyberbullying, offensive language, inadequate image and unknown contact) reflect the risk taxonomy presented in section 2.1.

4.3 Proposed architecture

The architecture was designed in layers, so that the content of the interactions remains in the device of the minor (Figura 1 e Quadro 3). Only the classification of the event (category, severity, application of origin, context and time) is sent to the responsible.

Table 3 – Layers of architecture and implementation situation

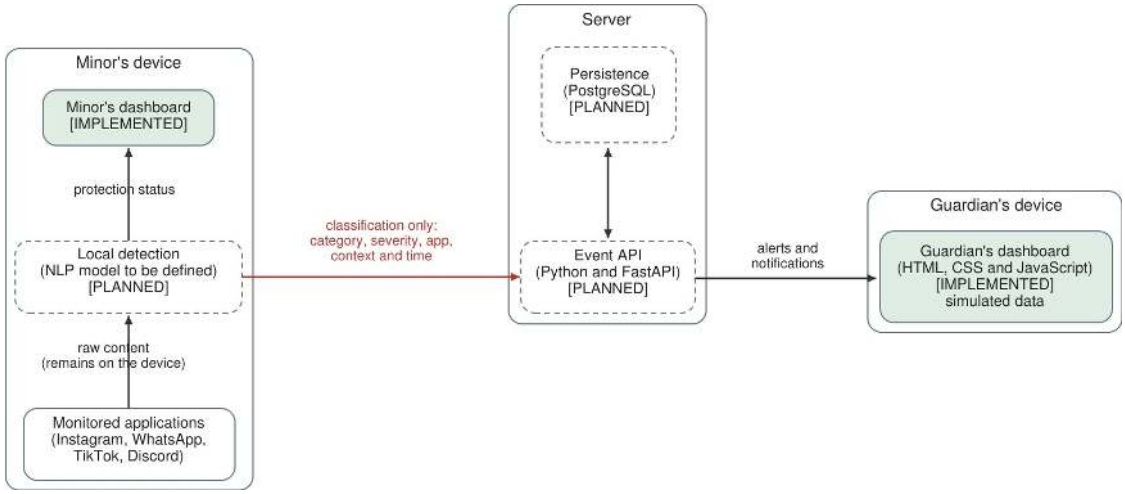
Layer	Function	Technology	Implementation status
Interface (minor and guardian profiles)	Screens and navigation	HTML, CSS and JavaScript	Implemented (simulated data)
Local detection	Classify content on the minor's device	Natural language processing (NLP) model to be defined	Planned
Event API	Receive and deliver event classifications	Python and FastAPI	Planned
Persistence	Store events and preferences	PostgreSQL	Planned

Source: elaborated by the authors (2026).

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

Figure 1 - Proposed architecture of GuardaMirim



Source: elaborated by the authors (2026). Traced blocks indicate predicted components, not yet implemented.

This organization seeks to meet the ECA Digital, which prohibits mechanisms of massive, generic or indiscriminate surveillance (BRASIL, 2025, art. 34, § 1º). What circulates between the layers is the classification of the event, never the text or the image that originated it. The policies of retention and disposal of event metadata still need to be defined, with deadlines, purpose, access control, exclusion and audit registration.

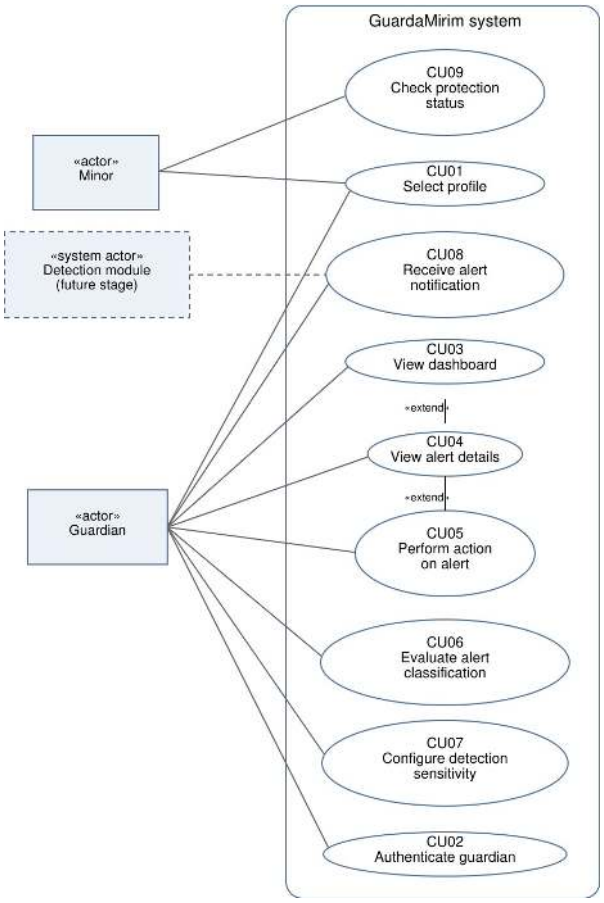
4.4 Modeling

The modeling identified two human actors, the minor and the responsible, and the detection module as a future systemic actor. Nine cases of use were raised (CU01 a CU09), presented in the diagram of Figure 2.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

Figure 2 - Diagram of cases of use of GuardaMirim



Source: elaborated by the authors (2026).

4.5 Implementation and interfaces

The prototype was implemented in HTML, CSS and JavaScript, as navigable flow of seven screens. Table 4 relates each screen to its function and to the corresponding privacy decision, and Figure 3 brings together the captures. The controller's flow runs through the panels (a) to (f): profile selection, authentication, panel, alert details, sensitivity configuration and notification on the blocking screen; the minor's flow corresponds to the panel (g). The prototype is available upon request from the authors.

Table 4 – Prototype screens and privacy decisions

Screen	Function	Privacy decision
Profile selection (Fig. 3a)	Separate access	Notice stating that analysis is local
Authentication (Fig. 3b)	Protect the guardian’s dashboard	Minimum barrier before access to sensitive data

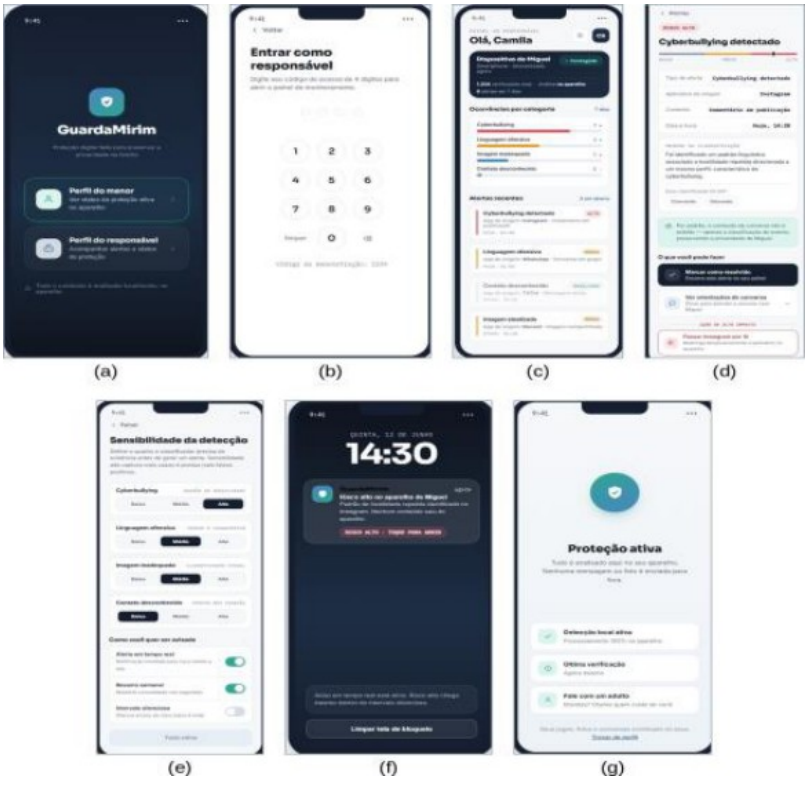
Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

Screen	Function	Privacy decision
Guardian's dashboard (Fig. 3c)	Status, category chart and alert feed	Displays metadata only
Alert details (Fig. 3d)	Metadata, summary and actions	Explicit notice that content is not displayed; actions support family mediation
Detection sensitivity (Fig. 3e)	Adjust detection strictness and monitored categories	The guardian controls the monitoring level without accessing content
Notification (Fig. 3f)	Notify the guardian of a new alert on the lock screen	Message with category and severity, without a conversation excerpt
Minor's dashboard (Fig. 3g)	Confirm active protection	Transparency about what is analyzed

Source: elaborated by the authors (2026).

Figure 3 - Screens of the GuardaMirim prototype (a a g)



Source: elaborated by the authors (2026).

4.6 Internal evaluation of interface consistency

The evaluation of this stage was internal: the team reviewed the coherence between the privacy discourse communicated to the user and what each screen actually displays. This review resulted, for example, the inclusion of privacy warning on the alert details screen. The prototype

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

has not yet been submitted to usability tests or validation with families, and there was no measurement of classification performance.

5 Results and discussion

5.1 Comparison of protection mechanisms on platforms

Table 5 synthesizes the platform × criterion constructed according to section 3.5. Each cell indicates whether the mechanism was identified in official source (O), only in journalistic source (J) or if it was not identified in consulted sources (–). As foreseen in the method, the third situation indicates absence of information in the analyzed documents, and not absence of the resource on the platform.

Table 5 - Platform matrix × criterion: mechanisms identified in the analyzed sources

Criterion	Discord	Instagram	TikTok	WhatsApp
Age verification	J - registration with date of birth; the Union claims that it is insufficient (MIGALHAS, 2026)	- Unidentified in the analyzed source (META, 2026b)	O – minimum age of 13 years; birth date form and account detection models of under 13 years (TIKTOK, 2026)	- Unidentified; the source does not describe method (META, 2026a)
Content moderation and reporting	O - reporting and blocking tools (DISCORD, [2026?]). J – the action calls for moderation in Portuguese and detection of risky livestreams (MIGALHAS, 2026)	O – channel of reporting of illegal content in Brazil, since March 2026 (META, 2026b)	O - technology and human review; reporting by anyone; cases of child sexual exploitation are reported (TIKTOK, 2026)	O – reporting and blocking of contacts and messages; context warning in messages of strangers; bans for violations of the Terms of Service, which prohibit content of abuse, intimidation or threat (WHATSAPP, [2026?])
Parental control	O - Family Center: account linking; the guardian sees friends, servers and recent contacts (DISCORD, [2026?])	O - activable supervision without authorization of the adolescent; alerts of searches for suicide and self-harm (META, 2026b); centralized management in the Family Center, with a unique invitation to Instagram, Facebook, Messenger and Meta Horizon (META, 2026c)	O – Family Pairing: time limit, content restrictions and message management (TIKTOK, 2026)	O – accounts managed by parents, for pre-adolescents (META, 2026a). J – new controls for adolescents, announced on September 30, 2026 (CANALTECH, 2026)

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

Criterion	Discord	Instagram	TikTok	WhatsApp
Privacy of minors' content	O - the guardian does not access the content of private messages (DISCORD, [2026?])	– Unidentified; the source cites only topics of conversations with Meta AI (META, 2026b)	O – accounts from 13 to 17 years old private by default, with publications visible only to approved followers; direct messages unavailable for 13 to 15 years (TIKTOK, [2026?]; TIKTOK, 2026). The sources do not clarify if parents see the messages	O – conversations with end-to-end encryption (META, 2026a). J – parents do not see the content of conversations (CANALTECH, 2026)
Incident response and transparency	J – Public Civil Action of AGU, with a request for compensation of at least R$ 500 million (MIGALHAS, 2026)	– Unidentified; the source does not mention transparency reports (META, 2026b)	O – quarterly reports at the Transparency Center; access of researchers to public data of minors (TIKTOK, 2026)	- Unidentified in the analyzed sources

Legend: O = official source; J = only journalistic source; - = not identified in the consulted sources. Source: elaborated by the authors (2026), based on the documents cited in each cell.

It was observed that parental control and reporting channels were the criteria with an official source of information on the four platforms. In contrast, the age check was described in official source only in TikTok. In the case of Discord, the only reference found comes from a journalistic source, which reports the claim of the Union that the declaration of the date of birth is insufficient, and for Instagram and WhatsApp, the analyzed sources do not detail the method. As for transparency reports, only TikTok describes them in official source; on Discord, the source found deals with regulatory action, and not with report.

As for privacy, Discord and WhatsApp inform that the responsible does not access the content of conversations, and, in WhatsApp, tip-to-end cryptography prevents even the company's access, according to the Goal (META, 2026a). This convergence is compatible with the prohibition of mass surveillance (BRASIL, 2025, art. 34, § 1º) and confirms the tension between protection and privacy of section 2.3: the native controls deliver to the responsible configurations and information of use, but not the content. The identified alerts refer to specific events, such as searches for suicide and self-mutilation on Instagram (META, 2026b) and entry into groups in WhatsApp (CANALTECH, 2026). In the consulted sources, public descriptions of alerts to the responsible about cyberbullying signs or grooming in private conversations were not identified, a gap that motivates GuardaMirim.

It is important to highlight the limits of the comparison. The matrix depends on the documents available until November 1, 2026, and the platforms change their resources frequently; the Meta note on Instagram, for example, was updated on October 29, 2026. The

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

information on Discord failures stems from Union claims reported by the press, and not from audit.

5.2 Limitations of the prototype and future validation steps

This subsection presents the limitations of the prototype and the future steps of GuardaMirim. The results described here were not obtained in this research: they are objectives of development and evaluation. The prototype demonstrates the interface and the navigation flow with simulated data; it does not implement a classifier, it has not been validated with families and does not prove access to content from other applications. The future steps are: (i) to evaluate the feasibility of a classifier in the device; (ii) to study the improvement of the model with protection of local data; (iii) to validate the experience with guardians and adolescents; and (iv) to specify data governance. The numerical criteria of approval should be defined before the confirmatory tests, based on a pilot study, and not by transposition of results from other data sets or devices.

The first step is to determine if it is feasible to replace the simulated data by a classifier executed on the user's device. Local inference is an architecture property, but it does not solve how the application would obtain the text to analyze. It will be necessary a proof of concept that identifies the source of the content, the required permissions and the technical and privacy restrictions. An initial scope alternative is to analyze content that the user himself shares voluntarily with GuardaMirim, without assuming unrestricted reading of other applications.

For Brazilian Portuguese, there are models and datasets that can support research: BERTimbau, a pre-trained model (SOUZA; NOGUEIRA; LOTUFO, 2020); ToLD-Br, of toxic language publications, whose reference work reports about 76% of macro-F1 in binary task (LEITE et al., 2020); and HateBR, comments from Instagram noted for offensiveness and hate speech (VARGAS et al., 2022). These numbers serve as a reference to compare published results, and not as a meta transferable to GuardaMirim.

The evaluation should first specify what will be classified, since offensive language, hate speech and cyberbullying are related but not equivalent phenomena. It will require operational categories, annotation guidelines, test data separated from the training and reporting of macro-F1, precision and recall by category, with examination of false positive and negative. The scarcity of representative labeled bases is one of the central challenges of the area (SALAWU; HE; LUMSDEN, 2020), and slang and irony of adolescents will require adequate ethical protocol. The execution on mobile devices is another test: MobileBERT reduced size

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

and latency in specific benchmarks (SUN et al., 2020), which does not predict the performance of a model in Portuguese in the target audience. Should be measured size, latency, memory and energy consumption, with limits fixed after the pilot study.

Another step is to evaluate whether the buttons "Concordo" and "Discordo" of the alert detail screen (RF10) can generate useful feedback to improve the classifier without sending the original message to a server. This feedback is not equivalent to a truth label: 'Discordo' indicates disagreement, but does not prove a false positive. The federated learning, in which the training occurs in the devices and only parameter updates are aggregated (MCMAHAN et al., 2017), is an alternative to investigate; it has already been applied to text forecasting on mobile keyboards (HARD et al., 2018), without proving its suitability to GuardaMirim. Model updates can also reveal indirect information, so that "not sending the text" does not guarantee absolute privacy.

The third stage is to evaluate the usability and understanding of the prototype with guardians and adolescents (seção 4.6). Studies on parental control applications point to tension between supervision, security, autonomy and privacy of adolescents (WISNIEWSKI et al., 2017; GHOSH et al., 2018a), and another study associated the use of monitoring applications with more online risk experiences, while direct involvement of those responsible was associated with fewer problems; because it is contextual evidence, it should not be taken as causal or generalized (GHOSH et al., 2018b). A pilot study with the two profiles is expected, combining observed tasks, semi-structured interviews and the System Usability Scale (SUS) (BROOKE, 1996). The average of about 70 points and the value of 71.4 associated with the adjective "good" (BANGOR; KORTUM; MILLER, 2009) serve as interpretive parameters, and not as a universal limit of approval.

The SUS will be complemented by indicators such as correct completion of tasks, errors, understanding of what is and what is not analyzed and perception of privacy. As it involves adolescents, the plan will be submitted to institutional ethical evaluation before recruitment, with the consent of those responsible, assent of adolescents and protection of records; the framework should be confirmed with the Research Ethics Committee, including the CNS Resolution No. 510/2016 (BRASIL, 2016).

5.3 Data governance, privacy and future work

The fourth stage is to specify a data policy consistent with the architecture and the rights of children and adolescents. Even without the transmission of the content of the messages, alert metadata such as category, source and time can be personal data when related to a minor, and

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

purpose, need and best interest should be evaluated for each data and treatment stage (BRASIL, 2018, art. 14; BRASIL, 2025, arts. 7º e 16).

Law No. 15.211/2025 determines that the controller maps and seeks to mitigate the risks of data processing of children and adolescents and elaborate an impact report, and Decree No. 12.880/2026 deals with the impact assessment on the safety and health of children (BRASIL, 2025, art. 16; BRASIL, 2026, art. 47). The project should examine the incidence of these requirements on its design and conclude the risk assessment before any pilot, as a measure of governance of his own work. The architecture should also be analyzed in the light of the prohibition of massive, generic or indiscriminate surveillance mechanisms (BRASIL, 2025, art. 34, § 1º, e art. 37, parágrafo único), evaluating which contents the system reaches, which metadata retains and who can consult alerts.

6 Final considerations

The present research aimed to analyze the technological mechanisms of protection of children and adolescents in the digital environment, considering the risks associated with cyberbullying and other harmful behaviors, as well as the provisions of the General Law on Personal Data Protection (LGPD) and the Digital Statute of Children and Adolescents (ECA Digital). From the analysis carried out, it was possible to understand the importance of reconciling security, privacy and parental supervision in the protection of this public.

The analysis of the security mechanisms of the Discord, Instagram, TikTok and WhatsApp platforms showed that, although there are resources for content moderation, age verification and parental control, these tools have limitations, especially in the identification of risk situations in private conversations. Thus, it was observed the need for complementary solutions that help those responsible for the prevention of harmful behaviors, respecting the privacy and autonomy of children and adolescents.

In this context, the prototype of the Guarda Mirim application was developed, a technological proposal aimed at supporting parental supervision. The work resulted in the definition of requirements, the elaboration of an architecture and the implementation of seven navigable screens, with functionalities aimed at presenting alerts and guidelines to those responsible. The proposal seeks to allow the identification of risk situations without exposing the original content of the conversations, prioritizing the principles of security and data protection.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

However, it is important to highlight that the prototype uses simulated data and does not yet have automated detection mechanisms implemented. In addition, no tests were performed with families or practical evaluations of their effectiveness. Thus, the hypothesis presented was sustained in the conceptual scope, but it still depends on validation. As a continuation of the research, it is recommended to investigate the feasibility of automated content classification, perform usability tests with guardians and adolescents and deepen the mechanisms of privacy and data protection.

Finally, it is emphasized that technology does not replace family dialogue and digital education, but can act as a tool to support the construction of a safer virtual environment.

References

AGÊNCIA BRASIL. ECA Digital começa a valer nesta terça; confira principais pontos. Brasília, 17 mar. 2026. Available at: https://agenciabrasil.ebc.com.br/politica/noticia/2026-03/eca-digital-comeca-valer-nesta-terca-confira-principais-pontos. Accessed on: 5 set. 2026.
AGÊNCIA NACIONAL DE PROTEÇÃO DE DADOS (ANPD). Plataformas digitais acessadas por crianças e adolescentes precisam publicar relatório de transparência até 17/09. Brasília, DF, 11 ago. 2026. Available at: https://www.gov.br/anpd/pt-br/assuntos/noticias/plataformas-digitais-acessadas-por-criancas-e-adolescentes-precisam-publicar-relatorio-de-transparencia-ate-17-de-setembro. Accessed on: 18 out. 2026.

BANGOR, Aaron; KORTUM, Philip; MILLER, James. Determining what individual SUS scores mean: adding an adjective rating scale. Journal of Usability Studies, v. 4, n. 3, p. 114–123, maio 2009. Available at: https://uxpajournal.org/wp-content/uploads/sites/7/pdf/JUS_Bangor_May2009.pdf. Accessed on: 18 out. 2026.

BARDIN, Laurence. Análise de conteúdo. São Paulo: Edições 70, 2016.

BRASIL. Conselho Nacional de Saúde. Resolução nº 510, de 7 de abril de 2016. Dispõe sobre as normas aplicáveis a pesquisas em Ciências Humanas e Sociais. Diário Oficial da União, Brasília, DF, 24 maio 2016.

BRASIL. Decreto nº 12.880, de 18 de março de 2026. Regulamenta a Lei nº 15.211, de 17 de setembro de 2025, e institui a Política Nacional de Promoção e Proteção dos Direitos da Criança e do Adolescente no Ambiente Digital. Diário Oficial da União, Brasília, DF, 18 mar. 2026, edição extra. Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2026/decreto/d12880.htm. Accessed on: 18 out. 2026.

BRASIL. Lei nº 13.185, de 6 de novembro de 2015. Institui o Programa de Combate à Intimidação Sistemática (Bullying). Diário Oficial da União, Brasília, DF, 9 nov. 2015. Available at: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13185.htm. Accessed on: 18 out. 2026.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, Brasília, DF, 15 ago. 2018. Available at: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Accessed on: 19 out. 2026.

BRASIL. Lei nº 14.811, de 12 de janeiro de 2024. Institui medidas de proteção à criança e ao adolescente contra a violência nos estabelecimentos educacionais ou similares, prevê a Política Nacional de Prevenção e Combate ao Abuso e Exploração Sexual da Criança e do Adolescente e altera o Código Penal e outras leis. Diário Oficial da União, Brasília, DF, 15 jan. 2024. Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/lei/l14811.htm. Accessed on: 19 out. 2026.

BRASIL. Lei nº 15.211, de 17 de setembro de 2025. Dispõe sobre a proteção de crianças e adolescentes em ambientes digitais (Estatuto Digital da Criança e do Adolescente). Diário Oficial da União, Brasília, DF, 17 set. 2025, edição extra. Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/l15211.htm. Accessed on: 19 out. 2026.

BROOKE, John. SUS: a “quick and dirty” usability scale. In: JORDAN, Patrick W. et al. (ed.). Usability evaluation in industry. London: Taylor & Francis, 1996. p. 189–194.

CANALTECH. WhatsApp ganha novos controles para pais de adolescentes: veja o que muda. [S. l.], 30 set. 2026. Available at: https://canaltech.com.br/apps/whatsapp-ganha-novos-controles-para-pais-de-adolescentes-veja-o-que-muda/. Accessed on: 1 nov. 2026.

DISCORD. Discord Family Center. [S. l.], [2026]. Available at: https://discord.com/safety-family-center. Accessed on: 1 nov. 2026.

FANTTI, Bruna. Morte de adolescente de 13 anos leva polícia a grupo suspeito de induzir jovens ao suicídio pela internet. Jornal de Brasília, Brasília, DF, 4 ago. 2026. Reportagem da Folhapress. Available at: https://jornaldebrasilia.com.br/noticias/brasil/morte-de-adolescente-de-13-anos-leva-policia-a-grupo-suspeito-de-induzir-jovens-ao-suicidio-pela-internet/. Accessed on: 1 nov. 2026.

FEDERAL TRADE COMMISSION. Children’s Online Privacy Protection Rule (COPPA). Washington, DC, 2025. Available at: https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa. Accessed on: 18 out. 2026.

FUNDAÇÃO ABRINQ PELOS DIREITOS DA CRIANÇA E DO ADOLESCENTE. ECA Digital: o que muda para crianças e adolescentes. São Paulo, 6 jul. 2026. Available at: https://www.fadc.org.br/noticias/eca-digital-o-que-muda. Accessed on: 5 set. 2026.

GHOSH, Arup Kumar et al. A matter of control or safety? Examining parental use of technical monitoring apps on teens’ mobile devices. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI), 2018, Montreal. Proceedings [...]. New York: ACM, 2018b. DOI: 10.1145/3173574.3173768.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

GHOSH, Arup Kumar et al. Safety vs. surveillance: what children have to say about mobile apps for parental control. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI), 2018, Montreal. Proceedings [...]. New York: ACM, 2018a. DOI: 10.1145/3173574.3173698.

GIL, Antonio Carlos. Como elaborar projetos de pesquisa. 6. ed. São Paulo: Atlas, 2017.

HARD, Andrew et al. Federated learning for mobile keyboard prediction. arXiv, arXiv:1811.03604, 2018. Available at: https://arxiv.org/abs/1811.03604. Accessed on: 18 out. 2026.

HEVNER, Alan R. et al. Design science in information systems research. MIS Quarterly, v. 28, n. 1, p. 75-105, 2004. DOI: 10.2307/25148625.

LEITE, João Augusto et al. Toxic language detection in social media for Brazilian Portuguese: new dataset and multilingual analysis. In: CONFERENCE OF THE ASIA-PACIFIC CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (AACL-IJCNLP), 1., 2020, Suzhou. Proceedings [...]. [S. l.]: ACL, 2020. p. 914–924. Available at: https://arxiv.org/abs/2010.04543. Accessed on: 18 out. 2026.

LINGUISTICS, 58., 2020. Proceedings [...]. [S. l.]: ACL, 2020. p. 2158–2170. DOI: 10.18653/v1/2020.acl-main.195.

MCLINDON, Elizabeth et al. Timelines of psychological, physical and sexual intimate partner violence among a nationally representative sample of Australian women. Women's Health, v. 21, p. 1-12, 2025. DOI: 10.1177/17455057251329640.

MCMAHAN, Brendan et al. Communication-efficient learning of deep networks from decentralized data. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS (AISTATS), 20., 2017, Fort Lauderdale. Proceedings [...]. [S. l.]: PMLR, 2017. v. 54, p. 1273–1282.

META. Apresentamos as contas gerenciadas por pais no WhatsApp. [S. l.], 11 mar. 2026a. Available at: https://about.fb.com/br/news/2026/03/apresentamos-as-contas-gerenciadas-por-pais-no-whatsapp/. Accessed on: 1 nov. 2026.

META. Iniciativas para manter pais brasileiros informados sobre as contas de adolescente e supervisão parental no Instagram. [S. l.], 15 jun. 2026b (atualizado em 29 out. 2026). Available at: https://about.fb.com/br/news/2026/06/iniciativas-para-manter-pais-brasileiros-informados-sobre-as-contas-de-adolescente-e-supervisao-parental-no-instagram/. Accessed on: 1 nov. 2026.

META. Supervisão parental simplificada em um só lugar. [S. l.], 12 maio 2026c. Available at: https://about.fb.com/br/news/2026/05/supervisao-parental-simplificada-em-um-so-lugar/. Accessed on: 18 out. 2026.

MIGALHAS. Governo processa Discord, pede R$ 500 mi e cobra proteção a menores. [S. l.], 26 ago. 2026. Available at: https://www.migalhas.com.br/quentes/463309/governo-processa-discord-pede-r-500-mi-e-cobra-protecao-a-menores. Accessed on: 5 set. 2026.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR (NIC.br) (ed.). Pesquisa sobre o uso da Internet por crianças e adolescentes no Brasil: TIC Kids Online Brasil 2025. 1. ed. São Paulo: Comitê Gestor da Internet no Brasil, 2026. ISBN 978-65-6165-007-6.

PEFFERS, Ken et al. A design science research methodology for information systems research. Journal of Management Information Systems, v. 24, n. 3, p. 45-77, 2007. DOI: 10.2753/MIS0742-1222240302.

PIAUÍ HOJE. MPF investiga segurança do Discord para proteger menores. [S. l.], 19 ago. 2026. Available at: https://piauihoje.com/index.php/noticias/brasil/mpf-investiga-seguranca-do-discord-para-proteger-menores-463658.html. Accessed on: 5 set. 2026.

SALAWU, Semiu; HE, Yulan; LUMSDEN, Joanna. Approaches to automated detection of cyberbullying: a survey. IEEE Transactions on Affective Computing, v. 11, n. 1, p. 3–24, 2020. DOI: 10.1109/TAFFC.2017.2761757.

SHOWMETECH. Governo federal proíbe o Discord para menores de 18 anos. [S. l.], 15 jun. 2026. Available at: https://www.showmetech.com.br/governo-federal-proibe-o-discord-para-menores-de-18-anos/. Accessed on: 5 set. 2026.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS), 9., 2020, Rio Grande. Proceedings [...]. Cham: Springer, 2020. (Lecture Notes in Computer Science, v. 12319), p. 403–417.

SUN, Zhiqing et al. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL

TIKTOK. Nosso compromisso inequívoco com a segurança dos jovens no TikTok. [S. l.], 17 mar. 2026. Available at: https://newsroom.tiktok.com/nosso-compromisso-inequvoco-com-a-segurana-dos-jovens-no-tiktok?lang=pt-BR. Accessed on: 1 nov. 2026.

TIKTOK. Privacidade e segurança para usuários menores de 18 anos. [S. l.], [2026?]. Available at: https://support.tiktok.com/pt_BR/account-and-privacy/account-privacy-settings/privacy-and-safety-settings-for-users-under-age-18. Accessed on: 18 out. 2026.

UNIÃO EUROPEIA. Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016. Relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados e que revoga a Diretiva 95/46/CE (Regulamento Geral sobre a Proteção de Dados). Jornal Oficial da União Europeia, Luxemburgo, L 119, p. 1-88, 4 maio 2016. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj. Accessed on: 18 out. 2026.

VARGAS, Francielle et al. HateBR: a large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE (LREC), 13., 2022, Marseille. Proceedings [...]. Paris: ELRA, 2022. p. 7174–7183. Available at: https://aclanthology.org/2022.lrec-1.777. Accessed on: 18 out. 2026.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VII, v.1 2026 | submissão: 15/11/2026 | aceito: 22/11/2026 | publicação: 27/11/2026

WHATSAPP. Como manter a segurança no WhatsApp. [S. l.], [2026?]. Available at: https://faq.whatsapp.com/1313491802751163/?locale=pt_BR. Accessed on: 18 out. 2026.

WHITTLE, Helen et al. A review of online grooming: characteristics and concerns. Aggression and Violent Behavior, v. 18, n. 1, p. 62-70, 2013.

WIENER, Cassandra. Seeing what is 'invisible in plain sight': policing coercive control. The Howard Journal of Crime and Justice, v. 56, n. 4, p. 500-515, 2017. DOI: 10.1111/hojo.12227.

WISNIEWSKI, Pamela et al. Parental control vs. teen self-regulation: is there a middle ground for mobile online safety? In: ACM CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK AND SOCIAL COMPUTING (CSCW), 2017, Portland. Proceedings [...]. New York: ACM, 2017. p. 51–69. DOI: 10.1145/2998181.2998352.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.