# Towards Proactive Cyber Social Security: A Phase-Aware AI Architecture for Online Grooming Prevention

Alessia Anna Catalano[1,*,†], Roberto Vadacca[2,†], Christian Catalano[3], Corrado Aaron Visaggio[4] and Antonietta Curci[5]

[1]University of Salento, Dept. Law Science, Via per Monteroni 165, Lecce, 73100, Apulia, Italy
[2]University of Salento, Dept. of Innovation Engineering, Via per Monteroni 165, Lecce, 73100, Apulia, Italy
[3]University of Bari Aldo Moro, Dept. of Computer Science, Via E. Orabona 4, Bari, 70125, Apulia, Italy
[4]University of Foggia, Dept. of Agricultural Sciences, Food, Natural Resources and Engineering, Via Napoli 25, Foggia, 71122, Apulia, Italy
[5]University of Bari Aldo Moro, Dept. of Education Science, Psychology and Communication Science, Via Scipione Crisanzio 42, Bari, 70122, Apulia, Italy

**Abstract**

Online grooming represents one of the most pervasive forms of cyber-enabled harm against minors, yet existing computational approaches remain predominantly reactive and limited to binary detection. Such models fail to capture the relational, progressive, and psychologically manipulative nature of grooming, which unfolds across distinct stages long before explicit sexual content appears. This paper proposes a prevention-oriented, human-centered framework for detecting the phases of grooming as defined by O'Connell through analysis of online conversations. The architecture integrates parental-consent–based data acquisition, large language models (LLMs) with in-context learning, chain-of-thought reasoning, and post-hoc explainability mechanisms. A phase-annotated dataset should be developed by re-labeling PAN12 conversations at the utterance level to overcome limitations of existing binary-focused corpora. The resulting system provides interpretable, context-aware alerts aimed at enabling timely intervention while ensuring ethical alignment, transparency, and user safety.

**Keywords**

Cyber Social Security, Symbiotic AI, Online Grooming Detection, Pattern Behavior, Child Sexual Exploitation Prevention

CEUR-WS.org/Vol-4329/paper10.pdf

## 1. Introduction

Online grooming represents one of the most pervasive forms of digital harm against minors, enabled by the widespread diffusion of communication platforms and the increasing integration of online and offline life [1]. Global evidence underscores the severity of the phenomenon: more than 300 million children are subjected to online sexual abuse or exploitation every year [2], and in Europe up to 20% of minors report unwanted sexual requests or grooming attempts before the age of 18 [1, 3]. Moreover, reports of child sexual exploitation material have increased by 87% between 2019 and 2023 [4], confirming that grooming constitutes a structural risk within digital ecosystems. Despite increased awareness and computational advances, responses still mainly focus on detecting explicit material after the fact. Such approaches fail to capture the progressive, relational, and psychologically manipulative nature of grooming, which often develops long before explicit sexual content emerges [5, 6]. This gap calls for a shift from reactive detection to preventive, interdisciplinary frameworks that view grooming as a phase-based process driven by manipulation, trust-building, secrecy, and escalating boundary violations. [7, 8].

*1st Workshop on Symbiotic AI for Cyber Social Security: Human–AI Collaboration for Social Resilience and Trustworthy Digital Ecosystems (SAI4CCS 2026), co-located with the 2nd Joint National Conference on Cybersecurity (ITASEC & SERICS 2026), February 09–13, 2026, Cagliari, Italy*

*Corresponding author.
†These authors contributed equally.
alessiaanna.catalano@unisalento.it (A. A. Catalano); roberto.vadacca@unisalento.it (R. Vadacca); christian.catalano@uniba.it (C. Catalano); corrado.visaggio@unifg.it (C. A. Visaggio); antonietta.curci@uniba.it (A. Curci)
0009-0007-6272-3924 (A. A. Catalano); 0009-0005-4766-6744 (R. Vadacca); 0000-0003-4038-2317 (C. Catalano); 0000-0002-0558-4450 (C. A. Visaggio); 0000-0002-0932-7152 (A. Curci)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-14

Within this perspective, the achievement of effective human–AI symbiosis depends critically on users' ability to trust AI-driven decisions and to meaningfully understand the reasoning behind them, making trustworthiness a foundational requirement rather than a secondary design goal. In this sense, Symbiotic AI (SAI) represents a promising direction for rethinking the role of artificial intelligence systems in child protection within digital environments. SAI emphasizes collaborative, explainable, and adaptive forms of human–machine interaction, in which AI augments human judgment, enhances situational awareness, and simultaneously protects children's rights [9, 10]. This approach aligns with emerging frameworks in Cyber Social Security (CSS), which interpret digital harms as phenomena situated at the intersection of technological infrastructures, social behavior, and institutional governance. In this article, we explore how computational approaches can be applied to classify grooming according to O'Connell's phase model [5], examining how linguistic, relational, and behavioral cues can support the development of preventive, interpretable, and ethically grounded SAI systems. By mapping computational methods onto the sequential dynamics of grooming, we identify opportunities and limitations in transitioning from reactive detection strategies to proactive early-warning paradigms. This work contributes to ongoing debates on trustworthy AI, digital child protection, and socio-technical security by proposing a theoretically grounded, phase-aware framework that integrates perspectives from AI, psychology, law, and the social sciences.

## 2. Background and Motivation

Online grooming is rooted in complex psychosocial processes that make early-stage manipulation difficult to detect. Offenders exploit developmental vulnerabilities and socio-emotional needs typical of childhood and adolescence, progressively transforming the interaction into a relationship of dependency and control [7, 1], through a strategic modulation consistent with relational abuse models and with forms of non-violent manipulation described in the psychological abuse literature [6, 5].

Research further highlights the use of interpersonal reinforcement, emotional gratification, and identity validation techniques, which are particularly persuasive during developmental stages in which self-construction is central [11, 12]. By presenting themselves as a trustworthy figure, the offender fosters trauma bonding, an emotional attachment dynamic generated through alternating attention, threat, and dependency [13, 14], compromising the minor's critical capacity and delaying abuse disclosure [15].

Digital social dynamics amplify these processes: perceived anonymity, idealization of the interlocutor, and peer pressure increase conversational intimacy and reduce risk assessment [16, 17]. Offenders take advantage of these conditions by constructing credible identity narratives tailored to the minor's expectations, often mimicking relational codes of emotional closeness or supportive interaction [1].

Online grooming also unfolds in environments characterized by marked informational asymmetries: offenders can gather and exploit information about the minor, whereas the minor lacks the means to recognize deceptive intentions or false identities [18]. This imbalance heightens minors' psychological and relational vulnerability, particularly when offline support networks are weakened or the adolescent is experiencing emotional fragility [12, 19].

A central element in the literature concerns environmental grooming, in which offenders manipulate parents, caregivers, or institutions to gain legitimacy, reduce supervision, and increase access to the victim [7, 20]. This process weakens protective barriers and creates institutional blind spots, further hindering early risk detection [3].

O'Connell's phase-based grooming model conceptualizes grooming as a progressive and dynamic process, articulated into multiple interconnected stages: the construction of friendship and trust, the strengthening of the emotional bond, risk assessment , victim isolation, and the gradual sexualization of the relationship. These phases are not rigidly sequential but may overlap or regress depending on the context, highlighting the adaptive nature of grooming. [5].

Because these dynamics emerge gradually through micro-variations in emotional tone, trust-building, and relational distance, detection approaches based on explicit signals, such as keywords or overtly sexual content, are inadequate in the initial stages of grooming [8]. Recent literature therefore calls for

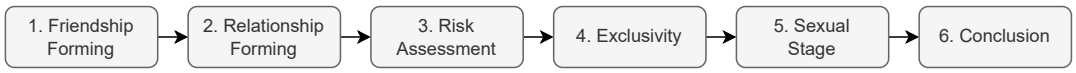



**Figure 1:** Schematic representation of the grooming stages as defined by O'Connell.

tools capable of capturing dynamic, multimodal patterns, including conversational shifts, emotional fluctuations, isolation strategies, and indicators of environmental grooming [3, 18].

Recent advances in CSS highlight the effectiveness of AI-driven detection systems that integrate automated analysis with empirical data collection, expert-informed evaluation, and educational outreach. Within this area, current research demonstrates how social sensing methods combined with multimodal GPT-4–based models can identify harmful online content and support preventive interventions in real-world scenarios [21]. Complementary work in the child-protection domain further shows that LLMs fine-tuned on targeted datasets can accurately detect age-specific linguistic patterns, sexual references, and abusive cues in textual conversations, thereby enhancing the interpretability and precision of automated safety tools [22].

Within the specific domain of online grooming detection, existing research has predominantly relied on binary classification approaches. Studies such as [23, 24, 25] report strong performance using traditional machine learning models, while more recent approaches based on transformer-based NLP architectures, including BERT and RoBERTa, achieve state-of-the-art results. The most widely adopted datasets for benchmarking include PAN12 [26], Perverted-Justice [27], and MovieStarPlanet [28]. PAN12 provides annotated chat transcripts from the "Sexual Predator Identification" task of the PAN 2012 lab, consisting of real conversations labeled with the presence of suspected predators. The Perverted-Justice dataset contains transcripts published by the volunteer organization of the same name, whose members posed as minors to expose adults seeking sexual interactions. The MovieStarPlanet dataset, collected from the popular online platform for children and adolescents, consists of real user conversations exported in collaboration with the service operators and annotated for risky behaviors, inappropriate requests, and grooming attempts.

Most studies, including those addressing early detection and prevention [29, 30, 31], continue to frame grooming as a binary classification problem. These models detect whether grooming is present, but they do not identify the distinct phases of the grooming process. Detecting the progression of phases is crucial for practical intervention, as it enables context-aware responses tailored to the level of threat, providing actionable guidance to potential victims and investigators. The most widely adopted datasets for benchmarking include PAN12 [26], Perverted-Justice [27], and MovieStarPlanet [28].

A further gap in the literature concerns the lack of integration strategies for deploying these algorithms in realistic operational scenarios. Existing studies primarily report classification performance on benchmark datasets without addressing how such models could be embedded into real-world monitoring or support systems. The aim of this work is to overcome this limitation by proposing a framework that performs phase-based grooming detection and outlines how such models can be incorporated into an operational system aimed at proactive online grooming prevention.

## 3. Proposal Architecture

The objective of this work is to develop a prevention-oriented tool capable of identifying the specific stages of online grooming as defined by O'Connell (Figure 1). Framing grooming as a multi-stage phenomenon enables timely, context-appropriate interventions rather than issuing a generic alert once a conversation is already severely compromised.

The use case falls within the broader domain of parental-control technologies. Any operational deployment requires installation and consent from a parent or guardian, who authorizes the collection and processing of a child's online chat data.

In a real-world scenario, the system operates through a browser or application extension capable of extracting conversational streams from messaging platforms, online games, forums, and social media.

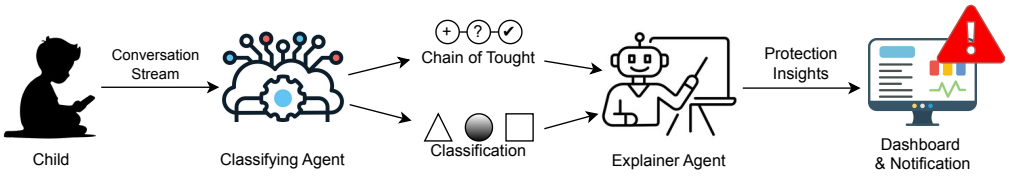



**Figure 2:** Pipeline showing conversation ingestion, stage classification with chain-of-thought, and post-hoc educational insights delivered via dashboard notifications.

With explicit parental permission, the extension captures the textual content of ongoing chats and forwards it to the detection module. This architecture enables low-latency processing and supports classification necessary for preventive action.

A critical component of the framework is the availability of a suitable dataset for multi-stage grooming classification, which is currently lacking. Existing resources such as PAN12 provide annotated conversations identifying the presence of sexual predators but do not include labels aligned with O'Connell's stages. Filling this gap would require a dedicated re-annotation effort involving domain experts to assign each utterance to a grooming phase. The resulting dataset would support both supervised evaluation and the exploration of in-context learning strategies.

A high-level overview of the proposed pipeline is shown in Figure 2. For the classification task, the system leverages LLMs capable of in-context learning. This approach allows the classifier to adapt dynamically to conversational context without requiring extensive task-specific fine-tuning. The chain-of-thought mechanism guides the model to articulate intermediate reasoning steps, which are later distilled into user-friendly explanations to support transparency and auditability. Model performance will be assessed using the newly annotated dataset, ensuring evaluation under conditions aligned with the intended operational environment.

From an architectural perspective, the entire pipeline integrates data ingestion from the parental-control extension, LLM-based stage classification, and generation of human-readable explanations. The system is designed to operate with minimal intrusion, providing escalating forms of guidance depending on the detected grooming stage. Within the proposed architecture, SAI plays a central role in coordinating automated phase detection with human-centered intervention strategies. Rather than acting as an autonomous decision-maker, the system is designed to function as an assistive layer that interprets conversational dynamics and translates model inferences into intelligible, stage-aware insights for guardians.

## 4. Conclusion

This work argues for moving beyond traditional binary classification approaches to online grooming detection toward a phase-based and preventive paradigm grounded in O'Connell's model. Within a Symbiotic AI framework, effective protection requires systems that capture the evolving dynamics of conversations, enabling earlier and more context-aware interventions.

Our analysis shows that grooming often develops through subtle socio-relational cues that are overlooked by datasets focused primarily on explicit sexual content. To overcome this limitation, we propose an architecture capable of identifying grooming phases as they unfold, integrating consent-based data acquisition, LLMs through in-context learning, and explainability mechanisms to ensure transparency and ethical alignment.

Key contributions of this study are the formulation of a phase-aware framework for online grooming prevention, the articulation of human-centered, explainable LLM-based analysis as a means for early intervention, and the motivation for utterance-level re-annotation of existing datasets to enable phase-based evaluation within a CSS and SAI perspective emphasizing ethics, transparency, and consent.

Future work should focus on the empirical validation of the proposed architecture, evaluating explainability for non-expert users, refining data governance strategies, and testing the integration

within parental control tools in real-world scenarios.

## Declaration on Generative AI

The authors have not employed any Generative AI tools.

## References


[1] H. Whittle, C. Hamilton-Giachritsis, A. Beech, G. Collings, A review of online grooming: Characteristics and concerns, Aggression and violent behavior 18 (2013) 62–70.

[2] J. Hill, Losing it: Can we stop violence against women and children?, Quarterly Essay (2025) 1–121.

[3] S. Webster, J. Davidson, A. Bifulco, P. Gottshalk, V. Caretti, T. Pham, J. Grove Hills, C. Turley, C. Tompkins, S. Ciulla, et al., European online grooming project, 2012.

[4] C. Setter, Estimates of childhood exposure to online sexual harms and their risk factors, 2023.

[5] R. O'Connell, A typology of child cybersexploitation and online grooming practices, Cyberspace Research Unit, University of Central Lancashire (2003) 6–10.

[6] K. V. Lanning, Child molesters: A behavioral analysis for professionals investigating the sexual exploitation of children, 2010.

[7] S. Craven, S. Brown, E. Gilchrist, Sexual grooming of children: Review of literature and theoretical considerations, Journal of sexual aggression 12 (2006) 287–299.

[8] G. M. Winters, E. L. Jeglic, Stages of sexual grooming: Recognizing potentially predatory behaviors of child molesters, Deviant behavior 38 (2017) 724–733.

[9] D. Dellermann, P. Ebel, M. Söllner, J. M. Leimeister, Hybrid intelligence, Business & Information Systems Engineering 61 (2019) 637–643.

[10] B. Shneiderman, Human-centered artificial intelligence: Reliable, safe & trustworthy, International Journal of Human–Computer Interaction 36 (2020) 495–504.

[11] W. Meeus, Adolescent psychosocial development: A review of longitudinal models and research., Developmental psychology 52 (2016) 1969.

[12] J. Nesi, The impact of social media on youth mental health: challenges and opportunities, North Carolina medical journal 81 (2020) 116–121.

[13] D. G. Dutton, S. Painter, Emotional attachments in abusive relationships: A test of traumatic bonding theory, Violence and victims 8 (1993) 105.

[14] D. F. Walker, H. W. Reid, T. O'Neill, L. Brown, Changes in personal religion/spirituality during and after childhood abuse: A review and synthesis., Psychological Trauma: Theory, Research, Practice, and Policy 1 (2009) 130.

[15] H. C. Whittle, An exploration of young people's vulnerabilities to online grooming and sexual abuse, Ph.D. thesis, University of Birmingham, 2014.

[16] S. Livingstone, P. K. Smith, Annual research review: Harms experienced by child users of online and mobile technologies: The nature, prevalence and management of sexual and aggressive risks in the digital age, Journal of child psychology and psychiatry 55 (2014) 635–654.

[17] E. Staksrud, K. Ólafsson, Tilgang, bruk, risiko og muligheter. norske barn på internett. resultater fra eu kids online-undersøkelsen i norge 2018, 2019.

[18] B. G. Westlake, M. Bouchard, Criminal careers in cyberspace: Examining website failure within child exploitation networks, Justice Quarterly 33 (2016) 1154–1181.

[19] S. Livingstone, A. Third, Children and young people's rights in the digital age: An emerging agenda, 2017.

[20] K. V. Lanning, P. Dietz, Acquaintance molestation and youth-serving organizations, Journal of Interpersonal Violence 29 (2014) 2815–2838.

[21] A. A. Catalano, C. Catalano, E. Ingusci, D. Caivano, Ai and digital awareness: An integrated

approach for the recognition and prevention of cyberbullying through visual content, Workshop: Cyber Security Education for Industry and Academia (CSE4IA 2025) (2025).

[22] M. T. Baldassarre, V. Santa Barletta, V. Bavaro, D. Caivano, A. P. De Matteis, A. Lippolis, A. Piccinno, Llms to detect cyber child abuse in textual conversations, Proceedings of A²ID — 1st International Workshop on Augmented Artificial Intelligence: a Symbiotic Approach to Enhance Decision-Making (2025).

[23] P. Rezaee Borj, K. Raja, P. Bours, Detecting online grooming by simple contrastive chat embeddings, in: Proceedings of the 9th ACM international workshop on security and privacy analytics, 2023, pp. 57–65.

[24] H. Waezi, R. Barzegar, H. Fani, Osprey: A reference framework for online grooming detection via neural models and conversation features, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 2024, pp. 4101–4106.

[25] J. Street, I. K. Ihianle, F. Olajide, A. Lotfi, Enhanced online grooming detection employing context determination and message-level analysis, Intelligent Systems with Applications 28 (2025) 200607. URL: https://www.sciencedirect.com/science/article/pii/S2667305325001334. doi:https://doi.org/10.1016/j.iswa.2025.200607.

[26] G. Inches, F. Crestani, Pan12 deception detection: Sexual predator identification, 2012. URL: https://doi.org/10.5281/zenodo.3713280. doi:10.5281/zenodo.3713280.

[27] P. J. Eide, perverted-justice, WebPage, 2002. URL: http://www.perverted-justice.com/.

[28] Y.-G. Cheong, A. K. Jensen, E. R. Guðnadóttir, B.-C. Bae, J. Togelius, Detecting predatory behavior in game chats, IEEE Transactions on Computational Intelligence and AI in Games 7 (2015) 220–232. doi:10.1109/TCIAIG.2015.2424932.

[29] A. P. López-Monroy, F. A. González, M. Montes, H. J. Escalante, T. Solorio, Early text classification using multi-resolution concept representations, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1216–1225. URL: https://aclanthology.org/N18-1110/. doi:10.18653/v1/N18-1110.

[30] D. F. Milon-Flores, R. L. Cordeiro, How to take advantage of behavioral features for the early detection of grooming in online conversations, Knowledge-Based Systems 240 (2022) 108017.

[31] D. Kim, T. Kim, J. Yang, Early detection of online grooming with language models, in: Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, 2025, pp. 963–970.