Symbiotic AI for Detecting and Mitigating Online Child Sexual Abuse Material

Gabriel Cellammare[1,3], Francesca Mantini[1], Vito Marchionna[1], Massimiliano Morga[2] and Davide Pio Posa[1]

[1]Cyber Academy for Security and Intelligence (CrASte)
[2]SER&Practices, Spin-off of the University of Bari Aldo Moro
[3]IMT School for Advanced Studies Lucca

Abstract

The increasing convergence between digital technologies and social ecosystems necessitates cybersecurity research that addresses not only technical vulnerabilities but also social, psychological, and ethical harms. Among these, Online Child Sexual Abuse and the global proliferation of Child Sexual Abuse Material represent one of the most critical challenges for Cyber Social Security. Building upon a systematic review of AI-based CSAM detection methods, including cryptographic hashing, source camera identification, metadata-based detection, and machine learning models, this work examines the strengths, limitations, and ethical constraints of current approaches. Therefore, the paper analyzes the need for human-centered, explainable, and ethically aligned Artificial Intelligence (AI) systems. The SAI4CSAM (Symbiotic Artificial Intelligence for Detecting and Monitoring Child Sexual Abuse) framework is proposed, as an adaptive, resilient, and ethically aligned logical framework for detecting and mitigating CSAM. This framework identifies four layers, the integration of which highlights how symbiotic AI can effectively enable collaborative human-AI ecosystems that integrate human-in-the-loop validation, multimodal social sensing, and transparent decision pipelines to support trustworthy socio-digital environments.

Keywords

Child Sexual Abuse, Cyber-Social Security, SAI, CSAM, Human-AI

1. Introduction

Compared to the past decade, there is undoubtedly greater sensitivity and active interest in cybersecurity research, along with a growing awareness that it can no longer be limited to analyzing and addressing technical vulnerabilities alone, but must necessarily broaden its scope. Modern threats have evolved and begun to increasingly target the cognitive and social dimensions of users, exploiting psychological manipulation and social trust dynamics [1]. This evolution requires a consequent change in the traditional cybersecurity paradigm in favor of what is known as Cyber Social Security[1] (CSS), a multidisciplinary field that addresses security not as a purely technical safeguard but takes into account the human factor, cognitive models, and the impact of cyber threats on society, culture, and politics [2]. Acknowledging how individuals often represent the weakest point in the cybersecurity landscape, CSS focuses on analyzing everything that has an impact on IT within a social context and on designing IT infrastructures that protect society from cyber threats [3, 4].

Within this field of study, Child Sexual Abuse (CSA) stands out as one of the most critical social threats [5, 6]. Despite the ongoing efforts of law enforcement agencies (LEAs), the scale of the problem is reaching alarming proportions, further exacerbated by the proliferation of Child Sexual Abuse Material (CSAM) online, the recent increase in content generated by Artificial Intelligence (deepfakes), and

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

gabriel.cellammare@imtlucca.it (G. Cellammare); f.mantini@studenti.uniba.it (F. Mantini); v.marchionna@studenti.uniba.it (V. Marchionna); m.morga@serandp.com (M. Morga); d.posa3@studenti.uniba.it (D. P. Posa)
0009-0001-1570-4487 (G. Cellammare); 0009-0009-7337-5018 (F. Mantini); 0009-0009-0257-3367 (D. P. Posa)

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

[1]https://craste.uniba.it/cyber-social-security/

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

CEUR-WS.org/Vol-4329/paper11.pdf

children's increasingly early access to the internet[2]. Both the International Criminal Police Organization[3] (INTERPOL) and the National Center for Missing and Exploited Children[4] (NCMEC) have recorded and highlighted an exponential increase in reports of CSAM online compared to the previous decade. This is certainly partly attributable to greater detection capabilities and methods. Still, it is also undoubtedly a figure that needs to be analyzed while remaining aware of how it reflects only a limited part of the true extent of the phenomenon. CSAM represents an especially severe form of re-victimization [7], as it not only creates a permanent record of the abuse suffered, but also exposes the victim to the possibility of its continued circulation and persistence online. This enduring availability amplifies the psychological trauma, forcing the victim to confront the distressing awareness that the documentation of their abuse may be accessed and viewed by an unlimited number of individuals.

Therefore, it is necessary to analyze the current methodologies mainly adopted to support CSAM detection [8]. Recent research examines the dual role of Artificial Intelligence in CSAM, highlighting both its potential for facilitating crimes and detecting abuse. Petrini and Guana [9] conducted qualitative analyses to demonstrate how Artificial Intelligence (AI) facilitates the production and distribution of CSAM, to reveal inadequate legislation at national and international levels, and to emphasize the need for collaborative efforts among governments, organizations, and companies to establish preventive regulatory frameworks. Wolbers et al. [10] performed a rapid evidence assessment of 33 empirical studies from 2010 to 2024, finding that all research focused exclusively on AI for CSA prevention and disruption, specifically detecting CSAM and investigating offenders, with no studies examining AI's role in offending. Instead, in [11], Singh and Nambiar conducted a systematic review identifying 35 papers on AI interventions for online child sexual abuse, revealing three major themes: current AI trends, algorithm evaluation, and recommendations for technical advancements, though they noted scarce evidence proving AI intervention effectiveness. The contribution of these techniques is certainly fundamental, but, given the technical challenges, such as robustness against adversarial attacks, the increase in deepfake content, and encryption, and considering the ethical and legal constraints that arise from them, it is clear that it is not sufficient to properly manage the nuances and sensitivity of CSAM detection in a socio-cybernetic context.

In order to fill these gaps, this research proposes the adoption of Symbiotic Artificial Intelligence (SAI) to enhance the identification and analysis of CSAM within a CSS perspective [12]. SAI goes beyond simply placing humans at the core by fostering ongoing interaction where both entities enhance one another [13]. SAI starts from the concept of Human-In-The-Loop, and upgrades it by promoting a collaborative ecosystem in which human intelligence (investigators, psychologists, legal experts) and artificial intelligence co-evolve. By combining automated perceptual capabilities with human cognitive judgment, the proposed SAI approach seeks to establish a more adaptive, resilient, and ethically aligned framework for CSAM detection, ultimately strengthening the broader objectives of Cyber Social Security.

The main contribution of this research work is to identify a logical framework called SAI4CSAM (Symbiotic Artificial Intelligence for Detecting and Monitoring Child Sexual Abuse) that identifies four layers (Figure 1): (i) Detection Layer; (ii) Symbiotic AI Layer; (iii) Socio-Digital Contextualization Layer; (iv) Governance and Ethical Layer. The integration of these layers contributes to the identification of online material relating to CSA, but also sextortion of minors, revenge porn, and cyberbullying.

The paper is structured as follows: Section 2 overviews the multi-modal engines for content-based, metadata-driven, forensic, and hashing-based CSAM detection. Section 3 presents the Symbiotic AI Layer, combining human oversight, co-learning, and transparency mechanisms. Section 4 introduces the Socio-Digital Contextualization Layer, enabling cross-platform behavioral analysis, grooming-pattern detection, and dark web intelligence. Section 5 addresses the Governance and Ethical Layer, detailing the embedded legal and regulatory requirements. Section 6 reports a practical use case integrating the NudeNet model into the proposed architecture, while Section 7 concludes the paper and outlines future

[2]https://clusit.it/rapporto-clusit/

[3]https://www.interpol.int/Crimes/Crimes-against-children/International-Child-Sexual-Exploitation-database

[4]https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata

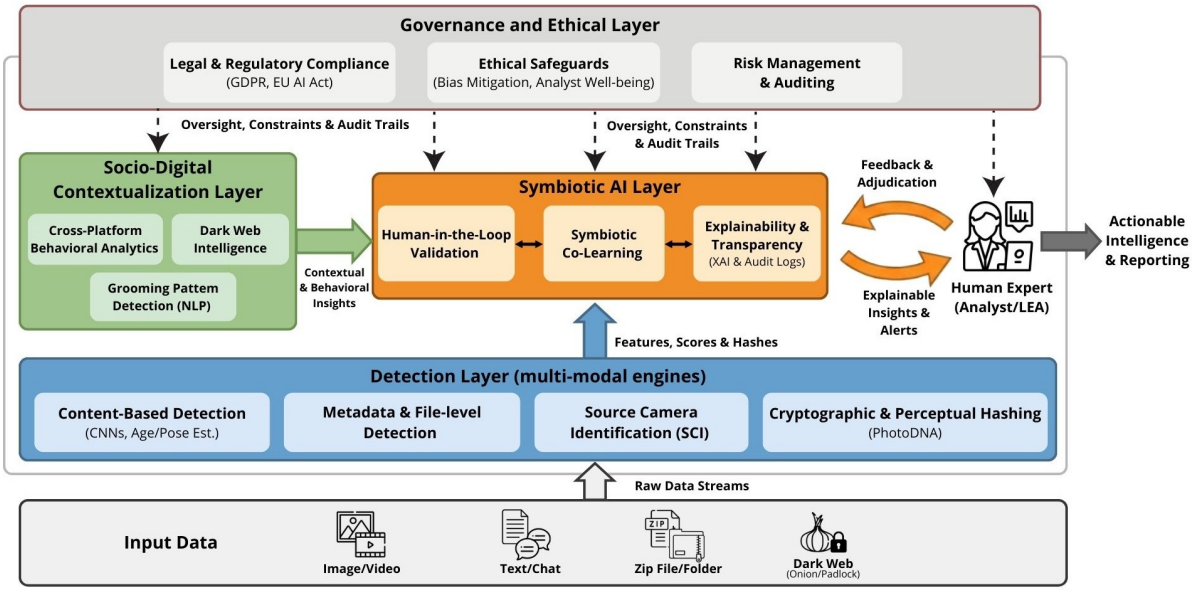


Figure 1: The SAI4CSAM Framework Architecture: layers and component interaction

research directions.

2. Detection Layer (multi-modal engines)

The Detection Layer constitutes the technical core of the SAI4CSAM framework. It integrates heterogeneous computational methods for detecting CSAM by combining traditional forensic approaches, modern deep learning techniques, and privacy-preserving strategies. This section provides a structured description of the four primary components of the layer: (i) *Content-Based Detection*, (ii) *Metadata and File-level Detection*, (iii) *Source Camera Identification*, and (iv) *Cryptographic and Perceptual Hashing*.

2.1. Content-Based Detection

Content-based approaches analyze the visual and textual characteristics of images, videos, and conversational data to identify potential CSAM. Convolutional Neural Networks (CNNs) represent the dominant paradigm in this domain due to their ability to extract discriminative representations from complex visual scenes [14]. As demonstrated by Rondeau et al. [15] and Vitorino et al. [16], CNNs trained on explicit and non-explicit corpora enable efficient classification, localization of sensitive content, and semantic filtering. Age estimation models play a central role in distinguishing between minors and adults, particularly when explicit content is present. Architectures such as DenseNet-161[5] and the attention-driven AttM-CNN [17] leverage deep feature extraction and spatial attention to produce accurate estimations of apparent age, improving the reliability of CSAM detection pipelines. The pornography-detection and age-group-classification networks are integrated to support CSA detection through two complementary strategies: decision-level fusion for the binary classification of CSA content, and score-level fusion for prioritizing and re-ordering suspicious images.

Complementary techniques include nudity and pose estimation, which support the semantic interpretation of body configurations and the identification of risky visual patterns. These models, used in conjunction with CNN-based classifiers, enhance contextual understanding and reduce false positives [15]. Given the legal and ethical implications of CSAM detection, explainability mechanisms are essential. Techniques such as Grad-CAM [18] provide visual evidence of model decisions by highlighting

[5]https://www.kaggle.com/datasets/pytorch/densenet161

salient regions of interest, supporting forensic auditability and analyst trust [19].

2.2. Metadata and File-level Detection

Metadata-based and file-level detection approaches provide an alternative path for identifying CSAM without direct exposure to explicit visual content. Pereira et al. [20] demonstrated that machine learning models trained on file paths can achieve 97% accuracy in CSAM detection while maintaining a low false positive rate (0.002) and showing resilience against adversarial attacks. This metadata-based approach circumvents ethical and legal constraints associated with collecting actual CSAM imagery for training purposes. In [21] is shown how multi-modal approaches combining deep learning with various detection methods, including filename and metadata analysis, yield superior results compared to single-method approaches. Westlake & Guerra [22] analyzed dark web CSAM distribution patterns, finding that websites prioritize organization over obfuscation, with explicit naming conventions and identifiable file structuring patterns that can enhance automated detection. Instead, Ngo et al. [23] developed systems capable of real-time CSAM classification while extracting valuable metadata insights about creator behaviors and victim characteristics from dark web forums.Therefore, file name and file-path classification methods model strings containing abbreviations, obfuscated vocabulary, and criminal jargon through short-text machine learning techniques [24]. These methods are particularly suitable for initial triage, reducing analyst exposure and supporting privacy-sensitive workflows.

2.3. Source Camera Identification

Source Camera Identification (SCI) is a digital forensic technique used to determine which camera or acquisition device produced a given image or video. It exploits the peculiarity of cameras, which have unique physical and electronic properties, resulting in "imperfections" or "noise" in the captured content that can be treated as the device's distinctive "signature." A common approach is Pattern Noise Analysis, which investigates the characteristics of the electronic noise generated by the sensor. At the same time, Fixed Pattern Noise (FPN) is identified, which represents the constant and non-uniform structural defects present in the sensor. The chrominance and luminance components of the image are then processed through the application of machine learning algorithms to extract distinctive features that unequivocally attribute the image to a specific device. Irshad et al.[25] introduce a new camera identification methodology that addresses the challenges posed by modern image manipulation methodologies. In a context where digital editing and seam carving techniques [6] are increasingly being used to disguise the origin of images, especially in the CSAM market, this method uses a One-Dimensional Convolutional Neural Network (1D CNN) to analyze the PRNU (Photo Response Non-Uniformity) characteristics of images, where PRNU refers to a technique of identifying the source of an image based on the numerical fingerprints generated by the sensor of a digital camera. Each image sensor (CCD or CMOS) has slight imperfections and variations in the response of pixels to light, which generate a unique and almost invariable pattern. This structural 'noise', invisible to the human eye but detectable by digital processing, functions as a unique identification fingerprint of the device that captured the image. The methodology demonstrates high accuracy in attributing the camera source, achieving state-of-the-art performance in identifying three classes: original images, images with stitching removed, and images with stitching inserted. However, research in this field faces several challenges, such as the variability of acquisition conditions, image degradation, the complexity of identification algorithms, and the need for constantly updated reference databases.

2.4. Cryptographic and Perceptual Hashing

Traditional cryptographic hashing is an essential component of computer security. While it enables verification of a digital document's authenticity, integrity, and ensures that such documents can be

[6] Image processing technique that allows the reduction or dynamic adjustment of the size of an image while preserving its main content.

reliably associated with their authors, it is brittle against minor file alterations. In the specific field of Content Security, *Perceptual Hashing* is preferred as it allows for the identification of explicit content even after minor modifications [26]. An example, adopted as an industry standard, is the PhotoDNA technique, widely used by organizations such as NCMEC, Facebook, and Microsoft. Unlike cryptographic hashes, PhotoDNA computes a gradient-based signature robust to resizing, compression, and minor color alterations. However, as highlighted by Steinebach et al. [27], this approach has significant limitations: it fails to detect content that has undergone substantial transformations, such as cropping, heavy rotation, or watermark insertion, necessitating more adaptive approaches.

In encrypted environments, homomorphic hashing and private membership tests allow systems to verify whether a piece of content appears in a CSAM hash database without decrypting the original material. These approaches significantly enhance privacy preservation and operational scalability [28]. In addition, in distributed and cloud-based settings, fragment hashing further expands detection capabilities. In [29, 30], the authors demonstrate that partitioning files into fragments enables identification even when only partial or corrupted content is available, a common scenario in peer-to-peer and dark web ecosystems.

3. Symbiotic AI Layer

The Human–AI Symbiotic Layer constitutes the core of the proposed framework, enabling a collaborative interaction between human experts and AI systems. This layer emphasizes adaptive, explainable, and human-centered intelligence to ensure robust decision-making in sensitive cyber-social contexts.

The human-AI collaboration in deep learning combines AI computational power with human domain expertise to improve performance and create value [31]. Hao et al. [32] introduce Symbiotic Artificial Intelligence with Shared Sensory Experiences (SAISSE), which establishes mutually beneficial relationships through shared sensory input channels and personalized adaptation to individual users. Instead, in [33] are identified four key principles for designing Human-Centered AI systems that comply with the EU AI Act's regulatory framework: (i) *Transparency* ensures that AI operations are understandable and overseen by humans; (ii) *Fairness* ensures equality and inclusiveness while preventing bias and discrimination; (iii) *Automation Level*, that balancing automation and human control is crucial for safe AI-human interaction; (iv) *Protection* prioritizes safeguarding users against harm, ensuring privacy, safety, and security. As emphasized in [34], AI should augment rather than replace human intelligence, creating synergistic relationships with transformative potential across healthcare, education, and business sectors. This collective research highlights symbiotic AI as a promising direction for responsible AI development that enhances human capabilities.

3.1. Human-in-the-Loop Validation

The Human-in-the-Loop component integrates analyst-aware feedback loops to continuously refine model outputs. Expert annotations are leveraged to correct and enhance AI predictions, while real-time adjudication mechanisms allow humans to intervene in borderline or ambiguous cases. Such interactions ensure that AI outputs remain aligned with domain expertise and legal/ethical constraints, mitigating risks associated with fully automated decision-making.

From a governance perspective, Human-in-the-Loop validation is also critical for ensuring compliance with legal and ethical requirements, including due process obligations, proportionality principles, data protection norms, and the transparency criteria mandated by emerging regulatory frameworks (e.g., EU AI Act). Human oversight provides a safeguard against opaque, unexplainable, or discriminatory decisions, allowing experts to contest system outputs, adjust decision thresholds, and introduce contextual knowledge that AI systems may not capture. Therefore, this component transforms the human role from passive reviewer to active co-supervisor of AI systems. This symbiotic arrangement not only enhances operational performance, but also strengthens institutional trust and accountability, ensuring that automated systems remain aligned with human expertise, societal expectations, and the ever-evolving regulatory landscapes.

Below is an example of how this component works within CSAM detection. Consider a scenario in which the system encounters an encrypted archive named: `IMG_2024-03-xx_yng-special-ver2.zip`. Although the model detects several indicators historically associated with CSAM obfuscation — such as token sequences ("yng", "special", "verX"), the presence of a versioned ZIP container, and fuzzy hash distances around 0.70 relative to known CSAM filename clusters — the confidence score remains below the automated reporting threshold (e.g., 0.80). Rather than generating an automatic alert, the system escalates the case to a trained analyst.

The analyst, who never views image content, examines contextual non-visual indicators such as: the file-path structure (e.g., `/Downloads/TG-Backup/Forwarded/`, a directory commonly associated with messaging-platform extractions); metadata inconsistencies within the ZIP archive; multi-device EXIF traces suggesting aggregated sourcing; internal filename tokens such as `pic_set_0xA3_rare`, which align with known distribution patterns in offender networks.

Based on this evidence, the analyst classifies the item as *CSAM-likely* and provides structured justification codes (e.g., filename-pattern, ZIP-structure anomaly, EXIF discrepancy, hash-distance relevance). This human judgment is then fed back into the learning pipeline as a labelled correction.

The system uses the analyst's decision to:

- update filename-classification models, improving tokenization of emerging obfuscation patterns (e.g., `yng` + `verX` + `rare`);
- recalibrate fuzzy-hashing thresholds, especially for multi-indicator categories where borderline distances can legitimately represent illicit content;
- strengthen anomaly-scoring modules, which integrate metadata convergence, path analysis, and device-source irregularities;
- expand the knowledge base with newly discovered obfuscation tactics (e.g., versioned ZIP distributions or rare-token embeddings).

3.2. Symbiotic Co-Learning

Symbiotic Co-Learning extends Human-in-the-Loop mechanisms by establishing a continuous, bidirectional exchange through which human expertise and machine learning models adapt to each other. Rather than operating through sporadic corrections or post-hoc adjustments, this component enables AI systems to integrate analyst knowledge into their internal representations over time, while simultaneously providing experts with insights on emerging trends, anomaly patterns, and latent structures identifiable only through large-scale computational processing. This reciprocal learning dynamic supports resilience against adversarial evasion, accelerates model convergence, and ensures that conceptual drifts in offender behavior are rapidly detected and operationalized. Criminals frequently adopt new obfuscation patterns — such as numeric padding, hybrid alphanumeric codes (e.g., "s3t-y0ng-v12"), or context-dependent token insertion ("art_ref", "tutorial_set", "promo") — to evade automated classifiers. When analysts begin encountering a new obfuscation scheme during casework, they annotate these tokens as high-risk, medium-risk, or contextual indicators within structured labelling interfaces. These annotations are then ingested by the learning pipeline through active learning queries, where the model prioritizes uncertain or borderline examples and requests verification to refine its boundaries.

Simultaneously, the system identifies latent clusters of filename transformations that share structural similarities with known CSAM-related naming conventions, even if human analysts have not yet observed them. For example, the model may detect a newly emerging cluster of filenames following the pattern:

```
childproj_vX_setY_raw
kidrefX_archiveZ_full
xp_yng_pack_v2
```

Although these specific forms may not explicitly match previous obfuscation tactics, embedding-based similarity metrics (e.g., character-level n-gram embeddings, subword transformers, lexical anomaly scoring) reveal structural proximities to known illicit distributions. The system surfaces these findings

to analysts in the form of candidate risk clusters, prompting human validation. Upon confirmation, these patterns are added to the knowledge base and incorporated into the next retraining cycle.

Another key function of Symbiotic Co-Learning is Reinforcement Learning from Human Feedback (RLHF) applied to CSAM triage tasks. When analysts repeatedly confirm that certain metadata combinations — such as offsets between device timestamps and system timestamps, or recurrent EXIF anomalies — strongly correlate with CSAM provenance, the reward function is updated to increase the weighting of these features. Conversely, patterns that frequently lead to false positives, such as filenames containing tokens like "baby_shower", "kids_project", or "family_album", receive negative reinforcement, reducing their influence and improving system precision.

3.3. Explainability & Transparency

The Explainability and Transparency component provides the structural foundation for Trustworthy AI[7], ensuring that system outputs are interpretable, audit-ready, and aligned with legal and ethical requirements. In environments involving sensitive content, high-risk decisions, or regulatory oversight, explainability is not an optional feature, but a mandatory pillar supporting accountability and human supervision.

This component incorporates in itself a multi-level interpretability framework, which offers explanatory insights at distinct granularity levels.

- *Model-level interpretability* elucidates the overarching logic of the system, describing how models transform inputs into outputs and how decision pathways are structured.
- *Feature-level interpretability* provides visibility into which linguistic patterns, metadata attributes, behavioral patterns, or latent embeddings influenced a given prediction.
- *Decision-level interpretability* focuses on local explanations for individual outputs, helping analysts understand why a specific case was flagged, how confidence scores were derived, and what evidence contributed to the system's assessment.

To support governance and auditability, the component integrates comprehensive audit logs and traceability mechanisms to capture metadata on model versions, input transformations, human interventions (e.g., overrides, corrections), and decision rationales. Such records are critical for regulatory compliance, particularly under frameworks like the EU AI Act, which require demonstrable accountability and documentation of automated decision-making processes. Furthermore, the system integrates role-specific justification interfaces designed to meet the distinct operational and strategic needs of law enforcement agencies managing CSAM detection platforms.

4. Socio-Digital Contextualization Layer

The Socio-Digital Contextualization Layer provides a comprehensive framework for integrating cross-platform threat intelligence into human–AI symbiotic systems, particularly in domains such as CSAM detection and digital risk monitoring. Unlike models that operate on single-platform signals, this layer contextualizes data from diverse digital ecosystems, including social media, messaging platforms, file-sharing networks, and dark-web forums. By combining heterogeneous inputs, it offers a holistic understanding of threat actors, content propagation dynamics, and behavioral patterns.

By synthesizing multi-platform behavioral insights, grooming indicators, and dark web intelligence, this layer transforms isolated patterns into rich socio-digital understanding, enabling proactive detection of coordinated abuse, early intervention, and adaptive policy formulation. Its design emphasizes operational effectiveness, ethical safeguards, and compliance, making it a strategic backbone for trustworthy, human-aligned AI in high-risk cyber environments.

[7] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

4.1. Cross-Platform Behavioral Analytics

A core function of the layer is aggregating and analyzing user behavior across multiple platforms. Data streams are normalized and semantically aligned to allow temporal, linguistic, and structural correlations across disparate sources. Behavioral features, such as posting frequency, content modification patterns, device usage, and relational interactions, are extracted to construct multidimensional profiles of users. These profiles enable predictive modeling of emerging threats, early identification of high-risk actors, and the detection of coordinated campaigns that would be invisible when analyzing platforms in isolation. By integrating feedback from Human-in-the-Loop analysts, the system continuously adapts, refining its understanding of legitimate versus suspicious behaviors while preserving ethical and legal compliance.

4.2. Grooming Pattern Detection

This module focuses specifically on identifying grooming behaviors, where criminals attempt to manipulate, coerce, or build trust with potential victims. By leveraging cross-platform behavioral features and natural language analysis, the system detects patterns of interaction indicative of grooming, such as repeated private messaging, solicitations, or the use of age-appropriate obfuscations. It incorporates temporal analysis to identify sequences of escalating engagement, and semantic modeling to recognize subtle linguistic cues, including euphemisms, codewords, or manipulative phrasing. The output of this module is presented to analysts as actionable insights, with confidence scores, contextual explanations, and alert prioritization, enabling informed intervention without exposing them to visual content.

4.3. Dark Web Intelligence Module

The modules provide visibility into hidden or illicit digital environments where CSAM and related content are frequently distributed. The system monitors marketplaces, forums, file-sharing portals, and other anonymized services to identify emerging distribution patterns, obfuscation strategies, and offender networks. Metadata analysis, content embeddings, and network mapping are employed to correlate dark-web activity with signals observed on open platforms. By integrating these insights, analysts gain foresight into new threats, including novel file structures, distribution channels, and obfuscation techniques, which can then inform model updates and policy decisions.

4.4. Integration and Analyst Feedback

Information from cross-platform behavioral analytics, grooming detection, and dark web modules is presented through role-specific dashboards. Analysts receive detailed timelines of suspicious activity, network visualizations of actor relationships, and contextual explanations for flagged content. Human feedback, including validation, annotation, and adjudication, is incorporated into the models, forming a co-evolutionary loop that strengthens predictive accuracy and operational efficiency. This approach ensures that AI-assisted decision-making is interpretable, traceable, and aligned with legal and ethical obligations.

5. Governance and Ethical Layer

The Governance and Ethical Layer serves as a critical oversight mechanism within AI-enabled CSAM detection systems, ensuring that operational activities adhere to legal, regulatory, and ethical standards. This layer operates at the intersection of technical performance and societal responsibility, embedding compliance, accountability, and transparency into the entire detection workflow. Its primary objective is to guarantee that the deployment of AI technologies aligns with both human rights considerations and institutional obligations. From a legal and regulatory perspective, this layer ensures strict adherence to privacy laws, child protection statutes, and national and international data protection frameworks,

including GDPR[8](General Data Protection Regulation) and directives from relevant law enforcement authorities like EDPB-EDPS[9] (Joint Opinion 04/2022 on the Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse). From a legal and regulatory perspective, this layer ensures strict adherence to privacy laws, child protection statutes, and national and international data protection frameworks, including GDPR and directives from relevant law enforcement authorities. It enforces the minimization of human exposure to harmful content by promoting metadata-driven and content-blind analysis wherever possible, and it manages the lawful retention, sharing, and auditing of sensitive information across agencies and jurisdictions.

In terms of ethical compliance, the layer operationalizes principles such as proportionality, due process, and fairness. It establishes protocols to monitor and mitigate algorithmic bias, prevent discriminatory outcomes, and protect the psychological well-being of analysts by controlling exposure to potentially harmful material. The risk management function within this layer evaluates potential harms associated with false positives and false negatives, prioritizes intervention strategies, and ensures that automated actions are aligned with societal and institutional expectations. By integrating continuous monitoring of model performance, human-in-the-loop feedback, and cross-platform intelligence insights, the Governance and Ethical Layer provides a structured framework for balancing operational efficiency with legal, moral, and societal imperatives.

6. CSAM: Use Case

In line with the proposed Symbiotic Artificial Intelligence (SAI) paradigm, we introduce the *NudeNet* model into our detection workflow. This model leverages the high efficiency of the YOLOv8 architecture [35], and it has been specifically designed for advanced visual analysis of sexually explicit content [36]. The peculiarity of NudeNet lies in its detailed classification of images into distinct body parts (e.g., genitals, breast, buttocks, etc.), each one associated with a confidence score. The outputs are aggregated and processed, then the images are classified into **Nude** (image containing nude), **Sexy** (images containing partially nude), or **Safe** (image not containing nude). Also, thanks to its highly efficient processing capability, achieved by its capacity to classify and process multiple images simultaneously in batch mode. These features significantly enhance the explainability of the AI's decision process for human investigation. In an operational context, a multimedia file is processed by the NudeNet model, which classifies the content as either nude, sexy, or safe.

```
{
  input: { //image or video frame
    file: String,
    width: Number,
    height: Number,
  },
  person: Boolean, //has a person been detected?
  sexy: Boolean, //is the person depicted in a sexually explicit way?
  nude: Boolean, //is the person depicted as nude?
  parts: Array<{ //array of detected body parts
    score: Number, //confidence in detection
    id: Number,
    class: String, //label for body part
    box: Number[], //[x, y, width, height]
  }>,
}
```

[8]https://eur-lex.europa.eu/eli/reg/2016/679/oj

[9]https://www.edpb.europa.eu/system/files/2022-07/edpb_edps_jointopinion_202204_csam_en_0.pdf



Figure 2: Example of a positive classification with relative bounding box and scores

If the classification results as 'nude' or 'sexy', the system provides the respective bounding box and a suspicious score for the identified body parts (Figure 2).

Critically, the system forwards the original content augmented with the AI's detailed explanation to a human expert. This mechanism facilitates the application of human context, ethical judgment, and semantic understanding, the role explicitly assigned to the human element in the SAI framework. By focusing the human analyst's limited time and cognitive resources exclusively on the filtered subset of images and guiding their attention to the exact area of concern, the implementation of NudeNet aims to significantly increase the system's sensitivity and specificity in CSAM detection while mitigating the risk of cognitive overload and burnout for the investigators.

7. Conclusions and Future Work

This work proposed a logical Symbiotic Artificial Intelligence framework (SAI4CSAM) for detecting and mitigating Child Sexual Abuse Material within the broader perspective of Cyber Social Security. By integrating multimodal detection engines, human-centered validation mechanisms, cross-platform socio-digital intelligence, and a robust governance and ethical infrastructure, the framework aims to address both the technical and socio-cybernetic complexities that characterize contemporary CSAM ecosystems. The analysis demonstrates that purely automated methods are insufficient to capture the evolving obfuscation strategies, behavioral patterns, and contextual nuances surrounding online abuse. Instead, human-AI collaboration emerges as a necessary paradigm for ensuring resilient, transparent, and ethically aligned detection workflows.

Future research will expand the operational evaluation of the SAI4CSAM architecture through real-world case studies and controlled deployments with the Postal Policies department [37]. Efforts will focus on refining cross-platform behavioral analytics, strengthening privacy-preserving hashing in encrypted environments, and improving explainability mechanisms for high-risk decision-making contexts [38]. Moreover, the integration of advanced co-learning strategies, including reinforcement learning from analyst feedback and adaptive updates driven by emerging socio-digital intelligence, represents a promising direction for enhancing responsiveness to conceptual drift and offender innovation. Ultimately, the framework aims to contribute to a safer digital ecosystem by fostering trustworthy, human-aligned AI systems capable of supporting proactive intervention and long-term societal protection.

Acknowledgments

This work was partially supported by the following projects: SERICS - "Security and Rights In the CyberSpace - SERICS" (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU; Accordo Quadro CrASte - "Cyber Academy for Security and

Intelligence"; SETH: Security Education and Training Hybrid-Wargame, Avviso "Reti - Sostegno alla ricerca collaborativa" - Codice progetto BM755Y1.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] G. Desolda, F. Greco, R. Lanzilotti, C. Tucci, Morpheus: A multidimensional framework for modeling, measuring, and mitigating human factors in cybersecurity, arXiv preprint arXiv:2512.18303 (2025).

[2] V. Antoniol, V. S. Barletta, F. Battista, P. Buono, D. Caivano, G. Campesi, G. Cascione, A. Curci, M. de Gemmis, V. Gattulli, R. Scardigno, A. L. Sciacovelli, P. Sorianello, V. Tamburrano, Security and rights in cyberspace: Cyber social security, CEUR Workshop Proceedings 4086 (2025). URL: https://ceur-ws.org/Vol-4086/paper3.pdf.

[3] V. S. Barletta, D. Caivano, C. Catalano, M. de Gemmis, D. Impedovo, Cyber social security education, in: L. T. De Paolis, P. Arpaia, M. Sacco (Eds.), Extended Reality, Springer Nature Switzerland, Cham, 2024, pp. 240–248.

[4] F. M. Cau, G. Desolda, F. Greco, L. D. Spano, L. Viganò, Can large language models automate phishing warning explanations? a controlled experiment on effectiveness and user perception, arXiv:2507.07916 (2026).

[5] WeProtect Global Alliance, Global threat assessment 2025, 2025. URL: https://www.weprotect.org/wp-content/uploads/GTA-2025_EN.pdf.

[6] World Health Organization, Child maltreatment, ???? URL: https://www.who.int/news-room/fact-sheets/detail/child-maltreatment.

[7] A. Chu, A. Deprince, I. Mauss, Exploring revictimization risk in a community sample of sexual assault survivors, Journal of trauma dissociation: the official journal of the International Society for the Study of Dissociation (ISSD) 15 (2013). doi:10.1080/15299732.2013.853723.

[8] M. T. Baldassarre, V. S. Barletta, V. Bavaro, D. Caivano, A. P. De Matteis, A. Lippolis, A. Piccinno, Llms to detect cyber child abuse in the in textual conversations, volume 3978, 2025. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105008757382&partnerID=40&md5=91bfb48c91b5043174e33d19f2ed45dd.

[9] A. Petrini, M. Belén Gauna, Child sexual abuse material facilitated by artificial intelligence, EthAIca 3 (2024) 146. URL: https://ai.ageditor.ar/index.php/ai/article/view/146. doi:10.56294/ai2024146.

[10] H. Wolbers, T. Cubitt, M. Cahill, Artificial intelligence and child sexual abuse: A rapid evidence assessment, 2025. URL: https://doi.org/10.52922/ti77802. doi:10.52922/ti77802.

[11] S. Singh, V. Nambiar, Role of artificial intelligence in the prevention of online child sexual abuse: A systematic review of literature, Journal of Applied Security Research 19 (2024) 586–627. URL: https://doi.org/10.1080/19361610.2024.2331885. doi:10.1080/19361610.2024.2331885. arXiv:https://doi.org/10.1080/19361610.2024.2331885.

[12] M. T. Baldassarre, D. Caivano, B. Fernández Nieto, A. Ragone, Ethics-driven incentives: Supporting government policies for responsible artificial intelligence innovation, IEEE Intelligent Systems 40 (2025) 55 – 63. doi:10.1109/MIS.2025.3583222.

[13] V. S. Barletta, M. Calvano, A. Curci, C. Nadeem Faisal, A. Piccinno, Assessing usability and cybersecurity of ai systems through the human-centered design, volume 3978, 2025. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105008759499&partnerID=40&md5=509497253bed2368c7b68de7b93d618d.

[14] V. S. Barletta, D. Caivano, G. Dimauro, F. Mantini, M. Morga, Exploring artificial intelligence challenges for monitoring cyber child abuse, volume 3978, 2025. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105008760266&partnerID=40&md5=fed46dfcbf71cc59bd9344aec2c4f01b.

[15] J. Rondeau, D. Deslauriers, T. Howard III, M. Alvarez, A Deep Learning Framework for Finding Illicit Images/Videos of Children, Technical Report NCJ 307535, U.S. Department of Justice, Office of Justice Programs, 2022. URL: https://ojp.gov/library/publications/deep-learning-framework-finding-illicit-imagesvideos-children, deep learning for age estimation and nudity detection to identify sexually exploitative imagery of children.

[16] P. Vitorino, S. Avila, M. Perez, A. Rocha, Leveraging deep neural networks to fight child pornography in the age of social media, J. Vis. Comun. Image Represent. 50 (2018) 303–313. URL: https://doi.org/10.1016/j.jvcir.2017.12.005. doi:10.1016/j.jvcir.2017.12.005.

[17] A. Gangwar, V. González-Castro, E. Alegre, E. Fidalgo, Attm-cnn: Attention and metric learning based cnn for pornography, age and child sexual abuse (csa) detection in images, Neurocomputing 445 (2021) 81–104. URL: https://www.sciencedirect.com/science/article/pii/S092523122100312X. doi:https://doi.org/10.1016/j.neucom.2021.02.056.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.

[19] W. Oronowicz-Jaśkowiak, T. Kozłowski, M. Polańska, J. Wojciechowski, P. Wasilewski, D. Ślęzak, M. Kowaluk, Using expert-reviewed csam to train cnns and its anthropological analysis, Journal of Forensic and Legal Medicine 101 (2024) 102619. URL: https://www.sciencedirect.com/science/article/pii/S1752928X23001373. doi:https://doi.org/10.1016/j.jflm.2023.102619.

[20] M. Pereira, R. Dodhia, H. Anderson, R. Brown, Metadata-based detection of child sexual abuse material, IEEE Transactions on Dependable and Secure Computing 21 (2024) 3153–3164. doi:10.1109/TDSC.2023.3324275.

[21] H.-E. Lee, T. Ermakova, V. Ververis, B. Fabian, Detecting child sexual abuse material: A comprehensive survey, Forensic Science International: Digital Investigation 34 (2020) 301022. URL: https://www.sciencedirect.com/science/article/pii/S2666281720301554. doi:https://doi.org/10.1016/j.fsidi.2020.301022.

[22] B. Westlake, E. Guerra, Using file and folder naming and structuring to improve automated detection of child sexual abuse images on the dark web, Forensic Science International: Digital Investigation 47 (2023) 301620. URL: https://www.sciencedirect.com/science/article/pii/S2666281723001324. doi:https://doi.org/10.1016/j.fsidi.2023.301620.

[23] V. M. Ngo, R. Gajula, C. Thorpe, S. Mckeever, Discovering child sexual abuse material creators' behaviors and preferences on the dark web, Child Abuse Neglect 147 (2024) 106558. URL: https://www.sciencedirect.com/science/article/pii/S014521342300546X. doi:https://doi.org/10.1016/j.chiabu.2023.106558.

[24] M. W. Al-Nabki, E. Fidalgo, E. Alegre, R. Alaiz-Rodriguez, Short text classification approach to identify child sexual exploitation material, Scientific Reports 13 (2023) 16108. URL: https://doi.org/10.1038/s41598-023-42902-8. doi:10.1038/s41598-023-42902-8.

[25] M. Irshad, B. Law, K. Loo, S. Haider, Imgcat: An approach to dismantle the anonymity of a source camera using correlative features and an integrated 1d convolutional neural network, Array 18 (2023) 100279. doi:10.1016/j.array.2023.100279.

[26] J. Madden, M. Bhavsar, L. Dorje, X. Li, Robustness of practical perceptual hashing algorithms to hash-evasion and hash-inversion attacks, 2024. URL: https://arxiv.org/abs/2406.00918. arXiv:2406.00918.

[27] M. Steinebach, Robustness and collision-resistance of photodna, 2024. URL: https://publica.fraunhofer.de/handle/publica/466041. doi:10.13052/jcsm2245-1439.1339.

[28] A. Kulshrestha, J. Mayer, Identifying harmful media in End-to-End encrypted communication: Efficient private membership computation, in: 30th USENIX Security Symposium (USENIX Security 21), USENIX Association, 2021, pp. 893–910. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/kulshrestha.

[29] J. Garcia, A fragment hashing approach for scalable and cloud-aware network file detection, in: 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 2018, pp. 1–5. doi:10.1109/NTMS.2018.8328746.

[30] Y.-Y. Teing, A. Dehghantanha, K.-K. R. Choo, L. T. Yang, Forensic investigation of p2p cloud storage services and backbone for iot networks: Bittorrent sync as a case study, Computers Electrical Engineering 58 (2017) 350–363. URL: https://www.sciencedirect.com/science/article/pii/S0045790616302208. doi:https://doi.org/10.1016/j.compeleceng.2016.08.020.

[31] B. Mahmud, G. Hong, B. Fong, A study of human–ai symbiosis for creative work: Recent developments and future directions in deep learning 20 (2023). URL: https://doi.org/10.1145/3542698. doi:10.1145/3542698.

[32] R. Hao, D. Liu, L. Hu, Enhancing human capabilities through symbiotic artificial intelligence with shared sensory experiences, 2023. URL: https://arxiv.org/abs/2305.19278. arXiv:2305.19278.

[33] M. Calvano, A. Curci, G. Desolda, A. Esposito, R. Lanzilotti, A. Piccinno, Building symbiotic artificial intelligence: Reviewing the ai act for a human-centred, principle-based framework, Minds and Machines 36 (2025) 1. URL: https://doi.org/10.1007/s11023-025-09753-w. doi:10.1007/s11023-025-09753-w, published online 2025/11/18.

[34] B. Zohuri, F. Mossavar-Rahmani, The symbiotic evolution: Artificial intelligence (ai) enhancing human intelligence (hi) an innovative technology collaboration and synergy, Journal of Material Sciences and Applied Engineering 3 (2024).

[35] G. Jocher, Yolov8: Yolo series with object detection and segmentation, GitHub repository (2023). URL: https://github.com/ultralytics/ultralytics.

[36] B. Praneeth, NudeNet: A Deep Learning Model for Nude Content Detection and Classification, https://github.com/notAI-tech/NudeNet, 2023. Accessed: 2025-12-09.

[37] P. Ardimento, M. T. Baldassarre, D. Caivano, G. Visaggio, Multiview framework for goal oriented measurement plan design, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3009 (2004) 159 – 173. doi:10.1007/978-3-540-24659-6_12.

[38] T. Baldassarre, N. Boffoli, D. Caivano, G. Visaggio, Managing software process improvement (spi) through statistical process control (spc), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3009 (2004) 30 – 46. doi:10.1007/978-3-540-24659-6_3.