future internet

MDPI

Article

Leveraging Blockchain for Ethical AI: Mitigating Digital Threats and Strengthening Societal Resilience

Chibuzor Udokwu [1,*], Roxana Voicu-Dorobanțu [2], Abiodun Afolayan Ogunyemi [3], Alex Norta [3,4], Nata Sturua [1] and Stefan Craß [1]

[1] Austrian Blockchain Center, 1020 Vienna, Austria; nata.sturua@abc-research.at (N.S.); stefan.crass@abc-research.at (S.C.)
[2] Faculty of International Business and Economics, Bucharest University of Economic Studies, 010374 Bucharest, Romania; roxana.voicu@rei.ase.ro
[3] School of Digital Technologies, Tallinn University, 10120 Tallinn, Estonia; abnogn@tlu.ee (A.A.O.); alex.norta.phd@ieee.org (A.N.)
[4] Department of Informatics, University of Pretoria, Pretoria 0002, South Africa
* Correspondence: chibuzor.udokwu@abc-research.at

Abstract

This position paper proposes a conceptual framework (CF-BIAI-SXT) for integrating blockchain with AI to enhance ethical governance, transparency, and privacy in high-risk AI applications that ensure societal resilience through the mitigation of sexual exploitation. Sextortion is a growing form of digital sexual exploitation, and the role of AI in its mitigation and the ethical issues that arise provide a good case for this paper. Through a combination of systematic and narrative literature reviews, the paper first explores the ethical shortcomings of existing AI systems in sextortion prevention and assesses the capacity of blockchain operations to mitigate these limitations. It then develops CF-BIAI-SXT, a framework operationalized through BPMN-modeled components and structured into a three-layer implementation strategy composed of technical enablement, governance alignment, and continuous oversight. The framework is then situated within real-world regulatory constraints, including GDPR and the EU AI Act. This position paper concludes that a resilient society needs ethical, privacy-first, and socially resilient digital infrastructures, and integrating two core technologies, such as AI and blockchain, creates a viable pathway towards this desideratum. Mitigating high-risk environments, such as sextortion, may be a fundamental first step in this pathway, with the potential expansion to other forms of online threats.

Keywords: AI ethics; sextortion mitigation; societal resilience; AI for social good; blockchain governance; digital sexual exploitation; ethical AI frameworks; technological neutrality; decentralized systems; privacy-preserving AI

Academic Editors: Klitos Christodoulou and Massimiliano Garda

Received: 19 May 2025
Revised: 11 July 2025
Accepted: 15 July 2025
Published: 17 July 2025

Citation: Udokwu, C.; Voicu-Dorobanțu, R.; Ogunyemi, A.A.; Norta, A.; Sturua, N.; Craß, S. Leveraging Blockchain for Ethical AI: Mitigating Digital Threats and Strengthening Societal Resilience. *Future Internet* **2025**, *17*, 309. https://doi.org/10.3390/fi17070309

1. Introduction

In a society where individuals share a large part of their day online through various activities, the growing digital threats to individuals are exacerbated at a societal level. Add to this volatile environment the potential of ethical challenges brought on by the increasing use of artificial intelligence (AI) in socio-technical systems, and we are today in a position where the individual, particularly the young individual, is in a highly vulnerable position. A growing risk comes from the threat of sextortion, a coercive form of digital sexual exploitation with disproportionately severe consequences to young persons of any

gender and mainly minors. Briefly, sextortion [1] is defined as the act of threatening to disseminate explicit, intimate, or embarrassing images of a sexual nature without the consent of the victim, usually to obtain more images, sexual favors, money, or other forms of compliance. The scale of this issue is more than alarming at the global level, with, for instance, more than 10,000 confirmed victims in a year in a single country (South Korea [2]), although integrative statistics are often lacking due to unreported crimes. A fuller account of the scope and societal impact of sextortion is presented in Section 3.1.2.

The present article is conceived as a position paper. Therefore, it is not the goal of the approach to validate experimentally or at the implementation level at this point. Instead, its principal contribution is the conceptual articulation of a system architecture—CF-BIAI-SXT—which deploys federated learning, blockchain, and ethical AI principles to tackle sextortion risks. By leveraging the regulatory context, the literature, and a conceptual case study, we intend to provide a principled basis on which technical deployment and empirical validation can be built.

This phenomenon of digital sexual exploitation raises critical concerns about human dignity but goes beyond the individual level to undermine institutional trust and, ultimately, societal resilience. In this context, we refer to societal resilience as the collective ability of digital societies to absorb and adapt to technological disruptions, such as AI, without undermining ethical values, human rights, or trust in public institutions. In the online world of tomorrow, which in an ideal case is characterized by decentralization, transparency, and societal trust, technological innovation must mitigate such pressing issues as sextortion [3], in which perpetrators take advantage of sensitive private digital content to blackmail victims.

AI systems have a rather paradoxical relationship to sextortion. On the one hand, they provide tools, techniques, and methods to counter sexual exploitation in online environments where they occur, with AI systems being able to be trained with datasets of victims of sexual exploitation to identify and prevent incidents. Advanced AI techniques such as large language models (LLMs) have also been used in natural language chatbots and robots to collect data on such incidents and detect them early in online interactions [3]. On the other hand, AI has enabled sexual exploitation through image manipulations, resulting in blackmail and extortion of victims [4]. With AI systems being perceived as black boxes by users, significant concern is also raised regarding their trustworthiness and transparency [5], as well as their potential for reinforcing existing inequalities or privacy violations. This dual-use character of AI, with both positive transformational potential and significant risks, reflects a broader principle of technological neutrality—the ethics of any technological deployment depend not on the technology itself but on its governance, design, and implementation. The rapid emergence of new technologies, used indiscriminately by individuals and companies due to their obvious advantages, to the point where they become ubiquitous in everyday life, opens the door to a set of risks that we cannot fully comprehend, assess, or mitigate. In this line, going beyond technical performance, the ethical deployment of AI in socio-technical systems requires alignment with values such as fairness, privacy, accountability, and transparency. AI for social good is, thus, not merely a design ambition but a governance imperative, with the ethical infrastructure having to ensure that interventions do not unintentionally replicate harm or reinforce existing inequalities.

The increased reliance on AI in domains involving vulnerable populations makes ethical governance not just preferable but essential. Given the aforementioned paradox, the governance issue becomes imperative and hinges on a core dilemma on how AI systems can be designed to maximize their capacity to protect while minimizing their potential for abuse in such sensitive, high-stakes applications, such as in the prevention

of sexual exploitation. Blockchain technology, with its core capacities of immutability, decentralization, and auditability, is a possible proposition in this respect, not without its challenges. Although both technologies have been studied separately in the context of digital abuse, their integration specifically for enhancing the ethical governance of AI in sextortion mitigation remains insufficiently examined, as will be shown in the first stage of the research. The intersection of these topics is not yet fully addressed, and, when addressed, current frameworks fail to provide comprehensive guidelines for leveraging these complementary technologies to enhance societal resilience against digital sexual exploitation. Thus, notably, sextortion becomes a paradigmatic case due to its multi-layered risks and trans-jurisdictional legal gaps, in which AI and blockchain may interact with societal vulnerabilities and allow for proper testing for ethical governance frameworks.

Consequently, the goal of this position paper is to address this gap and propose a conceptual framework for blockchain-integrated AI (BIAI) operations that mitigate ethical issues in AI-based anti-sextortion tools (CF-BIAI-SXT). For this, the research question we focus on in this position paper is "*How can blockchain technology's core capacities be integrated with AI systems to address ethical challenges in sextortion detection to strengthen societal resilience through enhanced governance, transparency, and privacy mechanisms?*" The framework leverages blockchain's decentralization, immutability, and auditability to enhance AI transparency and ethical governance while preserving privacy, essential qualities in sensitive applications like sextortion mitigation. To properly address this question and develop the framework, we first propose the methodology of this research and its preliminaries in Section 2—*Methodological Approach and Conceptual Foundations*. Section 3 provides the initial results from the literature review carried out, outlining AI ethical issues and the role of blockchain in addressing them. Section 4 proposes a privacy-first framework for integrating blockchain into AI systems and relevant discussions that emerge. The conclusion of this work is presented in Section 5, including the limitations and future work.

2. Methodological Approach and Conceptual Foundations of This Position Paper

2.1. Scope and Purpose of the Position Paper

This paper aims to present a conceptual framework rather than empirical research or technical implementation. The concepts at hand bring together risk management and societal resilience, multi-jurisdictional public policy, and two highly intricate technologies in an extremely complex intersection. Due to this situation, the methodological approach was a conundrum. There were various possible paths to address the issues (for instance, the practical, empirical way of proposing a Proof of Concept (PoC), analyzing a simulation, or presenting a case study, or the more principle-driven but still practical way of designing a system), but these approaches are limited by access to data, which is highly sensitive and originated from high-risk digital environments.

A brief clarification is also needed on the chosen format of this paper. In this context, we refer to the "position paper" as a conceptually grounded, literature-informed work that articulates a normative opinion. It does so by proposing a structured analytical framework, rather than a presentation of empirical findings or technical implementations. Given all these constraints, the method we have chosen is the literature-supported position paper leading to the proposal of a conceptual framework, articulating a position rather than testing a hypothesis. The proposed privacy-first, societal resilience-oriented framework is founded on a narrative integrative review that synthesizes evidence across diverse study designs, enabling researchers to address complex, multi-faceted questions that connect theory, values, and varied methodologies. Position papers that rely on narrative literature reviews [6] and conceptual frameworks offer a flexible, broad synthesis of diverse

evidence, having the ability to provide both interpretation and critique and inform policy and practice [7]. This solution is nonetheless limited, as it may lack a systematic approach of other review types and may have the potential for bias in the selection and interpretation of literature. This particular limitation is mitigated by adding a systematic literature review in the three-phase methodological approach described below and visually presented in Appendix A—Figure A1.

2.2. Methodological Approach

We adopt a mix of methodologies by balancing an integrative review with a systematic literature review (SLR) on the topic, coupled with a conceptual model of a mixed technological framework to achieve the goal of answering the research question described above in the format of a position paper. The former helps to identify important ethical issues that affect AI systems related to sexual exploitation prevention, and the latter is adopted for a systematic description of some useful blockchain-integrated AI operations that address AI ethical issues. This approach has been adopted in several information system-related research works [8–10] and is considered appropriate for multi-layered socio-technical challenges involving technology ethics and digital governance. This combination of methodologies aligns with a pragmatic epistemological opinion, which allows for the evidence mapping (via SLR, as will be shown) to be complemented by an interpretive, theory-driven synthesis (via narrative review). This approach is particularly suitable for addressing complex, interdisciplinary, and under-theorized phenomena.

The research follows a three-phase approach with (1) a review of the literature at the intersection of AI ethics, digital abuse, and frameworks for AI for social good and societal resilience; (2) an assessment of blockchain's suitability as an ethical infrastructure complementary to AI, and (3) a proposed framework integrating both to strengthen digital trust in contexts of online sexual exploitation. A full flow chart of the methodology going from the research question to the literature review and then to the framing of CF-BIAI-SXT is presented in Appendix A—Figure A1. Furthermore, due to the interdisciplinary nature of our paper, some clarification of the technical elements (such as AI and blockchain-related concepts) is needed and provided in the form of Preliminaries in Appendix B.

2.2.1. Methodology for the Evidence Synthesis: Systematic and Narrative Review

The first part (1) builds on separate elements, as the literature on AI for Social Good and societal resilience is vast, and the intersections between the concepts are numerous and highly complex. Therefore, we split the review into two separate stages: the first is a systematic literature review of the concepts in the research question, and the second is a contextualization of the concepts based on a narrative exploration of the aforementioned topics in three steps.

The first stage of the review provides a systematic literature review (SLR) to start the analysis and identify underexplored intersections. The SLR part aims to identify ethical concerns, governance gaps, and emerging responses that align with frameworks such as AI for Social Good and societal resilience and to refine the conceptual foundations for the framework proposed in later sections. We used the Rayyan platform (Rayyan Systematic Literature Review Tool: https://www.rayyan.ai/, access on 14 July 2025) to conduct the SLR of a series of articles from *Scopus*, *Web of Science*, *IEEE Explore*, *and ACM Digital Library.* We considered including ArXiv in the database search; however, it is mainly a pre-print repository, and we aimed to include peer-reviewed articles and chapters. Based on the initial search on Scopus for the word *sextortion*, we identified the starting point for the period of publication of the articles collected as 2017–early 2025. Prior to 2017, mentions were scarce and mostly lacked relevance to the intersection of AI ethics or blockchain infrastructure. As

the analysis yields a limited number of articles for the AI–blockchain–sextortion intersection, proving the hypothesis that there is a gap in research, a supplementary SLR was deployed by expanding the search results by using the OR argument instead of AND. Moreover, for the first batch of articles, a formal quality scoring rubric was deemed less meaningful to establish thematic boundaries and identify research gaps. For the second batch used in the supplementary SLR (which adheres to PRISMA guidelines), building towards establishing the complex landscape of the topic, we used the same criteria for inclusion and focused on the period chosen. The articles were screened by each author of this position paper independently. The Boolean Logic and the inclusion/exclusion criteria are also detailed in Appendix A.

The second stage of the review is the narrative exploration, which details the concepts in the systematic review to build on the conceptual landscape. If the SLR established the rarity and limitations of existing literature at the intersection of AI, blockchain, and sextortion, this section is meant to extend these findings across ethical and governance dimensions. Together, both these methods form the empirical and theoretical foundation of the proposed CF-BIAI-SXT framework presented in Section 4. This exploration is an interpretive, theory-informed process that identifies, explains, and interrelates key concepts from literature to construct a conceptual model. This approach (underlined as a valid alternative in [11]) makes sense for complex issues, multi-perspective interdisciplinary spaces, like the intersection of AI, blockchain, and sextortion—areas where there is still a lack of clarity, definitions remain fluid, conceptual frameworks are still taking shape, and the evidence base remains too limited for broad empirical generalizations. The narrative exploration conceptualizes sextortion mitigation as embedded in a multilayered ethical ecosystem that includes AI for social good, blockchain-enabled governance, societal resilience, and supporting regulatory and equity frameworks. Thus, the process not only identifies relevant themes but also explains their interconnections within the ethical landscape of AI-driven digital abuse prevention. In this stage, the three steps of the narrative conceptual exploration were completed. The literature selection for this stage used thematic relevance criteria, as per Figure 1, by choosing recent studies and a purposive sampling approach (only literature that specifically addresses each component of their conceptual ecosystem). It uses three steps:

1. First, to identify digital technologies that provide the environment for sexual exploitation leading to sextortion.
2. Second, to identify the dual-use role of AI in preventing and countering sexual exploitation on digital platforms and the potential ethical issues in automated AI systems.
3. Third, to clarify the emergence of blockchain operations for societal resilience as anchors for ethical AI development.

In this respect, we conducted an iterative snowball search (2024–2025) starting from the SLR seeds and expanding through forward and backward citation chaining in Google Scholar and institutional repositories, and also by adding reference points to the landscape of concepts underlined through the conceptual mapping of the SLR seeds. We mapped this corpus with a coding sheet replicating that of the SLR (see more in Appendix A, Tables A1–A3). The coding protocol included the identification of peer-reviewed sources as more relevant than non-peer-reviewed sources. Each article is validated as to be included in the corpus by the authors, based on the steps mentioned above.

Following the narrative exploration, we position (see Figure 2) the conceptualization of the sextortion mitigation use case in the broader framework of AI for social good, building on the societal impact of the technology and of the threat, on the concept of Good AI and its ethical challenges, on the frameworks (including guidelines and regulations) existing

in the literature and on the potential for transformation and further participation. The figure synthesizes the thematic logic of the narrative exploration by mapping the sextortion use case within the ethical AI ecosystem and serves as a bridge to the construction of CF-BIAI-SXT.

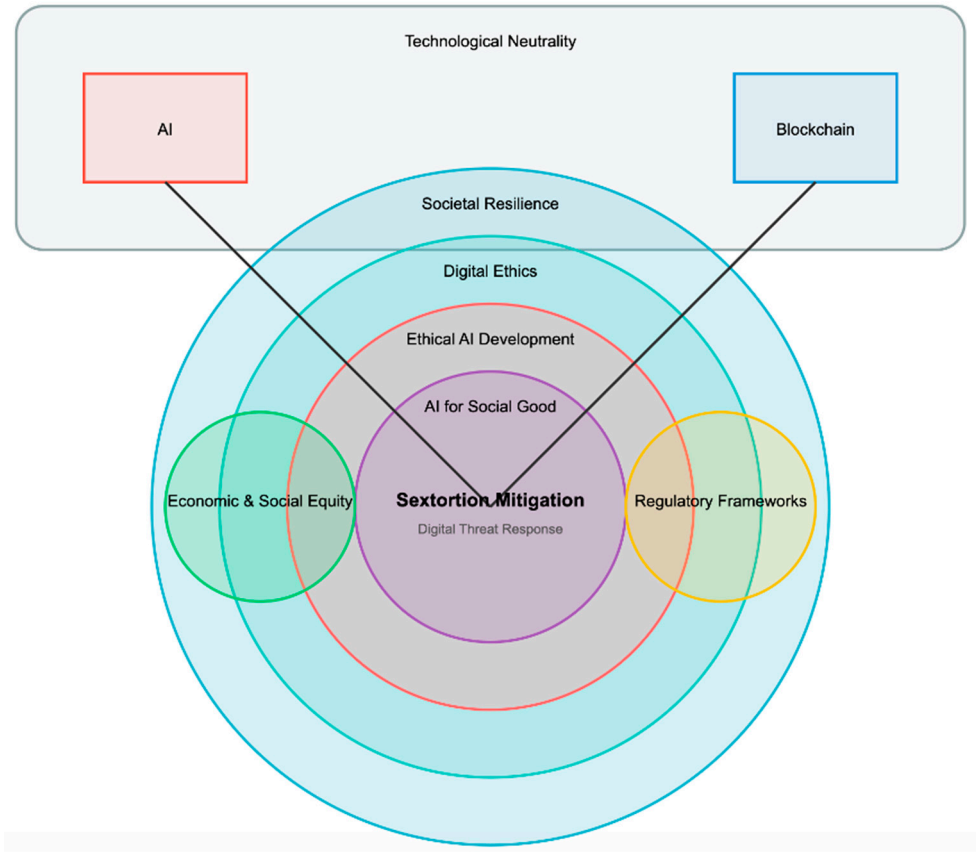

Figure 1. Conceptual map of the ethical ecosystem for AI-based sextortion mitigation.

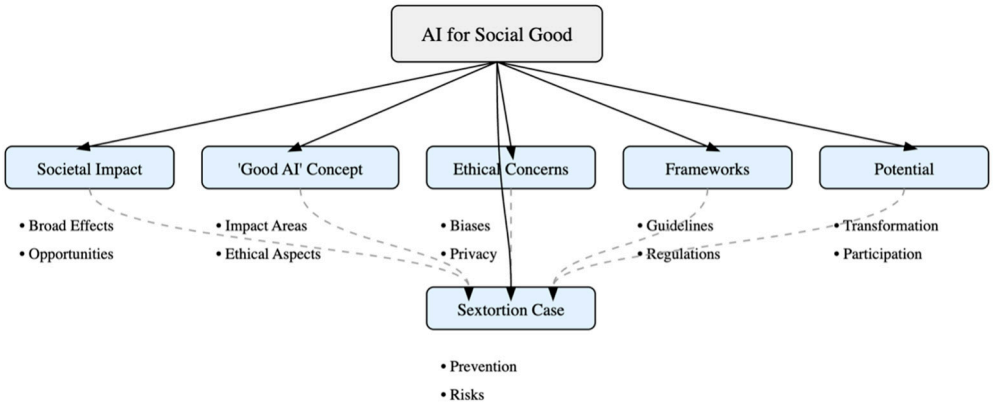

Figure 2. Conceptualization of the use case—Thematic mapping of AI for Social Good applied to the sextortion mitigation use case.

The narrative exploration uses four instruments to cover the complexity of the topics presented: conceptual mapping framework (as exhibited in Figures 1 and 2), multi-layered analysis (as exhibited in Figure 2), tabular comparison approach, and thematic synthesis. A structured summary of the findings from this narrative review—including digital technologies enabling sextortion, AI's roles in both enabling and mitigating such abuses, and corresponding blockchain capacities—is included in Appendix C.

2.2.2. Methodology for the Conceptual System Modeling: BPMN Framework Logic

The second part of this research design comprises conceptual BPMN (Business Process Model and Notation) models representing a set of blockchain-based operations integrated into AI processes for addressing AI ethical concerns. BPMN model representations are a well-established approach for conceptual modeling of software systems [12,13]. In this paper, BPMN elements are used to represent tasks and decision points in AI and blockchain processes. The goal is to realize models that, when implemented, will result in ethically aware AI applications that prevent sexual exploitation leading to sextortion. We adopt BPMN to the detriment of alternatives such as UML or Petri nets due to its suitability for modeling complex, multi-actor workflows involving both technical components (AI, blockchain) and institutional processes (e.g., permission checks and data governance). It offers an intuitive structure, accessible across technical and non-technical domains, to represent decision points, parallel actions, and traceable logic, critical for ethically aligned, auditable system design.

3. Initial Results

The results of this paper are structured in three phases, leading to the proposal of the conceptual framework:

1. A review of the literature at the intersection of AI ethics, digital abuse, and frameworks for AI for social good and societal resilience;
2. An assessment of blockchain's suitability as an ethical infrastructure complementary to AI
3. A BPMN-based conceptual model of blockchain operations that addresses AI ethical issues.

3.1. An Empirical and Conceptual Review of AI Ethics, Digital Abuse, and Frameworks for AI for Social Good and Societal Resilience

3.1.1. A Systematic Literature Review (SLR) of the Analyzed Concepts

To fully deploy the SLR part of the exploratory investigation, we searched Scopus, Web of Science, IEEE Explore, and ACM Digital Library using keywords spanning AI, blockchain, and sextortion (2017–2025) and identified 45 candidate publications. We then applied PRISMA 2020 guidelines to our screening process, as illustrated in Figure A2 in Appendix A, which shows the flow of information through identification, screening, and inclusion. As described in the methodology, to enhance the transparency and reproducibility of our literature review and ensure unbiased selection, we used a rigorous, multi-reviewer screening approach. Each author of this paper independently screened the records using the Rayyan platform, a web tool for systematic review management. We included only studies that met all criteria (addressing AI and blockchain in the context of digital sexual exploitation with an ethics focus) and excluded others (see Appendix A for detailed criteria).

To ensure that we keep the SLR as intended, at the intersection of AI, blockchain, ethics, and sextortion, we used Boolean logic in the search terms and presented them in Appendix A. Thus, in our systematic review, we included only sources that: (1) integrate AI and blockchain (technologies relevant to our research question); (2) focus on digital sexual exploitation (e.g., sextortion contexts); and (3) address ethical considerations (such as fairness, transparency, or privacy in AI use). Studies that failed to meet any of these criteria were excluded. For instance, we omitted works with no AI or no blockchain component (16 articles), those unrelated to sextortion or digital abuse (12 articles), and those lacking any AI ethics dimension (11 articles).

By using the defined search for the period 2017–2025, we find the following literature results:

- Scopus identified 6 papers from 2019 onwards, with 1 being a false report—5 articles;
- Web of Science returned 0 results (even without a period range);
- IEEE Explore returned 0 results (even without a period range);
- ACM Digital Library returned 55 results for the period 2017–2025 without defining the search by timeframe; therefore, we can identify two brief conclusions: those are all articles on the defined topic from the database and the chosen timeframe is valid.
- Although we initially excluded Arvix from the search, we nonetheless tried the query on this database as well and received 0 results. Similarly, Semantic Scholar returned 0 results for the same Boolean logic.

The limited number of relevant results may reflect a disciplinary/research dissemination skew caused by the fact that these databases tend to emphasize either highly technical implementations or formal computer science publications, with less coverage of interdisciplinary studies addressing ethics, governance, or social impact. It highlights a lack of holistic handling of macro issues with societal impacts by incentivizing researchers to focus on narrow issues, more suitable for publication. Moreover, the tight Boolean logic using AND operators may have narrowed retrieval too far for broader-indexing platforms. This limitation is to be mitigated by the supplementary SLR, as will be described further in this section.

We screened the 60 articles for duplicates and identified that ACM proposes both the Proceedings and the article in case there is a match, so we retained in the analysis just the article. This leads to retaining in the analysis only 40 references from the ACM Library. We applied the PRISMA methodology of literature review for the found articles (45 in total—5 from Scopus and 40 from ACM, all in English and peer-reviewed. See full list in Appendix B) and proceeded with a screening based on the inclusion/exclusion criteria detailed in Appendix A. Following the screening procedure identified that only 6 of the 45 articles fit the Boolean logic [14–19], an outcome that highlights a significant research gap at the intersection of AI, blockchain, and sextortion, rather than being an inadequacy. This limitation of the study is thus acknowledged explicitly and justifies our dual review approach: a targeted SLR to identify what little work exists, in combination with a broader narrative exploration to capture adjacent insights, as presented in the following Section 3.1.2. The use of a narrative review, as informing a position paper, is recognized in emerging or interdisciplinary fields where empirical studies are few as a suitable means to synthesize and contextualize knowledge, as previously mentioned. Thus, although the quantitative aspect of the SLR is limited, our study remains comprehensive through the integrative strategy proposed in the methodological section presented previously.

The main reasons for the exclusion of the rest of the articles were

- Lack of AI or blockchain relevance (16 articles)
- No focus on digital abuse or sexual exploitation (12 articles)
- Absence of ethical considerations such as fairness, transparency, or data privacy (11 articles)

We then coded the articles in Rayyan based on the following:

- AI function (mitigation, surveillance, moderation, etc.),
- Ethical focus (e.g., bias, explainability, misuse of data),
- Governance model (centralized vs. decentralized),
- Framework alignment (AI4SG, resilience, digital trust).

Considering that our review is exploratory and conceptual in nature, and our aim was to synthesize themes and inform a conceptual framework (rather than to evaluate an

intervention's effectiveness), we did not assign formal "quality scores" to the 6 articles included in the study. All selected works are peer-reviewed and from reputable databases, as per the initial requirement, and were deemed directly relevant to our research question, through the process of inclusion/exclusion by the authors. This inherently sets a baseline of credibility, in line with guidance on integrative reviews, relying more on thematic insight than on formal quality scoring. Moreover, in view of the nature of this study as a position paper, the quantitative risk-of-bias appraisal would add limited value.

However, for transparency, we have documented the qualitative coding protocol used in our review. Each of the 6 included studies was systematically coded along predefined dimensions: the role of AI (e.g., mitigation, detection), the ethical focus (fairness, privacy, etc.), the governance model (centralized vs. decentralized), and alignment with broader frameworks (AI for Social Good, societal resilience). The authors independently applied these codes to each article, then compared and reconciled any differences to ensure reliability, both for the SLR and the narrative review. In Appendix A, we included the tables (Tables A1 and A2) for the coding protocols for each stage. We also included a summary table (Table A3) to how each study in the SLR seed 6 maps to these categories, providing a clear audit trail from literature to topics. This coding scheme, in line with established thematic analysis techniques for building conceptual models, connects to the cluster map below (Figure 3) and ensures that the analytic process is traceable.

Figure 3. Cluster Map of the core SLR with Relationships Among Studies, Functions, Ethics, Governance, and Frameworks.

A cluster map of the 6 articles [14–19] considered for the systematic literature review is presented in Figure 3. However, the limited number of articles and the fact that the criteria for the intersection exclude a significant part of the overall picture of sextortion as a societal resilience threat in the context of AI and blockchain lead to the need for a more encompassing presentation of the concepts at hand. Thus, the methodological approach to include a narrative exploration is sound and offers the ability to cover the issue more comprehensively.

Before presenting the narrative exploration, we chose to test whether the literature gap holds across a broader research landscape. For this, we extended the Boolean logic by using OR operators instead of AND operations to include studies that addressed AI-blockchain integration with explicit emphasis on ethics, privacy, transparency, and sextortion-related contexts. This expansion led to additional papers that offered insights into adjacent topics, hence were reviewed thematically and also included in the narrative exploration in Section 3.1.2. We use Semantic Scholar and Rayyan as the main screening tools. It is to be noted here that this extra systematic review is meant to offer quantitative and thematic insights complementary to the SLR, not as a substitute. We screened 40 articles, and the findings of this supplementary SLR are structured in Figure 4.

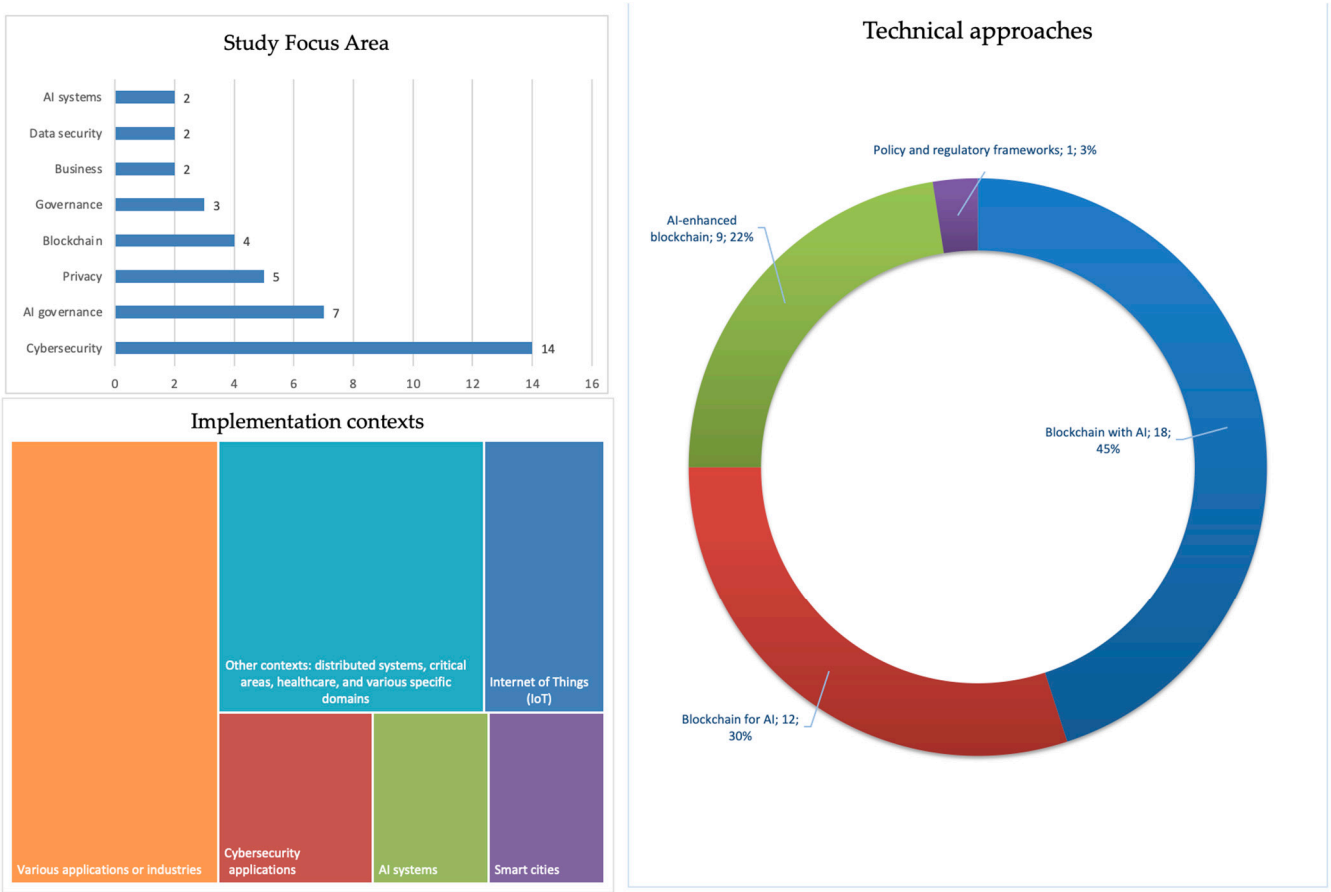

Figure 4. Supplementary SLR findings.

This supplementary SLR confirms the rarity of full-stack frameworks but points to building blocks (e.g., federated learning, explainable AI, smart contracts) that inform the conceptual model developed further on.

On the thematic analysis presented by this supplementary SLR, we identify the following four areas (as shown in Table 1), with blockchain-AI integration being the most developed.

Table 1. Themes in Blockchain, AI and Sextortion Research.

Thematic Area	Subtopic	Key Concepts/Mechanisms	Representative Studies
Blockchain-AI Integration	Privacy-preserving frameworks	Zero-knowledge proof, Homomorphic encryption, etc.	[20–22]
	Decentralized governance	Smart contracts, Identity management, Incentive distribution	[23,24]
	Technical architectures	Hyperledger Fabric, Ethereum, Reinforcement learning	[25–27]
Ethical Frameworks	Governance mechanisms	Balanced approaches, smart contract-based governance	[28,29]
	Transparency protocols	Smart contracts for ethics enforcement	[24,30,31]
	Privacy safeguards	Audit tracking, encryption, and de-identification	[32]
Sextortion Mitigation	Detection, Verification, Victim protocols	AI detection, blockchain verification	[3]
Societal Resilience	Stakeholder coordination	Cross-org data sharing, governance trust systems	[3,29]
	System effectiveness	Infrastructure performance	[27,33]
	Social impact	Equity, public trust, fairness	[29,34]

While several studies propose concrete architectures combining federated learning, secure multiparty computation, and smart contracts, ethical frameworks appear less consistently and are concentrated around privacy, consent, and bias, with fewer addressing explainability or transparency. In governance terms, centralized architectures dominate, although several studies point towards decentralization through smart contracts or identity management. Most articles address AI functionality in contexts of moderation, mitigation, or detection, while only a minority directly engage with sextortion as a phenomenon. Sextortion mitigation per se is rare, but societal resilience is growing in interest, especially related to trust and stakeholder coordination, though largely extrapolated from other domains. Finally, the alignment with frameworks like AI4SG, or digital trust, or societal resilience is often implicit, thus confirming the gap that CF-BIAI-SXT seeks to mitigate. Building on the list of articles analyzed in this stage (excluded and included) and adding articles to connect to the regulatory aspects, as well as AI for Good, AI governance, digital trust frameworks, and the role of blockchain in ethical infrastructures, we create the narrative exploration that explains in detail the concepts linking them to the conceptual framework proposed further on.

3.1.2. A Narrative Conceptual Exploration

The SLR in the previous section maps the limited empirical footprint at the intersection of AI, blockchain, and sextortion. This section, working on those findings, adopts a broader epistemic lens, engaging with adjacent topics to build towards the conceptual framework. This shift allows for the connection of the technical architectures with ethical governance and societal resilience perspectives, critical to the design logic of CF-BIAI-SXT.

To ensure traceability of this narrative review, we used a modified coding sheet, similar to the one used for the SLR, to map every narrative sources considered and referenced. This coding sheet is presented in Table A2 in Appendix A, to allow external researchers to follow and replicate or extend our thematic decisions, although it was not meant to

eliminate subjectivity. The snowball method with added reference points based on the concepts yielded a total number of 157 sources, of which 118 peer-reviewed and 39 non-peer-reviewed (news reports, policy/brief items, arXiv/working papers, or other gray-lit, such as the FBI web brief, OECD framework). These 157 sources include the 6 seed sources from the SLR and are listed in full in the references list. The corpus breakdown is as follows: 27% refer to digital-abuse technology, 19% to AI dual-use and risk, 22% to Ethical AI frameworks, 20% to blockchain capacity, and 12% to societal resilience. Their direct informing of the framework is detailed in the narrative review in this section, mostly in tabular format, with other sources linked indirectly being mentioned in other parts of the article. The corpus also includes survivor-voice studies, such as [1,35] that provide first-person evidence on coercion dynamics and thus, lead to informing the data-wallet consent logic of CF-BIAI-SXT.

The conceptualization of sextortion mitigation as embedded in a multilayered ethical ecosystem that includes AI for social good, blockchain-enabled governance, societal resilience, and supporting regulatory and equity frameworks is achieved in a three-step process.

Firstly, we identify the digital technologies that provide the environment for sexual exploitation leading to sextortion. Sextortion, defined as a form of sexual exploitation in which victims are, e.g., extorted with their sexual images [36], is a form of dissolution of societal resilience. It exploits vulnerabilities in social and economic structures and corrodes the integrity of society by breaking down structures meant to protect privacy and security. The harms of sextortion involve dignity, well-being, safety, and individual impact, touching on privacy, security, and societal consequences. It affects, at the global level, all social strata, although some may be more vulnerable than others.

Statistics on the topic is often missing, however, recent statistics from jurisdictions that collect data on the topic, such as those from the US Federal Bureau of Investigations (FBI) [37], indicate "more than 13,000 reports of online financial sextortion of minors [with] at least 12,600 victims and led to at least 20 suicides" between October 2021 and March 2023. The victims are minors, typically "men between the ages of 14 to 17"; however, the issue is affecting all genders. In a study in 2024 [37], the authors highlight that of close to 17,000 respondents, "14.5% reported victimization and 4.8% reported perpetration", with men, LGBTQ+, and young victims more likely to report victimization than other categories. The same study affirms that "experiencing threats to distribute intimate content is a relatively common event, affecting 1 in 7 adults". A report by the Korean Digital Sex Crime Victim Support Center [2] mentions 10,305 confirmed cases of digital sex crimes, with "teenagers and people in their 20s [accounting] for 78.7 percent of reported victims" with "1384 cases involving image manipulation using AI and advanced technology. This marks a sharp increase from the 423 cases reported in 2023—more than triple the previous year's figure". The landscape is dire and becoming riskier by the day. To add to the risks, this digital threat breeds in fertile environments for manipulation, grooming, and blackmail, such as social media, messaging, and virtual reality platforms [38,39]. These, hence, exacerbate this form of abuse with a significant emotional toll on its victims [40].

Although in most countries, sextortion is often not legally defined as a crime, authorities prosecute related crimes in varying ways between jurisdictions, categorizing it as child pornography, harassment, extortion, stalking, hacking, and violations of personal privacy [24]. For instance, in the USA, sextortion is defined in two primary ways: as a threat to share a victim's private sexual images to extort something from them or as coercion to make the victim send sexual material under threats. Federal law typically prosecutes sextortion as extortion or child pornography, depending on the victim's age [35]. The member states of the European Union also prosecute sextortion in a variety of ways

according to domestic legislation, often through extortion, privacy violation, and sexual harassment charges that account for individual and broad social impacts, from defamation and integrity infringements to media manipulation and gender inequalities [41]. For example, in France, illicit coercion of sexual favors faces civil and criminal charges such as sexual assault, extortion, blackmail, or corruption [41]. In addition to this, the Digital Republic Law covers issues related to the unauthorized use of personal data, including non-consensual sexual imagery and deepfakes [42]. Similarly, data protection laws are used to protect against sextortion in Germany, Spain, and Hungary, with legislation in the latter specifically including provisions for sexual exploitation, defined as coercing someone into sexual activities through threats, and, since 2013, sexual blackmail and extortion [41]. While the USA's prosecution approach through existing extortion laws provides flexibility, it fails to address the specific psychological harms of sextortion that specialized legislation like Germany's might better target. This legal fragmentation creates significant challenges for cross-border enforcement and technology-based solutions.

Digital technologies play a role in enabling sextortion by providing environments for sexual exploitation. Hence, it is necessary to understand these digital technologies, the roles they play in sexual exploitation, and the impacts on the victims. We examined the literature articles [38–40,43,44] in Table 2 to identify social media platforms, chatting platforms, dating websites, and virtual reality platforms where sexual exploitation activities are prevalent. The findings [38] show that mobile apps emphasize psychological impacts, but fail to differentiate between platform-specific risks, highlighting a critical gap when designing technological countermeasures. The broader categorization of platforms from [43] offers more nuanced insights into how different technological environments create unique exploitation opportunities, suggesting that a one-size-fits-all mitigation approach is insufficient. These platforms enable criminals to stalk, coerce, threaten, and harass victims into sharing sexually explicit images and videos. Social media platforms also provide grooming channels to lure victims into various romance scams. These explicit images are then stored on cloud services, web hosting platforms, and other file hosting services to extort the victims, thereby resulting in sextortion cases. Extortion of victims has several health implications, such as psychological, emotional, and financial, which can lead to depression, self-harm, and, in the worst cases, result in suicide.

In this context, a case study on sextortion as a form of cyber abuse and growing societal concern, particularly in the case of young people and/or minors, can be a niche illustration of the ethical challenges of AI. Efforts to combat this cybercrime may benefit from the use of AI [3,45], while some countries (Indonesia, for example) are setting regulatory frameworks in place to deal with sexual violence crimes, including sextortion (the TPKS law of 2022 [46]). On the opposite side of the spectrum, AI can be shown to leverage the sextortion efforts of perpetrators through dating apps [47] or deepfakes [4]. Despite timid advances in the topic, both in the literature and in regulatory and legal frameworks, the significant gap in empirical studies demonstrating the effectiveness of technological solutions (AI or other) in preventing and mitigating sextortion has yet to be addressed. This empirical gap stems from both methodological challenges in studying sensitive populations and the rapid evolution of digital technologies that outpace traditional research timelines. Furthermore, existing studies often focus on technical detection without adequately addressing the ethical dimensions of AI deployment in these contexts.

The manner in which these identified technological enablers of sextortion highlight how centralized data storage and insufficient user control over intimate content create exploitation vulnerabilities directly informs our framework's focus on decentralized control and enhanced privacy mechanisms.

Table 2. Digital technologies' roles and impact in enabling sextortion.

Digital Technologies	Roles in Enabling Sextortion	Impact	Referenced Studies
Mobile apps, virtual reality platforms, and social media.	Non-consensual taking, sharing, or threats to share personal, intimate images or videos.	Psychological and emotional, social, financial, and behavioral impacts, and physical harm	[38]
Social media platforms, messaging apps, online dating sites, camera and video-enabled devices, email, and online communication channels	Grooming, harassment, and non-consensual sharing of intimate images, cyberstalking, romance scams, revenge porn and sextortion, coercive messages	Violence, digital harassment, image-based abuse, sexual aggression and/or coercion, and gender-/sexuality-based harassment	[43]
Social networking, online hosting services, and advanced encryption techniques	Recruit victims, advertise services, store and share illicit content, protecting their communications and data from detection	Mental health risks, psychological terrorism	[40]
Social media, messaging apps, GPS tracking apps, online video-sharing platforms, cloud storage, and digital media sharing	Coercion, harassment, and dissemination of sexually explicit content, coerced sexting and sextortion, cyberstalking and monitoring of victims, recording and distribution of sexual assaults, and storing and disseminating explicit content.	Sexual violence and exploitation	[44]
Internet, online platforms, mobile phones, messaging apps, and live-streaming technology	Advertise victims and connect with potential clients, communicate covertly, coordinate logistics, broadcast exploitative content, and non-consensual explicit content	Fear, anxiety, anger, humiliation, shame, self-blame, suicidal ideation and suicide, depression, financial scam	[39]
Educational platforms, poor digital literacy infrastructure	Lack of early cybersafety education enables grooming and manipulation	Psychological and emotional harm, especially to minors	[48]
IoT devices and infrastructure	Potential for remote access to cameras, microphones, and private data, enabling covert monitoring and exploitation	Surveillance-driven coercion, data exposure, and non-consensual recordings	[49]
Socio-technical platforms (architecture-level vulnerabilities rather than just app-level usage)	Hosting sensitive user data that can be exploited, potential detection gaps	Privacy violations, data exploitation	[50]

Secondly, we identify the dual-use roles of AI in preventing and countering sexual exploitation on digital platforms and the potential ethical issues in automated AI systems. AI plays a role both in enabling and mitigating sexual exploitations that lead to sextortion cases. Even when AI is applied to help prevent sexual exploitation through various automated algorithms, there are also ethical concerns about the use and misuse of personal data. In Table 3, we analyzed the following related articles [3,4,14,17–19,49,51–70] to identify the roles of AI in enabling and mitigating sexual exploitation and collect the list of ethical concerns that are raised when AI is applied to address sexual exploitation problems. The role of AI in enabling sexual exploitation mostly involves the alteration of images and video

recordings where victims' faces or voices are used to replace original actors in sexually explicit content. These forms of alterations are commonly referred to as deepfakes. These digitally altered images and videos are then used to blackmail and extort victims on social media and chatting platforms, thereby resulting in sextortion. Other examples of AI usage in sexual exploitation include the sexualization of female AI agents.

Table 3. AI's roles in preventing sextortion.

Thematic Cluster	AI's Roles in Preventing Sextortion	Ethical Issues	Referenced Studies
Content Moderation and Identification of Exploitative Material	Content moderation on social media platforms; Image analyses for identification of sexual exploitation materials; Identification of sexualized content in public chatrooms; Deepfake images for blackmailing; Image-based sexual abuse through altered images and recordings	Bias, accuracy, freedom of expression, and privacy	[4,18,51–55]
Detection and Forensics	Detecting suspicious-transaction patterns related to sexual exploitation; Elimination of sexual exploitation materials; Data warehousing on sexual exploitation incidents; Semantic analysis for AI interpretation	Data retention, surveillance, misclassification, and trust	[49,56–60]
Reporting, Access, and Empowerment Tools	Natural-language chatbots for reporting incidents; Personalized education and victim support; Gamified behavioral interventions	Accessibility, agency, consent, and data sensitivity	[3,17,19,61]
AI-Blockchain Integration for Ethical Governance	Secure AI architecture with blockchain; Blockchain-enhanced explainable AI; Self-sovereign identity systems; Governance-by-design models	Transparency, decentralization, traceability, and explainability	[62–69]
Risks of AI Exploitation	Sexualization of female AI agents; Surveillance-enabled abuse; Generative AI for coercion	Discrimination, objectification, and unintended harm	[14,51,70]

According to the findings of the research in Table 3, AI can be used in many ways to limit, combat, and prevent sexual exploitation on digital platforms. The same Table 3 also highlights some ethical issues that arise, such as privacy issues due to data misuse and sensitivity, the accuracy of detection logic due to false positives, bias due to incomplete training datasets, security, and data confidentiality. Other ethical issues include transparency, trustability, and accountability of AI systems. There is also the problem of accessibility of AI tools and legal issues related to regulatory jurisdiction and different interpretations of sexual exploitation, issues that we explain in detail in the following paragraphs of this narrative exploration.

Thirdly, we clarify the emergence of AI for social good and societal resilience as anchors for ethical AI development.

AI for Social Good: With its ability to both help and harm, AI has immense promise in producing multidimensional impacts. We should properly harness its abilities, especially regarding social issues. This potential to address societal challenges is proven by solutions such as conversational AI tools being deployed for issues such as mental health awareness. For example, Ref. [71] shows how an AI-based emotional chatbot can be used to detect mental issues such as depression by analyzing facial expressions and textual content

produced by users, while studies [72,73] describe AI language models to identify depression and suicide prevention. However, there is a gap in understanding how AI may support the mitigation of long-term psychological effects and how it may be integrated into mental health frameworks. Often proposed individually, solutions discussed fail to consider how each component might work synergistically as part of a holistic support system, maximizing AI's ability to meaningfully impact individuals and society. A balanced and holistic approach to embedding AI within existing complex support networks can maximize its capacity to benefit individuals and society at large.

Societal Impact and Ethical Complexity: The societal impact of AI goes beyond the micro-level of individuals interacting with technology, and the literature is booming with analyses of various aspects of its potential uses for social good. Before 2021, there were fewer than 450,000 results in Google Scholar on AI and social good; as of March 2025, there were close to 6 million, with the potential for an increase in this number (see also the analysis by [74]). For example, Ref. [75] explores the "potential economic, political, and social costs", while [76] investigates the ethical applications of AI in social systems, paving the way for a nuanced discussion on the topic of AI use in addressing social challenges. Similar issues are addressed in [76], following an assertion that the wide implementation of AI systems goes beyond engineering to an intersection of technology and society, and proposes an illustration of the concept of "ethically designed social computing". From the positive aspects of "accuracy, efficiency, and cost savings" [77], issues such as privacy, trust, accountability, and bias must be considered [78]. Diverse aspects related to education and critical thinking can lead to unequal deployment of such technology and ultimately to greater societal polarization [79], an exacerbation of social inequality [80], and dissolution of societal resilience. Similarly to the previous assertion, the critical analysis of this wide range of insights lacks clarity on the way biases and inequities resulting from AI may be mitigated to enhance societal resilience. The literature so far remains at the status quo level and assessment without dwelling on the next steps of a risk management process: risk interconnection and mitigation or reduction.

Resilient societies and "Good AI": The idea of a 'good' AI society is not new. The work of [81] starts a conversation on ten potential areas of social impact: "crisis response, economic empowerment, educational challenges, environmental challenges, equality and inclusion, health and hunger, information verification and validation, infrastructure management, public and social sector management, security, and justice". In the same line, the work of [82] maps 14 ethical implications for AI in digital technologies: "dignity and well-being, safety, sustainability, intelligibility, accountability, fairness, promotion of prosperity, solidarity, autonomy, privacy, security, regulatory impact, financial and economic impact and individual and societal impact." All these implications and potential impacts are woven into a very complex ontological system of a society existing dually (in real and in digital) in which AI represents a new layer. An interplay of individual, institutional, and community capacities, societal resilience is based on coping, adaptation, and transformation and relies on holistic approaches, inter- and multi-disciplinary frameworks, and normative epistemological questions [83–85]. Although AI is a potential tool to improve individual resilience (particularly given the theory of resource conservation, as discussed by [86]), its role in community resilience is only beginning to be acknowledged. Firstly, societal resilience is promoted through adaptable and flexible systems and structures and innovation spaces [87], with AI-supported operational cyber resilience [88]. Secondly, AI's own resilience must be assessed, and one way to do so is through agent-based modeling [89]. However, more empirical research on how AI can be integrated to validly enhance social resilience without introducing new risks and vulnerabilities is needed.

Ethical Risks and Governance Challenges: In contrast, artificial intelligence (AI) is considered to pose significant risks to humans and societies if it is not ethically developed and used. Researchers, corporations, and NGOs, along with policymakers, explore maximizing AI's benefits and capabilities. Yet, fast progress and implementation of AI solutions may outpace understanding of unintended effects. The challenges posed by AI are diverse [78], ranging from algorithmic biases to the potential for humans to inherit AI errors. Fundamentally, AI must be fair, transparent, explainable, responsible, trustworthy, and reliable. Without these attributes, it remains a 'black box' where developers may themselves struggle to comprehend how the system generates its responses. It is of utmost importance that society (including all stakeholders) takes immediate steps to prevent AI tools from engaging in unpredictable behaviors and establish the credibility and trustworthiness essential to society.

Frameworks for Ethical AI, Guidelines and Ethical Principles, Regulatory and Legal Frameworks: As early as 2007, the authors in [90] emphasized that evaluating technologies in isolation is futile due to their social implications. In this view, a series of multi-layered frameworks have been developed to assess AI's impact potential for societal good. In [91], an analysis of AI solutions was performed using four criteria: breadth and depth of impact, potential implementation of the solution, risks of the solutions ("Bias/Fairness/Transparency Concerns" and "Need for Human Involvement"), and synergies in the area of opportunity. Although [92] claims that even before the widespread use of LLM, there was a need for a unified vision of the future of AI, the proposed guidelines fail to find a common thread. In their investigation of 84 guidelines, Refs. [93,94] highlight the consensus on fundamental AI ethics while noticing the high variation in how the principles are implemented. Based on [95], on the ethics of algorithms, Ref. [96] propose 7 essential factors for AI for Good: "(1) falsifiability and incremental deployment; (2) safeguards against the manipulation of predictors; (3) receiver contextualized intervention; (4) receiver-contextualized explanation and transparent purposes; (5) privacy protection and data subject consent; (6) situational fairness; and (7) human-friendly semanticisation". Similarly, Refs. [97–99] expand on the topic and refer to the need for a value-sensitive design, defined as a method to integrate values into technological solutions, while [100] advocates for a socially responsible algorithm and [101] proposes an ethics penetration testing for AI solutions.

Various countries are implementing AI regulations in a struggle against the black-box complexities of AI, particularly in balancing its benefits and potential harm. In the United States, the Biden administration has introduced an executive order advocating for 'Safe, Secure, and Trustworthy AI'; Canada proposed an Algorithmic Impact Assessment; the World Economic Forum an AI Procurement in a Box, and the OECD a Framework on AI Strategies [102]. The European Union is enacting its first AI regulations, although it requires a more integrated approach from the member states and governments [103–105]. The EU AI Act establishes four categories of prohibited AI practices: "(1) AI systems deploying subliminal techniques; (2) AI practices exploiting vulnerabilities; (3) social scoring systems; and (4) 'real-time' remote biometric identification systems" [105]. Still, existing frameworks, such as the Assessment List for Trustworthy AI (ALTAI), play a key role in guiding the development of fair and ethical AI [106]. ALTAI, in particular, protects people's fundamental rights [107], while other concurrent regulations, such as the General Data Protection Regulation (GDPR), protect user privacy [108]. The efficacy of these frameworks has yet to be completely determined. For example, Ref. [107] argues that we must interpret AI frameworks like ALTAI from a systems theory standpoint to be applied in various disciplines, allowing "the integration of a rich set of tools, legislation, and approaches". The scalability of ethical AI frameworks, in the context of their proper practical application in highly sensitive areas, such as sextortion, is also an area of potential

improvement for current research. Beyond ethical frameworks, regulating AI through data privacy laws as a clear instrument presents several challenges. These challenges include controlling personal data, ensuring the right to access personal data, adhering to the purpose limitation principle, and addressing the lack of transparency in AI decision-making processes. Furthermore, AI systems often introduce privacy risks by obscuring algorithmic biases, complicating the enforcement of the right to be forgotten, and affecting the right to object to automated decision-making [109]. Regardless of the scope, relevance, or enforcement of these regulations, AI applications pose significant risks beyond the range of data privacy laws.

Toward Blockchain-Enabled Ethical AI and Their Transformative Potential for a Resilient Society: The social good of AI goes beyond technology fixes; it calls for societal transformation. And ethical AI requires the participation of the communities that it seeks to enhance [110]. Although significant progress has been made in investigating the potential of AI for societal resilience, there are two noticeable gaps. The first is a critical gap in understanding the perpetuation of biases, especially in the context of sextortion, linking social, anthropological, and technological concerns. The second is a lack of critical analysis on the potential risks and consequences, with most of the literature corpus being polarized or presenting pinpointed solutions. Moreover, the current literature lacks insight into the scalability and practical implementation of ethical AI frameworks, particularly in combating sextortion. This status quo underscores the need for both an ethical framework as the one proposed in this position document, and for comprehensive studies that critically assess limitations.

Although blockchain characteristics and capacities are referenced as part of the ethical ecosystem (Figure 1), Section 3.2 offers a dedicated assessment of this technology's suitability as an ethical infrastructure (particularized on the case of sextortion), as the second phase of the research.

3.2. An Assessment of Blockchain's Suitability as an Ethical Infrastructure Complementary to AI-Blockchain Operations That Address Ethical and Trust Issues in AI Systems

The gaps identified in existing ethical AI frameworks are partially tackled by studies exploring the intersection of AI with blockchain as a promising solution for enhancing ethical oversight, data governance, and system resilience [111–117]. This research provides significant architectural and security perspectives. However, there is limited analysis of the potential of blockchain to act as an ethical infrastructure. This situation is ever more evident in specific cases, such as AI-driven sextortion mitigation systems—a gap this section of our research seeks to address.

3.2.1. Conceptual Foundations, Contrasting AI and Blockchain

Blockchain technologies and AI are considered to be complementary and converging technologies. On the other hand, both technologies operate within their own fundamentally distinct domains, each serving their own unique purposes with very diverging operational principles. Blockchain is inherently designed with a focus on establishing trust, verification, and transparency in decentralized systems that require secure and immutable record-keeping. As such, blockchain technologies achieve this by employing decentralized consensus mechanisms, e.g., Proof of Work (PoW), Proof of Stake (PoS), and Byzantine Fault Tolerance (BFT), and so on [118]. The goal is to ensure data integrity without any reliance on a central authority as a trusted third party. Thus, this way, blockchain yields the ability to provide auditability, traceability, and tamper-proof transaction histories [119]. Consequently, blockchain technologies have become indispensable in applications such as digital identity management, secure data sharing, and transparent financial transactions.

AI, on the other hand, has its roots in data-driven intelligence with a focus on predictive analysis, pattern recognition, and automated decision-making. Thus, AI employs

machine-learning algorithms and, additionally, large-scale datasets as well to discover insights, predict outcomes, and adapt to complex scenarios in real-time [120]. Different from blockchain, AI systems typically operate without transparency in their decision-making processes and instead, excel in domains such as computational adaptability and dynamic responsiveness [121]. AI is currently very widely applied in domains that require rapid and context-sensitive analysis, e.g., in fraud detection, personalized recommendations, and autonomous systems [122].

Despite these differences between blockchain technologies and AI, as well as building on previous studies [111–117], the integration exploration in this position paper does not aim to simply merge these two distinct domains into a single framework. Instead, this position paper aims to discover their complementary strengths. Blockchain's trust mechanisms ensure that data integrity, secure provenance, and decentralized governance are critical aspects of applications that handle sensitive information [123]. On the other hand, the adaptive intelligence of AI for pattern identification and the generation of actionable insights addresses dynamic and evolving challenges, e.g., digital threats [70]. Without compromising these technologies, the combination of blockchains and AI yields the possibility of addressing complex social issues, e.g., sextortion, privacy violations, and digital exploitation [24]. The goal of such an integrated approach is to ensure that the inherent strengths of blockchain and AI reinforce one another, and thus to create more robust and trustworthy systems [124].

3.2.2. Rationale for Integration and Ethical Synergies

If blockchain technology ensures that sensitive data is not tampered with and remains private, AI enhances the former's utility by processing large datasets in a privacy-preserving manner. While blockchain establishes trust, transparency, and tamper-proof data through decentralized records, AI focuses on dynamic data analysis and adaptive decision-making. When blockchain ensures data integrity by providing immutable and auditable records, which is ideal for sensitive information management, AI enhances utility with privacy-preserving techniques by leveraging secure, verified data to detect patterns and threats, such as in the case of sextortion [125]. The described integration allows blockchain to verify data inputs that are AI-generated. On the other hand, AI leverages blockchain to ensure accountability and traceability, thereby creating a system of synergies that is capable of addressing ethical and societal challenges.

Following the narrative conceptual exploration and the SLR, we build on existing literature towards the potential complementarity of blockchain as a mitigation to the ethical weaknesses and significant governance gaps of AI. In this section, we evaluate whether blockchain's key features—decentralization, immutability, traceability, and consensus mechanisms—can mitigate the aforementioned ethical issues. Rather than introducing the roles and impacts linked to these aspects in isolation or comparatively, we posit that they are complementary and thus, we integrate them into a unified perspective presented in Table 4 in Section 4. Nonetheless, here, we advocate may be used jointly to tackle significant societal resilience challenges, particularly in the case of digital abuse and sextortion. As shown in Table 4 below, ethical issues in AI applications that mitigate sexual exploitation leading to sextortion can be addressed with the integration of blockchain concepts. Blockchain technologies and concepts such as federated machine learning, data wallets, transaction auditability, DAO governance, Zero-knowledge proofs, and smart contracts may be adopted to address some of the identified ethical issues.

Table 4. Ethical issues in AI are addressed through blockchain concepts.

Ethical Issues in AI	Blockchain Concept	Blockchain Role in Mitigation	Referenced Studies
Accuracy (false positives)	Federated Machine Learning	Like other ensemble AI systems, it can improve the performance of the models and also provide more representative datasets since data is distributed across different locations, countries, and regions, ensuring that systems that mitigate sexual exploitation have higher performance	[126]
Privacy (data misuse)	Data wallet	Permission granting and revoking to prevent unauthorized data usage, thereby ensuring that victims of sexual exploitation maintain control and use of their data in training AI systems	[127]
Bias	Data auditability	Auditable datasets on the blockchain provide transparency and help in identifying biases in datasets	[128]
Censorship	DAO governance	Community-based governance of AI processes, instead of centrally controlled, ensures decisions on the mitigation of sexual exploitation are reached in a more democratic manner	[129]
Trustability	Smart contracts	On-chain logic is verifiable, and ethical guidelines can be encoded in smart contracts, ensuring that the basis for identifying sexual exploitations is verifiable and transparent	[24]
Security and confidentiality	Zero-knowledge-proof systems	Privacy-aware data processing systems such as homomorphic encryption can be adopted in processing confidential data such even if there are data breaches, information and data of victims of sexual exploitation remain hidden	[130]
Regulation on (legal) jurisdictions	Federated Machine Learning	Data are processed at the point where they are generated adhering to local regulations this ensures that different legal jurisdictions and regulations regarding the processing of sensitive data are adhered to specific to locations where the decentralized AI models run.	[131]

3.3. Blockchain Operations That Address Ethical Issues in AI

3.3.1. Federated Machine Learning and Blockchain for Privacy-Preserving and Censorship Resistance

Figure 5, adapted from [132,133], shows a simplified process representation of federated AI systems (a form of ensemble modeling) integrated with blockchain technologies. The process consists of several organizations {org1. . .n} in BPMN lanes, where each organization controls its own data and ML models used in the federated system. First, each organization prepares the models running in it by training and retraining with its datasets. A blockchain-based smart contract is used to verify the performance of the models from each of the organizations to select the models that meet the minimum defined model performance. Using a consensus mechanism of result aggregation, predictions from each of the selected models are ensembled into a single output. Operations in blockchain-enabled

federated machine learning can address some important ethical issues in AI, such as Regulation on (legal) jurisdictions, Accuracy (false positives), Censorship, and trustability.

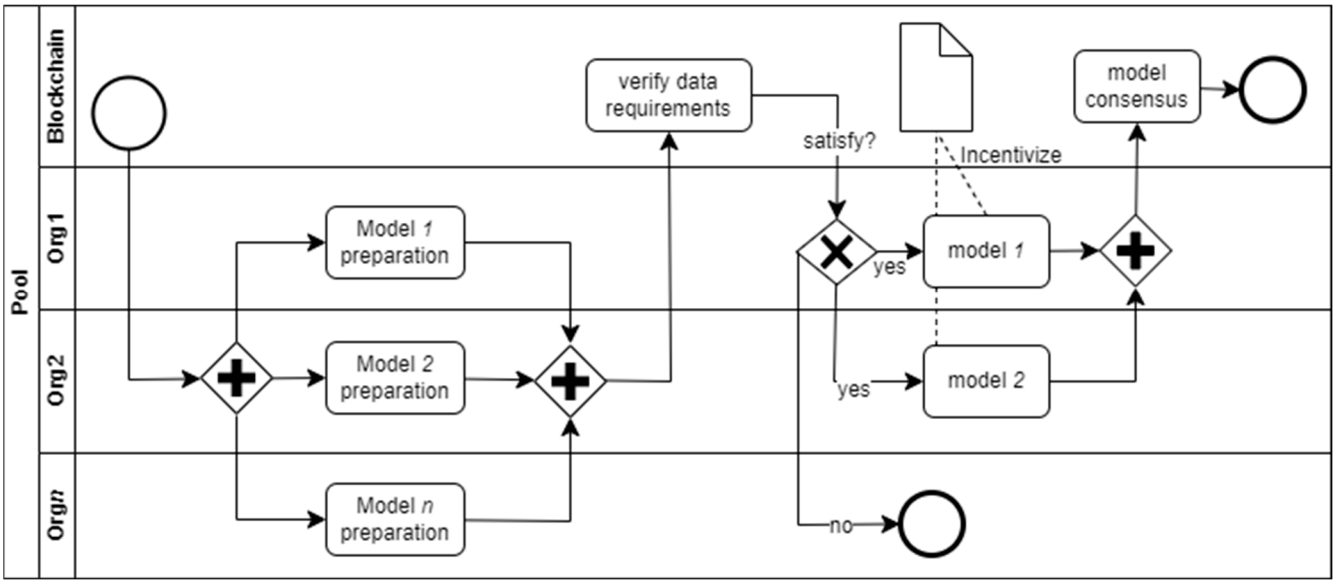

Figure 5. Decentralized Machine Learning Operations.

Using the federated learning approach, the sensitive data, such as healthcare records of sexual exploitation victims, used in training the models, does not leave the organizations (and legal jurisdictions) where they are generated. This ensures that data is not unnecessarily moved across legal jurisdictions and addresses *regulatory issues* in moving sensitive data during AI processes. Federated machine learning, being a form of ensemble approach, generally has higher performance than single model approaches, thereby increasing the *accuracy* of AI systems used in sexual exploitation use-cases. Furthermore, since the entities involved in the federated machine learning are independent organizations, a decentralized autonomous organization (DAO) can be established to govern this collaboration, thereby preventing *censorship*. The decentralized governance can determine the minimum data requirements and model performance for an organization to participate in federated learning. This ensures that no single organization can easily undermine or limit access to AI applications that address important societal problems like sextortion. Smart contract verification ensures that only models that meet the requirements specified by the DAO governance are used (ensembled) in sensitive AI applications, hence, improving the trustability of the system.

3.3.2. Data Wallets as Tools for Survivor-Controlled Consent and Data Use

Figure 6, adapted from [127], shows a BPMN representation of operations in a given organization (*org1*) where data containing user-sensitive information is preprocessed and prepared for use in AI applications for sextortion prevention. The process starts with data collection, and the data owners are incentivized to share their data for training AI systems that mitigate sexual exploitation. The training data, containing only datasets whose data owners have granted their permission, is aggregated for use in the AI model training. A smart contract on the blockchain verifies the permissions on the dataset, ensuring that no data with revoked permission is used during the AI training. The AI models are trained and retrained while adhering to the permission-granting and revoking rights of the data owners. The operations in the Data Wallet permission system address important ethical issues in AI systems, such as *Privacy* (*data misuse*) and *Trustability.*

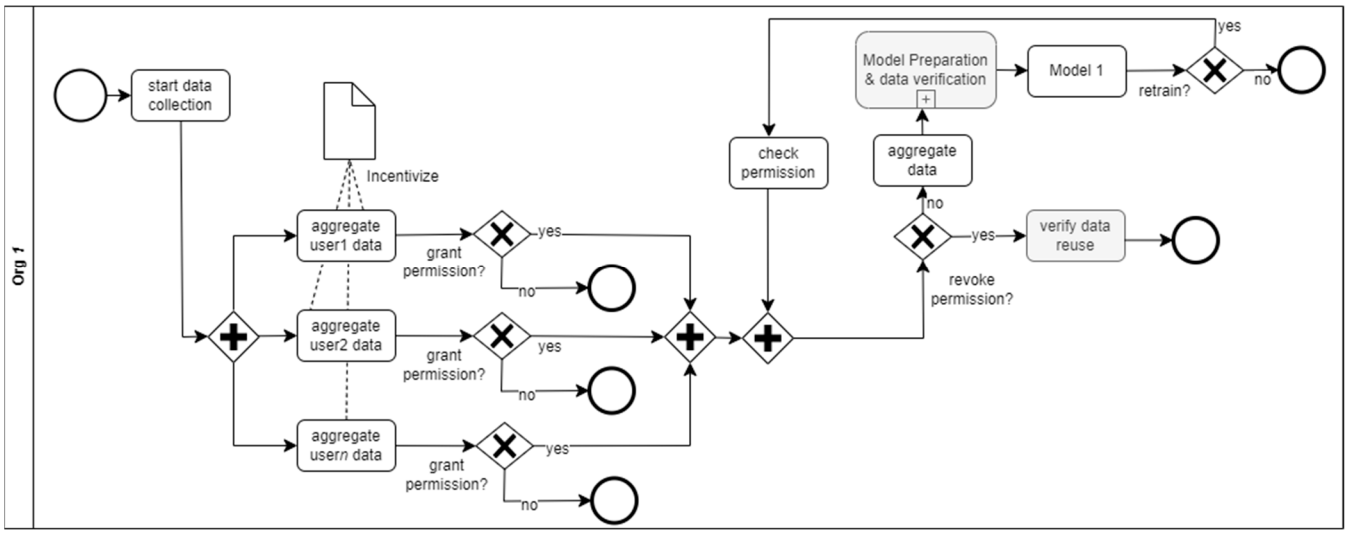

Figure 6. User Data Wallet Operations.

The permission-granting and revoking system in the Data Wallet improves the user privacy in AI applications by ensuring that users can share their data for the training of AI applications that they care about, such as applications that address societal issues like sextortion and sexual exploitation. The data owners also maintain full control over their data since they can also revoke previously granted permissions. Smart contract verification within the data wallet system ensures that operations of the AI are trustable and verifiable and that only the data with the right permissions is used in training the AI applications.

3.3.3. Smart Contracts and Zero-Knowledge Systems for Security and Confidentiality

The federated machine learning and data wallet operations in Figures 5 and 6 contain various smart contract verifications, such as data requirements, model performance, and validity of permissions. Performing these verifications both on the blockchain and off the blockchain can leak sensitive information about the data owners (such as the victims of sexual exploitation). Zero-knowledge proof (ZKP) provides a system for verifying claims without revealing sensitive information about the claim [130]. The formal proof about a claim that needs to be verified is stored on the blockchain, while a verification smart contract is used to validate the claim before proceeding to the next steps of AI operations. The blockchain-enabled ZKP system, when integrated into AI operations for mitigating sexual exploitation, can address *privacy* and *confidentiality* issues.

Verification of data used in training an AI system can leak sensitive information about the victims of sexual exploitation in the datasets. Permission claims on the datasets can be traced to the owners of the datasets, who are also the victims of sexual sextortion. With the use of ZKP, claims made by the organizations that control the federated AI models can be verified while hiding sensitive information. For instance, the formal proof of the permission claim is only stored on the blockchain, while the verification of this claim does not reveal the data owner who granted the permission. Furthermore, organizations may also not want to leak the performance of their individual AI models (in the federated set-up) to the general public, hence, homomorphic cryptographic mechanisms can be adopted to perform encrypted consensus arithmetic that produces the combined output from the ensembled models. This ensures the privacy of data owners (victims of sexual exploitation) and federating organizations (who host AI models that prevent sexual exploitation) while ensuring high-performance and trustability of the AI system.

3.3.4. Interplay of Federated Learning and ZKP—Compatibility Logic

Another critical issue that persists is the secure interaction between the federated learning (FL) components and the blockchain layer through zero-knowledge proof (ZKP). This interoperability is crucial to prevent local data from leaving local devices as well as to enable a distributed governance architecture for validating updates of models and operations on data without accessing the raw inputs.

The technical pipeline is as follows: the FL clients locally train the models on the private subsets of their data and calculate the intermediate updates (i.e., gradients or weights). These updates are subsequently encoded into small cryptographic attestations with zk-SNARKs, for example Groth16 or PLONK, which convince that the updates were computed correctly with respect to pre-agreed training parameters, while withholding the underlying information. The ZKPs are then published to the blockchain alongside the commitments of model hashes, giving validators or governance actors the means to verify that the calculation is not compromised, without transgressing privacy.

For the purpose of proving forgery or client malbehavior, each FL-ZKP bundle is bound by a per-session identifier and signed by the participant's identity key organized under a decentralized identity (DID) framework. Smart contracts are used on the blockchain to verify the proof and for future trust audit hashing. The DAO governance module can meanwhile utilize this proof of reliability to invoke reward distribution, punish hostile updates or invoke retraining procedures.

The architecture is designed in a way so that the heavy computations (FL model training and ZKP generation) are separated to off-chain infrastructures and the on-chain operations only entail lightweight proof checking and logging, which prevents the issue of performance bottleneck. Furthermore, modularity of design allows the FL model can be iterated independently of blockchain upgrades, improving its maintainability over the long term as well as its extension for cross-domain.

Another technical issue left as a challenge is the interoperability between diverse blockchain systems (e.g., due to data sovereignty, jurisdiction diversity, and governance diversity, it is expected that the multi-chain deployment will become a common practice). Even though CF-BIAI-SXT does not directly indicate communication protocols between the chains, this could be provided by decentralized interoperability protocols such as Cosmos IBC or Polkadot's relay-chain logic to establish secure and verifiable communication of user credentials and consent data between chains.

3.3.5. Token Models and Participation Incentives for Inclusiveness

As shown in Figures 5 and 6 in the federated machine learning and data wallet operations, the main entities in the systems are organizations that control the AI model and data owners, such as victims of sexual exploitation. The data owners need to be incentivized to grant the necessary permissions for AI systems to use their data for training and retaining AI models that prevent sexual exploitation. Blockchain-based tokens provide a system for implementing a reward token for entities that actively contribute to the federated learning system. Such a token system can be well grounded on token economics that ensures that the reward system properly incentivizes both the data owners and model owners to actively participate in the federated network while addressing their separate needs. Although the token economics of incentivizing entities in federated learning and data wallet operation does not directly address ethical issues in AI systems, it can reduce bias and also improve the performance of the AI systems.

By properly incentivizing data owners and victims of sexual exploitation, more individuals will share their data, hence providing a broader dataset for training the AI systems that address social problems. For organizations that host the AI models, proper incen-

tivization ensures that they always produce models that meet the minimum performance requirements for selection into an ensemble combination of models. For instance, an incentivization rule can ensure that only organizations whose models are selected are rewarded. This ensures that the organizations will produce high-performance models, which also translates to improved performance of AI systems that address sexual exploitation.

In reality, the very technical insights presented above may be easier understood through a narrative scenario: imagine a 16-year-old victim of sextortion in Romania uploading encrypted metadata of the high-risk event to a trusted app. The AI model flags suspicious coercive language patterns; the blockchain logs this action without exposing identity; a federated learning network updates detection models across jurisdictions without moving raw data; local NGOs may validate the high-risk event and initiate support. The DAO overseeing the federated platform adapts detection rules in response to emerging patterns. The CF-BIAI-SXT depicted in the following section, in line with the emerging requirements of the EU AI Act and the ALTAI framework, ensures that all these agents act in synergy to ensure that the victim is protected and the sextortion mitigation is achieved.

Although ethical soundness is highly emphasized in the design phase, the actual implementation of CF-BIAI-SXT should factor in the performance overhead and storage expenses in blockchain-AI integration. For example,

- Federated learning cost: On-chain use of smart contracts to verify performance on local models introduces latency (~200–500 ms/transaction on PoS chains like Polygon).
- ZKP and encryption overheads: Zero-knowledge protocols (e.g., zk-SNARKs) result in a computational overhead of 20–40× over regular inference pipelines.
- Storage bloat: On-chain anchoring of audit hashes without outsourcing to distributed file systems (like IPFS or Filecoin) could increase costs to $10–50 per GB per month on public blockchains.

These values are consistent with results of recent federated learning and blockchain-ZKP integration studies. Xing et al. show that the cost of complete zero-knowledge proofs in federated learning scenarios is less than one minute in time with the use of optimized zk-SNARK implementations [134]. Ogungbemi also discover that privacy preserving smart contracts written in ZKPs could be used to generate proof approximately in ~1.9 s, while gas cost is a bit higher for higher privacy levels [135]. Keshavarzkalhori et al. quantitative generalization is at least questionable that there are no inexpensive Ethereum-based FL systems, since audit costs and gas consumption both depend on the storage and complexity of models, and could explain the $10–50/GB/month claim [136].

To partially counter these issues, it is suggested that layer-2 scaling, to settle transactions off-chain (e.g., iExec), and credentials with selective disclosure protocols be used for better trade-off between verifiability and cost. These trade-offs will be tested empirically in future stages of implementation.

4. A Privacy-First, Societal-Resilience-Oriented Conceptual Framework Integrating AI and Blockchain as Complementary Technologies for Tackling Sextortion

4.1. Conceptual Framework: Integration Logic and System Architecture

Building on the previous foundational blocks of AI and blockchain in the context of high-risk digital environments, we propose a conceptual framework that aligns the technical capacities of both technologies with broader societal, ethical, and governance considerations. This framework is meant to structure their interaction in a manner supporting trust, resilience, and regulatory compliance in AI applications. This design, anchored in decentralized infrastructure with an ethical angle (privacy-first), contributes to strengthen-

ing societal resilience against digital threats such as sextortion, and its main design logic is anchored in the dual role of the two technologies as depicted in Table 5 and visually presented in Figure 7.

Table 5. AI and Blockchain Risk and Resilience Framework.

	Data Privacy and Security	Ethical Concerns	Operational Risk
Manifestation	Unauthorized data access or breaches	Bias, fairness, or lack of transparency in AI systems	System failure or misuse of AI/Blockchain
AI Intervention	Data encryption, secure multi-party computation	Transparent algorithm design, fair training datasets	Error detection and self-correction mechanisms
Blockchain Intervention	Immutable ledger for auditability, access control via smart contracts	Blockchain for ensuring transparency in decision-making processes	Decentralized, fault-tolerant architecture, consensus mechanisms
Integrated Benefit	Secure and verifiable data processing	Transparent, traceable, and fair decision-making	Robust and resilient system performance
Combined Impact	Improved data protection and user trust	Enhanced accountability and fairness in AI outcomes	Minimized operational downtime and systemic risk
Example or Tool	Privacy-preserving models, Blockchain-based consent systems	Fairness metrics, decentralized auditing systems	AI error detection tools, fault-tolerant consensus protocols

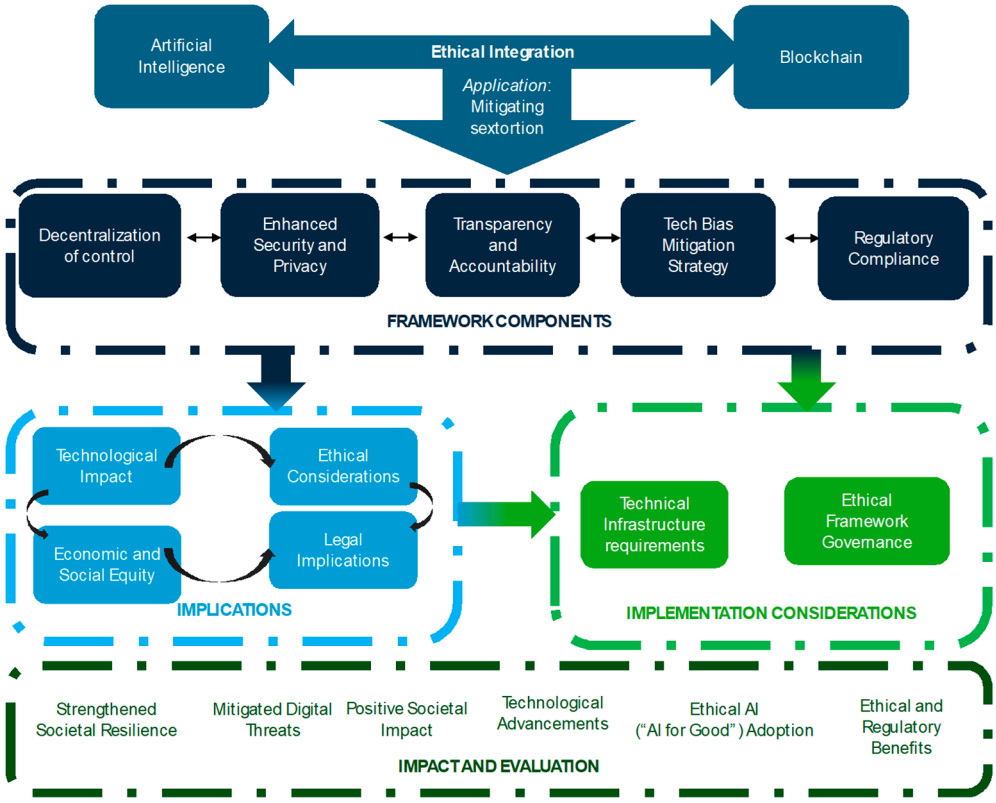

Figure 7. A Privacy-First, Societal-Resilience-Oriented Conceptual Framework Integrating Blockchain and AI for Ethical Mitigation of Sextortion.

The integration of AI and blockchain into a unified conceptual framework strengthens technical resilience and also advances ethical governance through traceability, decentralization, and privacy-preserving mechanisms. This path was also considered by [67], who argued for the necessity of combining AI and blockchain to address the escalating gover-

nance challenges posed by generative technologies. The integration is perceived, thus, as a structural mitigation measure or safeguard, and the CF-BIAI-SXT becomes a systems-level response to high-risk digital harms. Hence, the following characteristics: explainability, control, and institutional accountability, must be embedded by design. The framework has five core components (placed in the first layer in Figure 7):

- Decentralization of Control—local governance, DAO logic, user data ownership.
- Enhanced Security and Privacy—federated learning, ZKP, encryption.
- Transparency and Accountability—audit trails, explainable AI, immutable logs.
- Tech Bias Mitigation Strategy—data integrity, distributed validation.
- Regulatory Compliance—GDPR-aligned design, consent tracking, legal risk.

If, in the previous section, we referred to them in a mechanism-by-mechanism format, here we integrate them holistically, reassembling them in system-level integration logic. However, complementary to the technical efficacy of such a framework/high-risk AI system, the strength of assessing it in a Future Internet and/or societal resilience context relies on the multi-criteria evaluation lens detailed in the following sections and anchored by four dimensions: technological impact (e.g., architectural changes in AI workflows), ethical considerations (e.g., shifts from opaque to auditable systems), legal implications (e.g., evolving liability, compliance design) and economic and social Equity (e.g., data ownership, accessibility of AI protections).

4.2. Regulatory and Legal Harmonization

The proposed framework, which is mainly technical, must first be anchored in legal reality, and this sub-section explains how it fits into the real-world regulatory fabric—GDPR, AI Act, cross-border data, etc. Given the divergence between the two technologies and the resulting regulatory tensions highlighted in Table 6, specific challenges must be taken into account for effective deployment of blockchain-integrated AI systems [137].

Table 6. Key Legal Tensions between AI and Blockchain Regulatory Logics.

Issue	AI Systems	Blockchain Systems	Resulting Conflict
Data deletion (Right to Erasure)	Requires the ability to delete or modify user data (GDPR, AI Act)	Immutability prevents data deletion once recorded	Contradicts GDPR's "right to be forgotten"
Data controller accountability	Centralized or joint controllers are usually identifiable	Distributed nodes make accountability attribution difficult	Legal uncertainty over controller responsibility
Data minimization	Requires minimal, purpose-bound data collection	AI often relies on large, central datasets	Conflicts with blockchain's preference for minimal storage
Transparency and auditability	Emphasis on explainable models and traceability	Emphasis on immutable, transparent records	Overlap exists, but also tension when AI uses opaque models
Privacy protection (data sensitivity)	Strong emphasis on safeguarding personal data	Blockchain makes all transactions visible unless privacy-enhancing tech is used	Conflict unless privacy-preserving tech (ZKP, homomorphic encryption) is added
Cross-border data regulation	Subject to jurisdiction-specific rules (GDPR, CPRA, etc.)	Blockchain networks operate across jurisdictions with no centralized control	Regulatory compliance is complicated by decentralization

The reconceptualization of governance models requires aligning technical systems with legal and institutional architectures, as also shown by [138] in a context of policing of cybercrime. This is even more valid in the context of decentralized AI-blockchain systems, which challenge conventional accountability frameworks. However, there are numerous discrepancies and regulatory gaps, making the entire process of alignment extremely difficult and often bound for failure. One possible solution is the adoption of so-called regulatory sandboxes [139], providing a controlled environment for the testing of innovative blockchain-based AI systems in the context of relaxed regulations [140], and allowing for interdisciplinary collaboration to identify compliance issues and refine frameworks in real-world scenarios [141]. These sandboxes may be created under national data protection authorities or blockchain consortia, if supranational entities, such as the European Commission, fail to deploy them due to bureaucratic aspects. In parallel, self-regulatory principles aligned with frameworks like the OECD AI Principles and EU Blockchain Strategy offer an interim solution, enabling industry stakeholders to set governance standards while awaiting formal regulations [102]. Furthermore, cross-border regulatory divergence—particularly between the GDPR in the EU and the sectoral approach in the US, which proposes regulations for each type of the following: healthcare, finance, children's privacy rights, and education—complicates compliance for decentralized AI systems. Nonetheless, there are signs of regulatory convergence, as California adopted the Privacy Rights Act (CPRA), which incorporates many elements from the GDPR of the EU, including its treatment of automated decision-making.

4.3. The Pathways—A Layered Implementation Strategy for the CF-BIAI-SXT Framework

For the framework to be usable, apart from regulatory elements, technical and organizational preconditions must be met, alongside proper risk mitigation, adequate governance, and interoperability. Thus, we have three layers to consider for proper implementations: Technical Enablement, Governance and Stakeholder Alignment, and Trust and Oversight. The intersections of these dimensions, often creating design trade-offs -such as between transparency and computational efficiency, or between decentralization and scalability—must be properly tackled by governance issues and risk mitigation to ensure implementation success.

4.3.1. The Technical Enablement Layer (What Infrastructure Is Needed for the System to Work in Practice?)

This foundational layer ensures that privacy, scalability, and interoperability are embedded by design. It allows the BIAI system to operate effectively across jurisdictions and sensitive use cases. Key elements to be considered are trust-by-design and local data retention are crucial for mitigating sextortion.

Blockchain systems must record all AI-critical activities (such as model training, inference decisions, and data access) on immutable ledgers, guaranteeing verifiability, supporting post hoc auditing, and ensuring accountability, thus mitigating the risk of norm violations. Smart contracts can encode procedures for rule enforcement, compensation, and conflict resolution, reducing reliance on intermediaries and strengthening procedural fairness. Continuous system monitoring, through log review, performance audits, and advanced diagnostic tools using techniques like noise-aware sparse Gaussian processes and edge-AI analytics, may further enhance predictive accuracy and system responsiveness in real-time [142,143]. For example, the findings of [144] show that blockchain and AI in combination improve surveillance and tracking capabilities by ensuring detailed, secure, and immutable records of all transactions.

A series of key technical considerations for deploying BIAI systems in real-world, cross-jurisdictional contexts are summarized in Table A4 in Appendix C. These include the use

of federated learning, privacy-preserving computation (such as ZKPs and homomorphic encryption), and secure digital identities. In addition, also to be considered is the case of lower-resource jurisdictions (from which a significant number of victims emerge), where edge-based implementation may offer an alternative to cloud reliance, but may also require mobile-first or hybrid deployments.

4.3.2. The Governance and Stakeholder Alignment Layer (Who Runs It, Who Participates, and How Governance Legitimacy Is Ensured?)

This governance layer is central in translating the BIAI framework's technical elements into ethical context-aware systems capable of earning public trust. The key element to be considered is the importance of adaptive governance, essential when ethical risks evolve faster than regulation.

Additionally, effective governance frameworks are essential for the management of blockchain-integrated AI systems that tackle sexual exploitation issues. Such governance systems also need to be aligned with societal values and regulatory standards to deliver on their desiderate to strengthen societal resilience. Moreover, to properly ensure effective implementation of BIAI, they must fulfill several core functions:

- Support decentralized decision-making
  - Distributed authority prevents centralized abuse and improves transparency in high-risk contexts like sextortion.
  - DAO logic enables rule-setting via encoded, consensus-based processes [145].
- Integrate diverse stakeholders
  - Policymakers, ethicists, technologists, regulators, and victim advocacy groups must collaborate on design and oversight [146].
  - Such inclusion allows for democratic control and societal legitimacy.
- Encode and enforce ethical norms
  - Governance must ensure adherence to principles like fairness, privacy, accountability, and non-discrimination.
  - Recent frameworks offer guidance for ethical AI deployment [147].
- Enable dynamic policy adaptation
  - Regulatory rules must evolve with technology. Blockchain-based systems allow real-time enforcement without intermediaries.
  - Smart contracts can facilitate fluid, programmable policy [148].
- Maintain trust through institutional alignment
  - Ethical governance models must be aligned with legal norms and societal values to avoid deviation from initial commitments due to technical updates, commercial incentives, or regulatory gaps.
  - Trust frameworks should be embedded to guide system evolution [149,150].

4.3.3. The Trust, Oversight, and Continuous Evaluation Layer (How Legitimacy, Transparency, and Ethical Compliance Are Monitored?)

This third layer operationalizes the ethical backbone of the system, enabling the prevention and resolution in real-time of violations.

Key elements to be considered: accountability, resilience, and societal trust. Trust is central for BIAI in high-risk digital environments; therefore, ensuring it becomes crucial to the effectiveness of the framework. This depends on the capacity of the BIAI systems for transparent operation, reliable oversight, and responsive error correction, which may be achieved through a monitoring and auditing mechanism that fulfills three core functions:

Verifiability and Immutable Logging, Smart Contract–Based Enforcement and Dispute Resolution, and Real-Time Oversight and Adaptive System Monitoring. Complementary safeguards include system-level logging and auditability, ethics penetration testing, sandboxing and staged rollouts, third-party algorithmic audits, and participatory evaluation (users, survivors, civil society). In addition, these functions interact, as verifiability ensures post hoc accountability, smart contract enforcement addresses real-time disputes, and adaptive monitoring ensures forward-looking risk detection.

Lastly, the successful implementation of BIAI systems hinges not just on technical readiness but on their sustained alignment with evolving ethical, regulatory, and operational expectations. Although highly advanced techniques may improve performance, the societal impact resides in the capacity of the technical solution to be embedded within a responsive and accountable governance framework. Even more so in high-risk domains like sextortion, where maintaining public trust and resilience over time is as important as initial system effectiveness. Therefore, implementation must be acknowledged as a process of iterative socio-technical coordination, not a one-off deployment.

To address the widely noted issue on tension between blockchain immutability and legal rights such as the "right to be forgotten", the framework suggests a hybrid storage solution. PII (personally identifiable information) is never written to the chain. Instead, on-chain only hashed pointers or zero-knowledge proofs (ZKPs) are kept, linking to off-chain encrypted records stored in the possession of the respective users in what we call a data wallet. Data deletion can therefore be implemented through revocation of the keys and off-chain deletion, and the on-chain reference to the data is left ungatherable yet the audit integrity is maintained in the blockchain.

This is consistent with the position of the framing interpretations of the EU (e.g., EDPB guidance), which acknowledge a functional deletion or economic inaccessibility as a reasonable way of applying data erasure. Additional compliance is enforced through the provision of consent-granting timestamps, which allow fine-grained audit trails for legal attestation.

4.4. Discussions: Towards Real-World Application and Stakeholder Alignment

Once the CF-BIAI-SXT is articulated as an ethical infrastructure, it must also be fit into the real world. This part of the paper discusses this thematic point, especially how BIAI can promote equity and support institutional resilience in responding to digital sexual exploitation.

4.4.1. Empowerment Through Data Justice and Participation

Decentralization affects more than just control of individual information—it alters the very structure of power in data ownership. Historically, data were at risk of misuse, as access to them was gate-kept by a small number of large corporations and government agencies, as described in [151]. This status quo is rebalanced through blockchain technology, which allows individuals to own and control their digital assets, leading to consumer empowerment and fairness in making algorithmic decisions. For the use of legal and social services, this builds on a more inclusive dataset, providing a scenario in which algorithmic decision-making can be less biased than currently.

Using the concept of decentralization, as delivered by BIAI, alongside transparency and enhanced security, we can identify several directions in which economic and social equity in the context of mitigating sextortion may be discussed. The structure of the discussion moves from micro to macro and from victim to perpetrator, highlighting how BIAI can positively affect this chain. First, BIAI ensures democratized access to financial resources and digital services, promoting economic equity according to the token economic

model shown in the previous section. This feature may lead to providing a secure platform for sextortion victims to seek help without the danger of additional exploitation.

Second, BIAI supports social equity by increasing security through privacy-aware data processing enabled by ZKP systems. For victims of sextortion or potential victims, this level of security and accountability is essential because it gives them confidence that their data is being stored and handled properly and that there is accountability in such instances when a data breach occurs. This empowers victims, encouraging them to seek help and resources for protection; in addition, the blockchain side of BIAI facilitates decentralized reporting systems, which in turn permit sextortion victims to report incidents securely and anonymously. As a consequence, the risk of social ostracism is mitigated and social equity is advanced by safeguarding the well-being of victims.

4.4.2. Systemic Equity and Institutional Reform

The link of sextortion with economic and social equity is three-pronged: from causes to consequences and specific implications for policies and practices. Thus, sextortion may be driven by socioeconomic factors (such as economic uncertainty, gender issues, and patriarchal societies). However, it may also lead to micro and macroeconomic effects. For example, it can cause economic vulnerability and instability in victims, increase social inequality, and perpetuate stigma and ostracism of already marginalized individuals [152,153]. Third, when the victim does not come forward, BIAI can support identifying red flags of such an event occurring, empowering law enforcement with analytics to improve understanding of related crime patterns. By analyzing immutable logs and metadata trails, abuse patterns in transactions can be highlighted, allowing for the prevention of injustice and making the system proactive, rather than reactive. Fourth, in addition to this BIAI-supported recognition of victims, AI-supported educational and awareness programs may prove beneficial in supporting and empowering vulnerable individuals. These types of programs equip victims with the knowledge to recognize and respond to threats.

4.4.3. Alignment with Public Institutions and Law Enforcement

Fifth, on the law enforcement side, after the victim has been identified and the sextortion crime has been committed, BIAI can support the justice system in several ways, ethically, by following a supranational ethical framework:

- Transparent legal processes—BIAI may guarantee that legal decisions are transparent and auditable, which will help combat corruption [154]. In turn, corruption, through bribery and the influence of perpetrators, has been proven to affect effective prosecution of sextortion cases, for example, in South Africa with migrants [152].
- Tamper-proof evidence—Blockchain can ensure the integrity of the chain of custody [155], while AI can be employed in the analysis and management of evidence and digital forensics.
- Support for marginalized groups—One of the groups most affected by sextortion is represented by migrants and refugees [152], who are also affected by a lack of easily verifiable identification and may have access to legal aid and support through blockchain-based identity verification.

Sixth, while the crime has been proven and aid is directed towards the victim, BIAI may also bring efficiency and security. Blockchain-enabled anonymous payment mechanisms, further down the mitigation process, ensure that the aid aimed at victims reaches recipients securely and anonymously without the risk of misappropriation. Blockchain has been proven to improve social welfare programs and promote equitable resource distribution by ensuring that funds are not smuggled into intermediaries. Last, at a more macro level, we must consider a two-pronged perspective: on the one hand, the role of AI in the use of

decentralized data to improve decision-making, and, on the other hand, the transformative role of AI and blockchain for significant parts of society. This dual perspective leads us to reassert the need for an integrated approach to ensure the proper and ethical use of BIAI.

4.4.4. Risks and Governance Challenges in Blockchain-AI Integration

The proposed CF-BIAI-SXT framework highlights blockchain as a promising ethical infrastructure for AI-based sextortion mitigation. However, there is the risk of over hyping its benefits; therefore, an acknowledgement of potential risks and governance challenges that may arise from such integration, particularly in high-risk, high-stakes environments, is needed.

The first risk comes from the DAO Governance Vulnerabilities. DAOs are susceptible to governance capture or majority collusion in scenarios where voting power is token-based, as well-resourced actors might gather disproportionate influence and, thus, skew decision-making and undermine fairness. Potential mitigations for such risks are quadratic voting mechanisms, reputation-based access, or the introduction of veto rights for verified civil society nodes.

The second risk also comes from collusion (coordinated manipulation of model updates), but this time in federated learning that relies on distributed training across organizational nodes. This risk is enhanced by the opacity of local training environments. A potential mitigation is the use of smart contract-enforced audit trails. Another important risk derives from the token-based reward mechanisms. For instance, victims may be or feel pressured to contribute data to earn tokens, or bad actors might submit synthetic data to game model updates. The token incentive distortions may lead to a gamification of pseudo-consensual sextortion, inducing data sharing. A mitigation measure may be the use of real consent tracking via data wallets. Lastly, the complexity of BIAI systems brings a significant risk, with edge-case failures, such as smart contract bugs, ZKP misconfigurations, or identity spoofing, having outsized consequences and, more importantly, carrying the potential to retraumatize users or compromise legal evidence. This risk may be mitigated by ongoing third-party audits, staged rollouts, and real-time risk monitoring.

4.5. Operationalization of Ethical Principles: Privacy-First and Societal Resilience Metrics

Moving beyond high-level statements of principles means we need operational metrics to allow the empirical validation of guiding principles in the framework. Especially, the desires of being "privacy-first" and improving "societal resilience" need to be described in terms that can be monitored Benchmarkable in real or simulated deployments. These principles remain as normative ideals, unless they are operationalised as system requirements.

To be able to address these issues, we suggest a preliminary list of quantitative and qualitative indicators that can be used for assessing the performance of the system in further pilot implementations. These indicators are further consolidated by two overarching categories, identified as: (1) privacy protection and user control; and (2) resilience-enabling features, such as inclusivity, accessibility, and ethical response latency. The measures are to be employed in prototype testing of concept and eventual field implementations and stand as a foundation for ongoing audit and adaptive system governance.

These measures act as accountability checks as well as signposts for participatory improvements in later design iterations. This list will be extended in future work, accounting for the details of deployment within the individual systems and the input from stakeholders. Table 7 shows the relevant metrics and recommended thresholds.

Table 7. Operational Metrics for Evaluating "Privacy-First" and "Societal Resilience" Principles.

Principle	Metric	Definition	Evaluation Method	Target Threshold
Privacy-first	Data Minimization Ratio	Proportion of user data retained vs. total data accessed	Logging AI queries and ZKP audit trails	<0.10
Privacy-first	Consent Auditability	% of data operations traceable to explicit user consent	Smart contract-based data wallet logs	100%
Societal resilience	System Accessibility Index	% of system functions available in low-bandwidth environments	Network simulation and fallback logs	>90%
Societal resilience	Governance Inclusiveness Score	Ratio of system decisions made through DAO voting vs. centralized override	Smart contract vote audit logs	>75%
Societal resilience	Abuse Detection Latency	Time from abuse-report trigger to flagging response	AI + blockchain event log comparison	<5 min

5. Conclusions

Blockchain and AI integration represent a transformative step toward addressing complex societal and technological challenges, especially when used in addressing sexual exploitation. Practical applications such as using federated learning for sextortion mitigation and blockchain-audited AI decision-making demonstrate the potential for these technologies to align with ethical governance while solving real-world problems. In this paper, we are tackling this literature gap, identified through both related work analyses as well as a systematic literature review, by proposing a privacy-first, societal-resilience-oriented conceptual framework. Specifically, we investigated to what extent decentralized immutable traceability strengthens AI systems by addressing the ethical issues inherent in them. Since new Internet-based technology builds on previous ones to support social and economic activities, and thus positive societal evolution, the future of the Internet must be inextricably tied to the ethical design of its enabling technologies.

This article positions the CF-BIAI-SXT framework in the context of established AI ethics guidelines (such as the EU's High-Level Expert Group on AI proposal from 2019 of Ethics Guidelines for Trustworthy Artificial Intelligence or, from the same year, the OECD AI Principles), as presented in Section 3. The ethical gaps and solutions identified in our sextortion case study directly relate to principles of trustworthy AI and, hence, reinforce the idea that the CF-BIAI-SXT framework is designed to uphold the core values deemed essential for socially responsible AI deployment.

To conclude, our position paper highlights the potential of integrating AI and blockchain technologies to combat this ever-growing digital threat that is increasing for vulnerable societal categories, such as minors. In a multi-layered approach, we suggest using AI for the early detection of sextortion attempts and providing immediate support to victims, while blockchain ensures secure, anonymous reporting and evidence storage. We also correlate blockchain's transparency with increased trust in AI systems, which is crucial when handling sensitive issues. The study indicates that this technological integration can enhance privacy and data control for users, key factors in sextortion prevention; thus, we propose using smart contracts for secure case management and evidence collection. Overall, our findings suggest that combining AI and blockchain can significantly strengthen societal resilience against sextortion by empowering victims and supporting law enforcement efforts.

Ethical and governance risks: Nonetheless, as previously stated, there is a significant literature gap in empirical studies, with the goal of using AI to support societal resilience

thus hindered by a series of challenges researchers encounter when addressing this topic (sextortion), such as:

- *Limited Empirical Research at Risk of Obsolescence*: As both the AI-sextortion and blockchain-sextortion areas are in their infancy, empirical research remains very limited. Only a few studies deal with the use of AI (or blockchain) to detect, prevent, or mitigate the phenomenon, and even fewer focus on the nexus of the three concepts. This lack of transdisciplinary research holds back technology from having direct applicability in solving social problems and may have major implications in the long run. Studies might become obsolete or irrelevant as the time needed for academic research can be excessively long compared to the speedy pace of technological changes.
- *Interdisciplinary Complexity*: Apart from the two technologies considered, aspects related to psychology, sociology, anthropology, ethics, law, and economics must be integrated to gain a clear perspective and propose potential solutions. For this, we used two versions of reviews (a systematic and a narrative) and included a flow diagram and code protocols in Appendix A to facilitate independent verification of both.
- *Data Use and Sensitivity*: Extensive exploitation of data for educating AI or carrying out research is hindered by its sensitive nature, posing potential harm to victims, its distribution across multiple jurisdictions and platforms, in addition to the numerous ethical clearances needed from different local and global organizations. This international perspective is also affected by distinct legal and regulatory frameworks, leading to the inability to propose shared policies or resolutions.

However, this is also the stage in which, as Collingridge [156] mentions in his famous dilemma, society must intervene to ensure that a specific technology is safely deployed and is a solution, not a threat. Sextortion serves as a relevant example of this intersection. Both users and stakeholders need to understand well the ethical implications of BIAI systems by integrating extensive educational measures that are part of an integrated, comprehensive governance framework. These measures must always contain very clear information about the detailed use of data, the important AI decision-making processes, and the vital role of blockchain technologies that ensure transparency and security. Always encouraging a highly responsible use of BIAI systems with better informed consent requires the promotion of sophisticated ethical knowledge among users. Thus, a better and well-matched alignment of such complex systems with their related societal expectations and additional ethical standards can be achieved when a properly informed user base exists. However, we emphasize the need for careful ethical considerations in the implementation of these technologies to ensure they do not inadvertently cause harm or infringe on individual privacy rights. The essential imperative is to ensure that human rights are upheld and individual justice is promoted whenever blockchain technologies and AI are deployed together.

Limitations and research gaps: This position paper introduces the CF-BIAI-SXT architecture as a conceptual combination of blockchain and artificial intelligence toward ethical governance on high-risk AI applications, and more specifically for the prevention of sextortion. Even though the interdisciplinary approach and methodological framework are based on a systematic and narrative review, a number of limitations need to be considered, and avenues for further research and validation need to be stressed.

There are two reasons for this: First, empirical validation of the proposed framework is missing. The BPMN models do not provide active workflow instances, rather are conceptual abstractions, and the lack of simulation, prototype implementation, or evaluation is a shortcoming to the applicability and performance of the framework under operational conditions. Therefore, one cannot yet claim the scalability, efficiency, and real-world governance effect of the CF-BIAI-SXT design are given.

Second, despite us having abstracted out crucial architectural building blocks such as federated learning, ZKP systems, and smart-contract-based consent governance, we are yet to define these at an implementation level. No performance benchmarks, latency analysis, and throughput comparisons are found in the position paper. In addition, implementation trade-offs have not been quantified, e.g., storage overhead, blockchain energy, and cross-jurisdictional synchronization, which are the scope of further studies.

Third, the approach accepts but leaves unresolved legal frictions such as the tension between blockchain immutability and the GDPR's "right to be forgotten." While potential design patterns such as off-chain storage or revocation registries are well-known, these have not been incorporated or tested within the current model. Legal-technical co-design is emphasized as a future line of work.

Fourth, the principles of "privacy-first" and "societal resilience" are taken as axiomatic essentials rather than instruments for the measurability of a system. Better understanding how to translate such dimensions into actual evaluative numbers, such as consent traceability, fairness scores, and resilience metrics, would greatly enhance the practical utility and impact of the framework.

Fifth, although the current paper builds on a rich body of scholarly and gray literature, the central intersection of AI, the blockchain, and sextortion is topical and undertheorized in empirical research. This is supported by the low number of eligible articles from the systematic literature review, which is further supported by the need for a narrative synthesis. For a position paper, this deficit is acceptable, while on the other hand, such restrictiveness undermines the generalization and reproducibility of the findings.

Furthermore, a validation stemming from survivor-centered design and participatory prototyping would also enhance the CF-BIAI-SXT framework. If technologies are to be both socially attuned and grounded in the real world, they will need, and must incorporate, the feedback of affected stakeholders, from civil society, including NGOs, victims' advocates, and legal experts. Validation of this framework will involve proof of concept (PoC) development of a blockchain-integrated AI application that mitigates sexual exploitation. This PoC development will potentially face technical challenges associated with complex systems development and integration, including integration complexity, scalability, and interoperability challenges.

Future research imperatives. Further research is required to better understand the long-term effects of BIAI systems on societal behaviors and norms. Furthermore, up-to-date policy development is necessary to keep up with technological advances. For future work, it is important to perform an impact assessment and evaluation of the positions, concepts, and models proposed in this paper. This includes PoC demonstrations for various cases of BIAI applications that address social issues. Additional criteria for the evaluation of these developed PoCs include, as previously mentioned in our framework, ethical and regulatory benefits, technological advances, and AI for social good. The PoC implementation will adopt strategies to limit technical challenges associated with the development and integration of complex systems through modular architecture design and implementation that allows independent development and testing of various components that make up the system. The technology choices for the implementation will focus on blockchain stacks that have already demonstrated stable integration with external entities such as AI components and algorithms.

Future research should also explore further improvement options for societal resilience via the integration of blockchain technologies and AI interdisciplinary collaboration between technologists, ethicists, and policymakers. Such interdisciplinary study efforts aim to form a foundation for addressing emerging digital threats by ensuring that BIAI systems advance responsibly in ways that uphold human rights and promote social equity.

Author Contributions: Conceptualization, C.U., A.N. and R.V.-D.; methodology, C.U. and R.V.-D.; data curation, A.A.O., C.U. and R.V.-D.; writing—original draft preparation, C.U., A.N., R.V.-D., A.A.O. and N.S.; writing—review and editing, A.N., C.U., R.V.-D., A.A.O., N.S. and S.C.; project administration, C.U., R.V.-D., A.N. and S.C.; All authors have read and agreed to the published version of the manuscript.

Funding: This research is partially funded within the framework of the COMET center ABC, Austrian Blockchain Center, by BMK, BMAW, and the provinces of Vienna, Lower Austria, and Vorarlberg. The COMET program (Competence Centers for Excellent Technologies) is managed by the FFG. This research is also partially funded by the Estonian "Personal Research Funding: Team Grant (PRG)" project PRG1641, and the Estonian project MKM-POL21-2025.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ADM	Automated Decision-Making
AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy AI
AML	Anti-Money Laundering
BIAI	Blockchain-Integrated AI
BPMN	Business Process Modeling and Notation
CJEU	Court of Justice of the European Union
CPRA	California Privacy Rights Act
DAO	Decentralized Autonomous Organization
DCAP	Decentralized Conditional Anonymous Payment
DApp	Decentralized Application
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability
LLM	Large Language Model
ML	Machine Learning
NFT	Non-Fungible Token
NLP	Natural Language Processing
SLR	Systematic Literature Review
SVG	Scalable Vector Graphics
TPKS	Tindak Pidana Kekerasan Seksual (Indonesian law on sexual violence)
ZKP	Zero-Knowledge Proof

Appendix A

SLR detailed methods

To cover the topics proposed in the SLR, we used the following Boolean logic on the aforementioned databases:

("artificial intelligence" OR "machine learning" OR "AI" OR "LLM" OR "blockchain" OR "decentralized systems" OR "smart contracts")

AND

("sextortion" OR "digital sexual exploitation" OR "online coercion" OR "cyber abuse")

AND

("ethics" OR "privacy" OR "accountability" OR "governance" OR "transparency" OR "fairness" OR "bias" OR "trust" OR "digital trust" OR "AI for social good" OR "societal resilience" OR "social resilience" OR "AI ethics" OR "ethical AI" OR "AI governance").

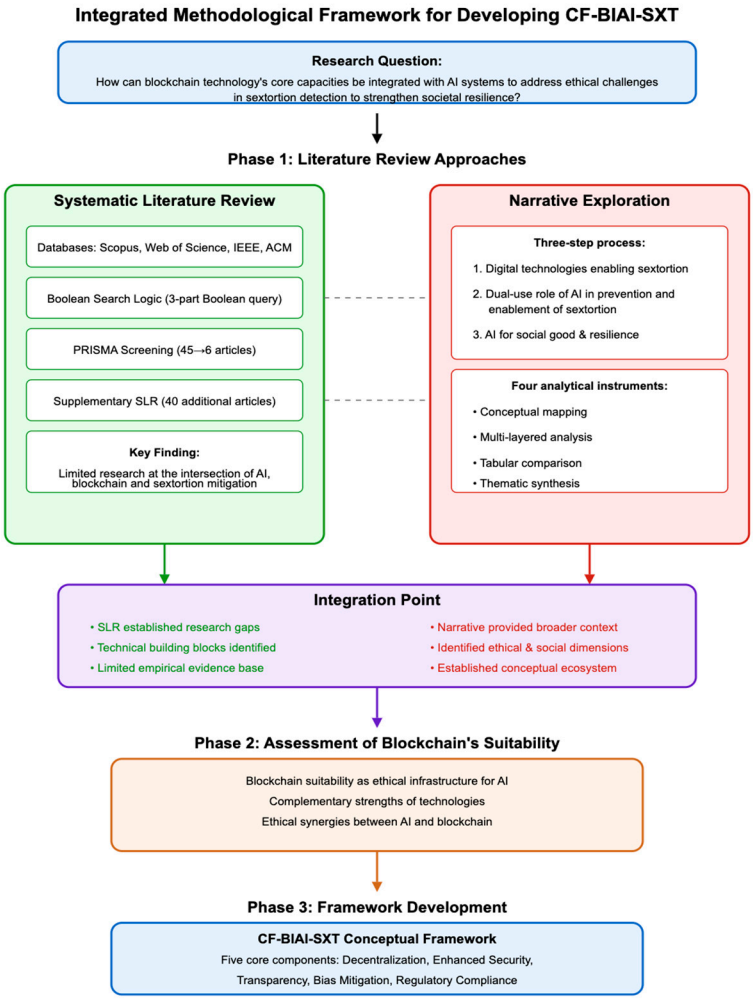

Figure A1. The Integrated Methodological Framework for developing CF-BIAI-SXT.

Inclusion criteria:

- Articles must focus on AI or algorithmic systems applied in contexts of digital abuse, especially sexual exploitation;
- Studies must engage with at least one ethical dimension (e.g., fairness, bias, data consent, transparency, accountability);
- Papers must be peer-reviewed and published in English.

Exclusion criteria:

- Purely technical or engineering-focused without reference to ethical, legal, or societal implications;
- Related to physical or non-digital forms of abuse;
- Gray literature or preprints.

We coded the articles in Rayyan based on the following:

- AI function (mitigation, surveillance, moderation, etc.),
- Ethical focus (e.g., bias, explainability, misuse of data),
- Governance model (centralized vs. decentralized),
- Framework alignment (AI4SG, resilience, digital trust).

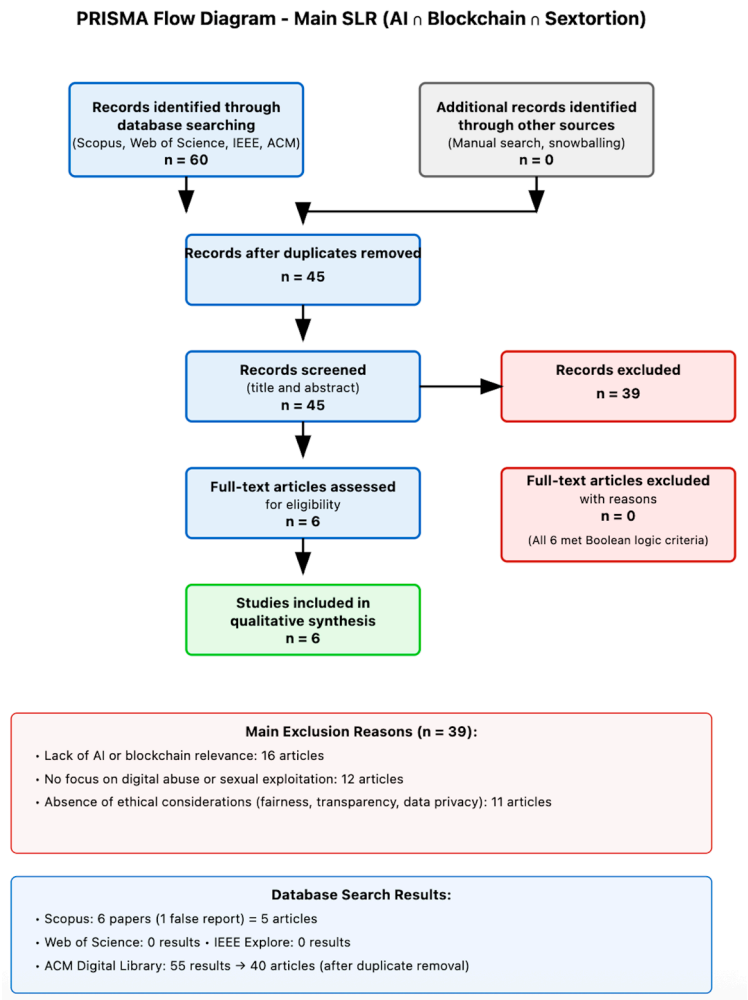

Figure A2. PRISMA flow diagram for the SLR part of the study for informing the CF-BIAI-SXT.

Table A1. SLR Coding sheet.

Field	Values
Study ID	FirstAuthor-Year
Published in	Journal/Conference/Chapter
Evidence quality *	**PR** (Peer-Reviewed Empirical or Technical)/**C** (Conceptual/Commentary)/**Qual** (Peer-Reviewed Qualitative Empirical)
AI link/Function	Detection/Mitigation/Moderation/Surveillance
Ethical focus	Privacy/Fairness/Transparency/Bias/Accountability
Governance model	Centralized/Consortium/DAO/None
Framework alignment	AI4SG (AI-for-Social-Good)/Digital-trust/Resilience/None
Gap mapped by study	Provenance/consent/Bias/accuracy/Opaque moderation/User control/explainability/Evidence integrity/Fairness drift
Blockchain capacity	Immutability/audit trail/Federated learning/Smart-contract audit/Data-wallet/DAO/Chain-of-custody/On-chain logging/None
Sextortion link/relevance	Direct/Related abuse/Conceptual
Methodology	Empirical/Prototype/Conceptual
Notes	Free text (e.g., dataset size)

* in lieu of formal scoring as explained in Section 3.1.

Table A2. Narrative Review Coding Sheet.

Field	Values
Study ID	FirstAuthor-Year
Source type	Journal/Conference/Policy-brief/News/Thesis/ArXiv
Peer-reviewed?	PR/NP
Region/scope	Country or Global
Core theme	Digital-abuse tech/AI dual-use and risk/Ethical-AI framework/Blockchain capacity/Societal-resilience
Sub-themes	Free list (e.g., Deepfakes; DAO governance; Survivor design)
Sextortion relevance	Direct/Related abuse/Conceptual
Ethical keyword	Privacy/Fairness/Transparency/Consent/Bias/Equity
Method type	Emp-quantitative/Emp-qualitative/Mixed/Conceptual
Other	Free notes

Table A3. Coding of the SLR seed 6.

Field						
Study ID	[14]	[15]	[16]	[17]	[18]	[19]
Published in	Journal	Conf.	Journal	Journal/Conf.	Conf.	Conf.
Evidence quality	C	PR	Qual	PR	C	PR
AI link/Function	— (commentary)	Detection	—	(risk-experience)	Mitigation	Detection
Ethical focus	Privacy; Transparency	Bias; Accuracy	Privacy	Consent; Privacy	Privacy; Accountability	Fairness; Explainability
Governance model	None	Centralized	None	None	Centralized	Centralized
Framework alignment	Digital-trust	Digital-trust	Resilience	AI4SG	Digital-trust	AI4SG
Gap mapped by study	Loss of provenance/consent in AI-generated nudes	Language-bias and accuracy drop in cyber-abuse detector	Victims distrust opaque moderation	Teens need data control and explainability	Need tamper-proof evidence for deepfake blackmail cases	Fairness drift in real-time risk models
Blockchain capacity	Immutability + audit trail for origin verification	Federated learning to enlarge/de-bias datasets	Smart-contract audit of moderation decisions	Data-wallet permissions; DAO youth governance	Blockchain chain-of-custody	On-chain model-performance logging
Sextortion link/relevance	Direct	Related abuse (cyber-harassment)	Related abuse (adult cyber-abuse)	Direct (adolescent sextortion)	Direct (deepfake blackmail)	Direct (adolescent sextortion-risk detection)
Methodology	Conceptual commentary	Empirical prototype	Empirical—qualitative	Empirical—mixed (survey + posts, $n \approx 195$)	Conceptual/scoping review	Prototype/design study (usability pilot, $n \approx 12$)
Notes	Essay on GenAI sextortion risks; no dataset	TRAC-1 multilingual dataset	Phenomenological interviews with K-12 teachers ($n \approx 25$)	Reddit/Tumblr data + teen survey; topic modeling	Reviews PH deepfake cases; outlines legal and technical counter-measures	Youth-centric co-design of risk-detection dashboard; early pilot

Appendix B

In this section, we provide the preliminaries for understanding the key essential technical concepts and frameworks that are foundational to this position paper. It is important to introduce in this way the core principles that drive the mentioned technological combination of blockchain technologies and AI.

Appendix B.1 Preliminaries—Technical Concepts—AI-Related Concepts

Both machine learning (ML) and LLM models are necessary to realize the AI application that addresses the use case of sextortion, which is the focus of this paper [157–160]:

ML and LLM Algorithms: ML algorithms are models trained to identify patterns in a dataset and provide predictions by classifying categories of values, predicting continuous values, or organizing data in entirely new clusters. Hence, ML algorithms can be categorized into classification, regression, and clustering algorithms. Yet, LLMs are algorithms that provide prediction by understanding patterns and relationships in an existing dataset to generate a completely new set of data in the form of natural language, coherent for human understanding. LLMs are a special type of natural language processor (NLP) realized from various deep learning algorithms focusing on generating human-readable content.

Model performance metrics: Evaluation of AI models facilitates an understanding of their accuracy, hence providing a layer of transparency to these algorithms that are often considered a black box. Similar metrics used in evaluating classification ML algorithms, such as accuracy score and F1 score, are also used in assessing the correctness of predictions produced by LLMs. Generally, the accuracy score checks the quantity of correct predictions, while the F1 score checks the quality of the model using properties such as model precision and recall. Additional metrics are incorporated specifically for LLMs to check the fluency, coherency, and relevance of predictions generated.

Algorithm execution steps: Both ML and LLM models follow similar execution processes involving data preparation, model training, and model execution. Data preparation involves all the steps of data pre-processing, such as data acquisition, aggregation, transformation, cleaning, normalization, etc., before the data is fed into a model for training. The models are then trained to identify patterns and relationships in a dataset to predict a result for a given set of data questions or to generate an entirely new set of data. The models are optimized to achieve a particular level of performance criteria and then deployed for execution in a real-life environment. One of the ways to optimize a model is by ensemble modeling, such that several (similar) models are separately trained and their results combined using a consensus algorithm to generate a final result.

Appendix B.2 Preliminaries—Technical Concepts—Blockchain-Related Concepts

For the blockchain-related concepts, first, we describe a typical blockchain network and the consensus algorithm that supports it. Then, we describe smart contracts that run on decentralized networks and token-based systems to exchange assets and values within a blockchain network.

Blockchain network and consensus mechanisms: The blockchain network comprises peers and nodes, that represent entities that execute transactions within the network. Transactions are organized in blocks, cryptographically linked with previous transactions, and are redundantly recorded across peers, ensuring consistency of the blockchain's state among all peers. Generally, blockchain networks fall into two categories: private and public. In private blockchains, permission is required to join the network, while in public blockchains, anyone can join and execute transactions on the network. Before any transaction is accepted into the network, it undergoes validation using a specified consensus method. Public blockchains commonly employ proof-based consensus methods,

such as proof of work and proof of stake, along with their variations. In contrast, private blockchains typically use voting-based consensus methods, often based on adaptations of Byzantine fault-tolerant systems.

Smart contracts and tokenization: The computer programs running on the blockchains are commonly known as smart contracts. Consequently, different rules and conditions can be encoded within a smart contract and are executed without the need to rely on a central entity for coordination. A blockchain application, also referred to as a decentralized application (DApp), can consist of several smart contracts. Smart contracts are also used to realize information assets and value exchange among the network participants. These values and digital assets can be represented in various types of tokens that exist in blockchain networks. Some common examples of tokens are utility and non-fungible tokens (NFTs). Utility tokens are fungible and are used to implement the ownership and transfer of values in blockchain networks. NFTs are commonly used to provide a unique representation of digital assets in a blockchain network.

Appendix C

Table A4. Technical Implementation Landscape for BIAI.

Dimension	Challenge/Requirement	Solution Approaches
Data Governance	Cross-border compliance, GDPR alignment	Federated ML, Data wallets, Consent logs
Architecture Choices	Public vs. Private chains, On-chain vs. Off-chain	Hybrid models, selective decentralization
Smart Contracts	Immutability vs. Upgradability, Vulnerabilities	Solidity audits, Best practices, Formal verification
Consensus Mechanisms	Trade-offs: speed, cost, energy, trust	PoW, PoS, BFT, context-aware selection
Interoperability	Fragmented ecosystems, siloed data	Standardized APIs, Cross-chain protocols
Scalability	Throughput, storage, compute bottlenecks	Layer-2s, Sharding, Off-chain solutions

References

1. Patchin, J.W.; Hinduja, S. Sextortion Among Adolescents: Results from a National Survey of U.S. Youth. *Sex. Abus.* **2018**, *32*, 30–54. [CrossRef] [PubMed]
2. The Korea Times. AI and Deepfake Technology Fuel Surge in Digital Sex Crimes in South Korea. 2025. Available online: https://www.koreatimes.co.kr/southkorea/law-crime/20250410/ai-and-deepfake-technology-fuel-surge-in-digital-sex-crimes-in-south-korea (accessed on 12 April 2025).
3. Norta, A.; Makrygiannis, S. Designing Artificial Intelligence Equipped Social Decentralized Autonomous Organizations for Tackling Sextortion Cases Version 0.7. *arXiv* **2023**, arXiv:2312.14090.
4. Okolie, C. Artificial Intelligence-Altered Videos (Deepfakes), Image-Based Sexual Abuse, and Data Privacy Concerns. *J. Int. Women's Stud.* **2023**, *25*, 11. Available online: https://vc.bridgew.edu/jiws/vol25/iss2/11/ (accessed on 12 April 2025).
5. von Eschenbach, W.J. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philos. Technol.* **2021**, *34*, 1607–1622. [CrossRef]
6. Sukhera, J. Narrative Reviews: Flexible, Rigorous, and Practical. *J. Grad. Med. Educ.* **2022**, *14*, 414–417. [CrossRef]
7. Oliver, S.; Harden, A.; Rees, R.; Shepherd, J.; Brunton, G.; Garcia, J.; Oakley, A. An Emerging Framework for Including Different Types of Evidence in Systematic Reviews for Public Policy. *Evaluation* **2005**, *11*, 428–446. [CrossRef]
8. Heeager, L.T.; Nielsen, P.A. A Conceptual Model of Agile Software Development in a Safety-Critical Context: A Systematic Literature Review. *Inf. Softw. Technol.* **2018**, *103*, 22–39. [CrossRef]
9. Abdel-Aty, T.A.; Negri, E. Conceptualizing the Digital Thread for Smart Manufacturing: A Systematic Literature Review. *J. Intell. Manuf.* **2024**, *35*, 3629–3653. [CrossRef]
10. Naeem, M.; Ozuem, W.; Howell, K.; Ranfagni, S. A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research. *Int. J. Qual. Methods* **2023**, *22*, 16094069231205789. [CrossRef]
11. Greenhalgh, T.; Thorne, S.; Malterud, K. Time to Challenge the Spurious Hierarchy of Systematic over Narrative Reviews? *Eur. J. Clin. Investig.* **2018**, *48*, e12931. [CrossRef]

12. Recker, J.; Indulska, M.; Michael; Green, P. How Good Is BPMN Really? Insights from Theory and Practice. In Proceedings of the 14th European Conference on Information Systems (ECIS 2006), Göteborg, Sweden, 12–14 June 2006; p. 135. Available online: https://aisel.aisnet.org/ecis2006/135 (accessed on 12 April 2025).
13. Karagöz, N.A.; Demirörs, O. Conceptual Modeling Notations and Techniques. In *Conceptual Modeling for Discrete-Event Simulation*; Robinson, S., Brooks, R.J., Kotiadis, K., Van der Zee, D.-J., Eds.; CRC Press: Boca Raton, FL, USA, 2010; pp. 195–226.
14. Pater, J.; McDaniel, B.T.; Nova, F.F.; Drouin, M.; O'Connor, K.; Zytko, D. A Commentary on Sexting, Sextortion, and Generative AI: Risks, Deception, and Digital Vulnerability. *Fam. Relat.* **2025**, *74*, 1109–1120. [CrossRef]
15. Malte, A.; Ratadiya, P. Multilingual Cyber Abuse Detection Using Advanced Transformer Architecture. In Proceedings of the 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 17–20 October 2019; pp. 784–789. [CrossRef]
16. Arantes, J. It's Too Late–The Post Has Gone Viral Already: A Novel Methodological Stance to Explore K–12 Teachers' Lived Experiences of Adult Cyber Abuse. *Qual. Res. J.* **2023**. [CrossRef]
17. Alsoubai, A.; Song, J.; Razi, A.; Naher, N.; De Choudhury, M.; Wisniewski, P.J. From 'Friends with Benefits' to 'Sextortion:' A Nuanced Investigation of Adolescents' Online Sexual Risk Experiences. *Proc. ACM Hum.-Comput. Interact.* **2022**, *6*, 1–32. [CrossRef]
18. Blancaflor, E.; Garcia, J.I.; Magno, F.D.; Vilar, M.J. Deepfake Blackmailing on the Rise: The Burgeoning Posterity of Revenge Pornography in the Philippines. In Proceedings of the 2024 9th International Conference on Intelligent Information Technology (ICIIT 2024), Tokyo, Japan, 24–26 February 2024; pp. 295–301. [CrossRef]
19. Alsoubai, A. A Human-Centered Approach to Improving Adolescent Real-Time Online Risk Detection Algorithms. In Proceedings of the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), Hamburg, Germany, 23–28 April 2023; Article 480, pp. 1–5. [CrossRef]
20. Li, Z.; Kong, D.; Niu, Y.; Peng, H.; Li, X.; Li, W. An Overview of AI and Blockchain Integration for Privacy-Preserving. *arXiv* **2023**, arXiv:2305.03928. [CrossRef]
21. Shanmugam, L.; Tillu, R.; Jangoan, S. Privacy-Preserving AI/ML Application Architectures: Techniques, Trade-Offs, and Case Studies. *J. Knowl. Learn. Sci. Technol.* **2023**, *2*, 398–420. [CrossRef]
22. Preuveneers, D.; Rimmer, V.; Tsingenopoulos, I.; Spooren, J.; Joosen, W.; Ilie-Zudor, E. Chained Anomaly Detection Models for Federated Learning: An Intrusion Detection Case Study. *Appl. Sci.* **2018**, *8*, 2663. [CrossRef]
23. Liu, Y.; Lu, Q.; Zhu, L.; Paik, H.Y. Decentralised Governance-Driven Architecture for Designing Foundation Model Based Systems: Exploring the Role of Blockchain in Responsible AI. *arXiv* **2023**, arXiv:2308.05962. [CrossRef]
24. Nassar, M.; Salah, K.; Ur Rehman, M.H.; Svetinovic, D. Blockchain for Explainable and Trustworthy Artificial Intelligence. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1340. [CrossRef]
25. Calvaresi, D.; Calbimonte, J.P.; Dubovitskaya, A.; Mattioli, V.; Piguet, J.G.; Schumacher, M. The Good, the Bad, and the Ethical Implications of Bridging Blockchain and Multi-Agent Systems. *Information* **2019**, *10*, 363. [CrossRef]
26. Nawaz, A.; Gia, T.N.; Queralta, J.P.; Westerlund, T. Edge AI and Blockchain for Privacy-Critical and Data-Sensitive Applications. In Proceedings of the 2019 Twelfth International Conference on Mobile Computing and Ubiquitous Network (ICMU), Kathmandu, Nepal, 4–6 November 2019; pp. 1–2. [CrossRef]
27. Khan, B.; Goh, K.; Khan, A.R.; Zuhairi, M.; Chaimanee, M. Integrating AI and Blockchain for Enhanced Data Security in IoT-Driven Smart Cities. *Processes* **2024**, *12*, 1825. [CrossRef]
28. Kalenzi, C. Artificial Intelligence and Blockchain: How Should Emerging Technologies Be Governed? *Front. Res. Metr. Anal.* **2022**, *7*, 801549. [CrossRef]
29. Manias, G.; Apostolopoulos, D.; Athanassopoulos, S.; Borotis, S.A.; Chatzimallis, C.; Chatzipantelis, T.; Corrales Compagnucci, M.; Draksler, T.Z.; Fournier, F.; Goralczyk, M.; et al. AI4Gov: Trusted AI for Transparent Public Governance Fostering Democratic Values. In Proceedings of the 2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), Coral Gables, FL, USA, 19–21 June 2023. [CrossRef]
30. Asif, R.; Hassan, S.R.; Parr, G. Integrating a Blockchain-Based Governance Framework for Responsible AI. *Future Internet* **2023**, *15*, 97. [CrossRef]
31. Calvaresi, D.; Mualla, Y.; Najjar, A.; Galland, S.; Schumacher, M. Explainable Multi-Agent Systems Through Blockchain Technology. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems. EXTRAAMAS 2019*; Calvaresi, D., Najjar, A., Schumacher, M., Främling, K., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; Volume 11763. [CrossRef]
32. Padmanaban, H. Privacy-Preserving Architectures for AI/ML Applications: Methods, Balances, and Illustrations. *J. Artif. Intell. Gen. Sci.* **2024**, *3*, 235–245. [CrossRef]
33. Renuka, G.B.; Patjoshi, P.K.; Aswal, U.; Manikandan, G.; Jayanthi, L.N.; Kaushal, A. Integrating Reliable AI to Boost Blockchain's Transparency and Accountability. In Proceedings of the 2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET), Rourkela, India, 2–4 August 2024; pp. 1–6. [CrossRef]
34. Chowdhury, R.H. Blockchain and AI: Driving the Future of Data Security and Business Intelligence. *World J. Adv. Res. Rev.* **2024**, *23*, 2559–2570. [CrossRef]

35. Henry, N.; Umbach, R. Sextortion: Prevalence and Correlates in 10 Countries. *Comput. Hum. Behav.* **2024**, *158*, 108298. [CrossRef]
36. Carlton, A. Sextortion: The Hybrid Cyber-Sex Crime. *N. Carol. J. Law Technol.* **2019**, *21*, 177.
37. Federal Bureau of Investigation (FBI). Sextortion: A Growing Threat Targeting Minors. 2023. Available online: https://www.fbi.gov/contact-us/field-offices/nashville/news/sextortion-a-growing-threat-targeting-minors (accessed on 12 April 2025).
38. Sheikh, M.M.R.; Rogers, M.M. Technology-Facilitated Sexual Violence and Abuse in Low and Middle-Income Countries: A Scoping Review. *Trauma Violence Abus.* **2024**, *25*, 1614–1629. [CrossRef]
39. Ray, A.; Henry, N. Sextortion: A Scoping Review. *Trauma Violence Abus.* **2025**, *26*, 138–155. [CrossRef]
40. Paat, Y.F.; Markham, C. Digital Crime, Trauma, and Abuse: Internet Safety and Cyber Risks for Adolescents and Emerging Adults in the 21st Century. *Soc. Work Ment. Health* **2021**, *19*, 18–40. [CrossRef]
41. Sorbán, K. An Elephant in the Room—EU Policy Gaps in the Regulation of Moderating Illegal Sexual Content on Video-Sharing Platforms. *Int. J. Law Inf. Technol.* **2023**, *31*, 171–185. [CrossRef]
42. Küpeli, C. Legal Analysis of Sextortion Crime in the Comparative Law and Turkish Law. *Health Sci. Q.* **2019**, *3*, 87–98. [CrossRef]
43. Makinde, O.A.; Olamijuwon, E.; Ichegbo, N.K.; Onyemelukwe, C.; Ilesanmi, M.G. The Nature of Technology-Facilitated Violence and Abuse among Young Adults in Sub-Saharan Africa. In *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*; Bailey, C., Flynn, A., Powell, A., Henry, N., Eds.; Emerald Publishing: Bingley, UK, 2021; pp. 83–101. [CrossRef]
44. Lyttle Storrod, M. "It Started with a Nude": Gangs & Technology-Facilitated Sexual Violence. *Int. Crim. Justice Rev.* **2024**, *34*, 245–261. [CrossRef]
45. Sunde, N.; Sunde, I.M. Conceptualizing an AI-Based Police Robot for Preventing Online Child Sexual Exploitation and Abuse: Part I—The Theoretical and Technical Foundations for PrevBOT. *Nord. J. Stud. Polic.* **2021**, *8*, 1–21. [CrossRef]
46. Transparency International. Sextortion: Undermining Gender Equality. 2020. Available online: http://ti.or.id/publikasi/sextortion/ENG_briefing_paper_sextortion.pdf (accessed on 12 April 2025).
47. Fletcher, R.; Tzani, C.; Ioannou, M. The Dark Side of Artificial Intelligence–Risks Arising in Dating Applications. *Assess. Dev. Matters* **2024**, *16*, 17–23. [CrossRef]
48. Futcher, L.; Thomson, K.L.; Kucherera, L.; Gcaza, N. Key Elements for Cybersafety Education of Primary School Learners in South Africa. In Proceedings of the International Symposium on Human Aspects of Information Security and Assurance (HAISA 2023), Kent, UK, 4–6 July 2023; Springer Nature: Cham, Switzerland, 2023; pp. 116–128. [CrossRef]
49. Ahakonye, L.A.C.; Nwakanma, C.I.; Kim, D.S. Tides of Blockchain in IoT Cybersecurity. *Sensors* **2024**, *24*, 3111. [CrossRef]
50. Chellappan, S.; Fisk, N. A Novel Privacy-Preserving Socio-Technical Platform for Detecting Cyber Abuse. In *Computational Data and Social Networks. CSoNet 2019*; Nguyen, N.T., Chbeir, R., Alhajj, R., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; Volume 11920, pp. 307–308. [CrossRef]
51. Phippen, A.; Bond, E. Image Recognition in Child Sexual Exploitation Material—Capabilities, Ethics and Rights. In *Policing in the Era of AI and Smart Societies*; Leong, C., Shanahan, M., Eds.; Springer: Singapore, 2020; pp. 179–198.
52. Borau, S. Deception, Discrimination, and Objectification: Ethical Issues of Female AI Agents. *J. Bus. Ethics* **2024**, *198*, 1–19. [CrossRef]
53. Ali, M.I. Strategies for Intervention and Prevention in Online Child Sexual Exploitation. *Teisė* **2024**, *132*, 145–155. [CrossRef]
54. Kumar, P.; Javeed, D.; Kumar, R.; Islam, A.N. Blockchain and Explainable AI for Enhanced Decision Making in Cyber Threat Detection. *Softw. Pract. Exp.* **2024**, *54*, 1337–1360. [CrossRef]
55. Wei, C. Enhancing AI Security: A Review of Blockchain-Integrated Solutions. In Proceedings of the 2024 4th International Conference on Computer Science and Blockchain (CCSB), Xi'an, China, 6–8 September 2024; pp. 561–569. [CrossRef]
56. Stockhem, O. Improving the International Regulation of Cybersex Trafficking of Women and Children Through the Use of Data Science and Artificial Intelligence. Ph.D. Thesis, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 2020. [CrossRef]
57. Demertzis, K.; Rantos, K.; Magafas, L.; Skianis, C.; Iliadis, L. A Secure and Privacy-Preserving Blockchain-Based XAI-Justice System. *Information* **2023**, *14*, 477. [CrossRef]
58. Chatziamanetoglou, D.; Rantos, K. Cyber Threat Intelligence on Blockchain: A Systematic Literature Review. *Computers* **2024**, *13*, 60. [CrossRef]
59. Saxena, R.; Gayathri, E.; Surya Kumari, L. Semantic Analysis of Blockchain Intelligence with Proposed Agenda for Future Issues. *Int. J. Syst. Assur. Eng. Manag.* **2023**, *14* (Suppl. S1), 34–54. [CrossRef]
60. Saputra, D.H.; Prakarsa, A. Enhancing Detection Mechanisms: Law Enforcement Strategies Identifying Suspected Financial Transactions of Child Sexual Exploitation Crimes. In *Proceedings of the ASEAN Conference on Sexual Exploitation of Children (ACOSEC)*; Atlantis Press: Dordrecht, The Netherlands, 2024; pp. 120–126. [CrossRef]
61. Rita, M.N.; Shava, F.B. Chatbot-Driven Web-Based Platform for Online Safety and Sexual Exploitation Awareness and Reporting in Namibia. In Proceedings of the 2021 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 5–6 August 2021; pp. 1–5. [CrossRef]
62. Saleh, A.M.S. Blockchain for Secure and Decentralized Artificial Intelligence in Cybersecurity: A Comprehensive Review. *Blockchain Res. Appl.* **2024**, *5*, 100193. [CrossRef]

63. Chavali, B.; Khatri, S.K.; Hossain, S.A. AI and Blockchain Integration. In Proceedings of the 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 4–5 June 2020; pp. 548–552. [CrossRef]
64. Ayissi, B.D.; Befoum, S.R.; Kombou, V. AI-Driven Blockchain: A Review of Pathways to Self-Sovereign Intelligence. *SSRN* **2023**. [CrossRef]
65. Jain, V.; Chouhan, S.; Kate, V.; Nigam, N.; Bhalerao, S. Enhancing Data Security and Data Sensitivity: Leveraging the Synergy of Blockchain Artificial Intelligence. In Proceedings of the 2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG), Indore, India, 8–9 December 2023; pp. 1–8. [CrossRef]
66. Ural, O.; Yoshigoe, K. Survey on Blockchain-Enhanced Machine Learning. *IEEE Access* **2023**, *11*, 145331–145362. [CrossRef]
67. Brewer, J.; Patel, D.; Kim, D.; Murray, A. Navigating the Challenges of Generative Technologies: Proposing the Integration of Artificial Intelligence and Blockchain. *Bus. Horiz.* **2024**, *67*, 525–535. [CrossRef]
68. Dillenberger, D.N.; Novotny, P.; Zhang, Q.; Jayachandran, P.; Gupta, H.; Hans, S.; Sarpatwar, K. Blockchain Analytics and Artificial Intelligence. *IBM J. Res. Dev.* **2019**, *63*, 1–14. [CrossRef]
69. Kumar, R.; Arjunaditya; Singh, D.; Srinivasan, K.; Hu, Y.C. AI-Powered Blockchain Technology for Public Health: A Contemporary Review, Open Challenges, and Future Research Directions. *Healthcare* **2022**, *11*, 81. [CrossRef]
70. Salama, R.; Al-Turjman, F. AI in Blockchain towards Realizing Cyber Security. In Proceedings of the 2022 International Conference on Artificial Intelligence in Everything (AIE), Montreal, QC, Canada, 2–4 August 2022; pp. 471–475. [CrossRef]
71. Joshi, M.L.; Kanoongo, N. Depression Detection Using Emotional Artificial Intelligence and Machine Learning: A Closer Review. *Mater. Today Proc.* **2022**, *58*, 217–226. [CrossRef]
72. Kaywan, P.; Ahmed, K.; Ibaida, A.; Miao, Y.; Gu, B. Early Detection of Depression Using a Conversational AI Bot: A Non-Clinical Trial. *PLoS ONE* **2023**, *18*, e0279743. [CrossRef]
73. Martins, R.; Almeida, J.J.; Henriques, P.R.; Novais, P. Identifying Depression Clues Using Emotions and AI. In Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021), Online Streaming, 4–6 February 2021; Volume 2, pp. 1137–1143. Available online: https://www.scitepress.org/PublishedPapers/2021/103328/103328.pdf (accessed on 12 April 2025).
74. Benefo, E.O.; Tingler, A.; White, M.; Cover, J.; Torres, L.; Broussard, C.; Shirmohammadi, A.; Pradhan, A.K.; Patra, D. Ethical, Legal, Social, and Economic (ELSE) Implications of Artificial Intelligence at a Global Level: A Scientometrics Approach. *AI Ethics* **2022**, *2*, 667–682. [CrossRef]
75. Acemoglu, D. *Harms of AI*; NBER Working Paper No. w29247; National Bureau of Economic Research: Cambridge, MA, USA, 2021. [CrossRef]
76. Garcia, P.; Darroch, F.; West, L.; Brooks-Cleator, L. Ethical Applications of Big Data-Driven AI on Social Systems: Literature Analysis and Example Deployment Use Case. *Information* **2020**, *11*, 235. [CrossRef]
77. Dignum, V. Responsibility and Artificial Intelligence. In *The Oxford Handbook of Ethics of AI*; Dubber, M.D., Pasquale, F., Das, S., Eds.; Oxford University Press: Oxford, UK, 2020; Volume 4698, p. 215.
78. Kamila, M.K.; Jasrotia, S.S. Ethical Issues in the Development of Artificial Intelligence: Recognizing the Risks. *Int. J. Ethics Syst.* **2025**, *41*, 45–63. [CrossRef]
79. Jacobs, J. The Artificial Intelligence Shock and Socio-Political Polarization. *Technol. Forecast. Soc. Change* **2024**, *199*, 123006. [CrossRef]
80. Hagerty, A.; Rubinov, I. Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence. *arXiv* **2019**, arXiv:1907.07892. [CrossRef]
81. Wamba, S.F.; Bawack, R.E.; Guthrie, C.; Queiroz, M.M.; Carillo, K.D.A. Are We Preparing for a Good AI Society? A Bibliometric Review and Research Agenda. *Technol. Forecast. Soc. Change* **2021**, *164*, 120482. [CrossRef]
82. Ashok, M.; Madan, R.; Joha, A.; Sivarajah, U. Ethical Framework for Artificial Intelligence and Digital Technologies. *Int. J. Inf. Manag.* **2022**, *62*, 102433. [CrossRef]
83. Cote, M.; Nightingale, A.J. Resilience Thinking Meets Social Theory: Situating Social Change in Socio-Ecological Systems (SES) Research. *Prog. Hum. Geogr.* **2012**, *36*, 475–489. [CrossRef]
84. Kou, C.; Yang, X. Improving Social Resilience amid the COVID-19 Epidemic: A System Dynamics Model. *PLoS ONE* **2023**, *18*, e0294108. [CrossRef]
85. Keck, M.; Sakdapolrak, P. What Is Social Resilience? Lessons Learned and Ways Forward. *Erdkunde* **2013**, *67*, 5–19. Available online: https://www.jstor.org/stable/23595352 (accessed on 12 April 2025). [CrossRef]
86. Hu, Q.; Lu, Y.; Pan, Z.; Wang, B. How Does AI Use Drive Individual Digital Resilience? A Conservation of Resources (COR) Theory Perspective. *Behav. Inf. Technol.* **2023**, *42*, 2654–2673. [CrossRef]
87. Moya Velasco, J.; Goenechea Domínguez, M. Variables Involved in the Development of Social Resilience in Sustainable Economies: From Individual Resilience to Societies' Resilience. *Vis. Rev. Int. Vis. Cult. Rev.* **2022**, *12*, 1–15. [CrossRef]

88. Cody, T.; Beling, P.A. Towards Operational Resilience for AI-Based Cyber in Multi-Domain Operations. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications V*; SPIE: Bellingham, WA, USA, 2023; pp. 368–373. [CrossRef]
89. Moskalenko, V.; Kharchenko, V.; Moskalenko, A.; Kuzikov, B. Resilience and Resilient Systems of Artificial Intelligence: Taxonomy, Models and Methods. *Algorithms* **2023**, *16*, 165. [CrossRef]
90. Adomavicius, G.; Bockstedt, J.C.; Gupta, A.; Kauffman, R.J. Technology Roles and Paths of Influence in an Ecosystem Model of Technology Evolution. *Inf. Technol. Manag.* **2007**, *8*, 185–202. [CrossRef]
91. Brockman, B.; Hersh, S.; Hoyer Gosselink, B.; Maganza, F.; Berman, M. Investing in AI for Good. *Stanf. Soc. Innov. Rev.* **2021**. Available online: https://ssir.org/articles/entry/investing_in_ai_for_good (accessed on 13 April 2025).
92. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Vayena, E. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef]
93. Jobin, A.; Ienca, M.; Vayena, E. The Global Landscape of AI Ethics Guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
94. Whittlestone, J.; Nyrup, R.; Alexandrova, A.; Dihal, K.; Cave, S. *Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research*; Nuffield Foundation: London, UK, 2019; pp. 1–59.
95. Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The Ethics of Algorithms: Mapping the Debate. *Big Data Soc.* **2016**, *3*, 205395. [CrossRef]
96. Floridi, L.; Cowls, J.; King, T.C.; Taddeo, M. How to Design AI for Social Good: Seven Essential Factors. In *Ethics, Governance, and Policies in Artificial Intelligence*; Floridi, L., Ed.; Philosophical Studies Series; Springer: Cham, Switzerland, 2021; Volume 144. [CrossRef]
97. Cowls, J.; Tsamados, A.; Taddeo, M.; Floridi, L. A Definition, Benchmark and Database of AI for Social Good Initiatives. *Nat. Mach. Intell.* **2021**, *3*, 111–115. [CrossRef]
98. Floridi, L.; Cowls, J. A Unified Framework of Five Principles for AI in Society. In *Machine Learning and the City: Applications in Architecture and Urban Design*; Wiley: Hoboken, NJ, USA, 2022; pp. 535–545. [CrossRef]
99. Umbrello, S.; van de Poel, I. Mapping Value Sensitive Design onto AI for Social Good Principles. *AI Ethics* **2021**, *1*, 283–296. [CrossRef]
100. Cheng, L.; Varshney, K.R.; Liu, H. Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. *J. Artif. Intell. Res.* **2021**, *71*, 1137–1181. [CrossRef]
101. Berendt, B. AI for the Common Good?! Pitfalls, Challenges, and Ethics Pen-Testing. *Paladyn J. Behav. Robot.* **2019**, *10*, 44–65. [CrossRef]
102. OECD. *OECD Framework for the Classification of AI Systems*; OECD Publishing: Paris, France, 2022; Available online: https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html (accessed on 14 April 2025).
103. Vesnic-Alujevic, L.; Nascimento, S.; Polvora, A. Societal and Ethical Impacts of Artificial Intelligence: Critical Notes on European Policy Frameworks. *Telecomm. Policy* **2020**, *44*, 101961. [CrossRef]
104. Laux, J.; Wachter, S.; Mittelstadt, B. Trustworthy Artificial Intelligence and the European Union AI Act: On the Conflation of Trustworthiness and Acceptability of Risk. *Regul. Gov.* **2024**, *18*, 3–32. [CrossRef] [PubMed]
105. Neuwirth, R.J. Prohibited Artificial Intelligence Practices in the Proposed EU Artificial Intelligence Act (AIA). *Comput. Law Secur. Rev.* **2023**, *51*, 105798. [CrossRef]
106. High-Level Expert Group on Artificial Intelligence. *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment*; European Commission: Brussels, Belgium, 2020; Available online: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (accessed on 14 April 2025).
107. Stahl, B.C.; Leach, T. Assessing the Ethical and Social Concerns of Artificial Intelligence in Neuroinformatics Research: An Empirical Test of the European Union Assessment List for Trustworthy AI (ALTAI). *AI Ethics* **2023**, *3*, 745–767. [CrossRef]
108. European Commission. *Legal Framework of EU Data Protection*; European Commission: Brussels, Belgium, 2022; Available online: https://commission.europa.eu/law/law-topic/data-protection/legal-framework-eu-data-protection_en (accessed on 14 April 2025).
109. Chaturvedi, A. Defining Legal Responsibility in the Age of AI: Addressing Gaps in Data Privacy Regulation. *Indian J. Integr. Res. Law* **2023**, *3*, 1. Available online: https://heinonline.org/HOL/LandingPage?handle=hein.journals/injloitd4&div=128&id=&page= (accessed on 14 April 2025).
110. Bondi, E.; Xu, L.; Acosta-Navas, D.; Killian, J.A. Envisioning Communities: A Participatory Approach towards AI for Social Good. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, 19–21 July 2021; pp. 425–436. [CrossRef]
111. Ressi, D.; Romanello, R.; Piazza, C.; Rossi, S. AI-Enhanced Blockchain Technology: A Review of Advancements and Opportunities. *J. Netw. Comput. Appl.* **2024**, *224*, 103858. [CrossRef]
112. Alzoubi, M.M. Investigating the Synergy of Blockchain and AI: Enhancing Security, Efficiency, and Transparency. *J. Cyber Secur. Technol.* **2024**. [CrossRef]

113. Kuznetsov, O.; Sernani, P.; Romeo, L.; Frontoni, E.; Mancini, A. On the Integration of Artificial Intelligence and Blockchain Technology: A Perspective About Security. *IEEE Access* **2024**, *12*, 3881–3897. [CrossRef]
114. Fadi, O.; Zkik, K.; El Ghazi, A.; Boulmalf, M. A Survey on Blockchain and Artificial Intelligence Technologies for Enhancing Security and Privacy in Smart Environments. *IEEE Access* **2022**, *10*, 93168–93186. [CrossRef]
115. Malik, V.; Mittal, R.; Mavaluru, D.; Narapureddy, B.; Goyal, S.B.; Martin, R.J.; Srinivasan, K.; Mittal, A. Building a Secure Platform for Digital Governance Interoperability and Data Exchange Using Blockchain and Deep Learning-Based Frameworks. *IEEE Access* **2023**, *11*, 70110–70131. [CrossRef]
116. Sonthi, V.K.; Nagarajan, S.; Murali Krishna, M.V.B.; Giridhar, K.; Lalitha, V.; Mohan, V. Imminent Threat with Authentication Methods for AI Data Using Blockchain Security. In *Blockchain Security in Cloud Computing*; Springer: Cham, Switzerland, 2021. [CrossRef]
117. Wylde, V.; Rawindaran, N.; Lawrence, J.; Balasubramanian, R.; Prakash, E.; Jayal, A.; Khan, I.A.; Hewage, C.; Platts, J. Cybersecurity, Data Privacy and Blockchain: A Review. *SN Comput. Sci.* **2022**, *3*, 160. [CrossRef]
118. Zhang, C.; Wu, C.; Wang, X. Overview of Blockchain Consensus Mechanism. In Proceedings of the 2020 2nd International Conference on Big Data Engineering (BDE '20), Shanghai, China, 17–19 June 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 7–12. [CrossRef]
119. Zhang, C.; Wang, R.; Tsai, W.T.; He, J.; Liu, C.; Li, Q. Actor-Based Model for Concurrent Byzantine Fault-Tolerant Algorithm. In Proceedings of the 2019 International Conference on Computer, Network, Communication and Information Systems (CNCI 2019), Wuhan, China, 25–26 May 2019; Atlantis Press: Paris, France, 2019; pp. 552–558. [CrossRef]
120. Gérin, B.; Halin, A.; Cioppa, A.; Henry, M.; Ghanem, B.; Macq, B.; De Vleeschouwer, C.; Van Droogenbroeck, M. Multi-Stream Cellular Test-Time Adaptation of Real-Time Models Evolving in Dynamic Environments. *arXiv* **2024**, arXiv:2404.17930. [CrossRef]
121. Sachan, S.; Yang, J.B.; Xu, D.L.; Benavides, D.E.; Li, Y. An Explainable AI Decision-Support-System to Automate Loan Underwriting. *Expert Syst. Appl.* **2020**, *144*, 113100. [CrossRef]
122. Marevac, E.; Patković, S.; Žunić, E. Decision-Making AI for Customer Worthiness and Viability. In Proceedings of the 2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 15–17 March 2023; pp. 1–6. [CrossRef]
123. Kannan, K.; Singh, A.; Verma, M.; Jayachandran, P.; Mehta, S. Blockchain-Based Platform for Trusted Collaborations on Data and AI Models. In Proceedings of the 2020 IEEE International Conference on Blockchain (Blockchain), Rhodes Island, Greece, 2–6 November 2020; pp. 82–89. [CrossRef]
124. Salah, K.; Rehman, M.H.U.; Nizamuddin, N.; Al-Fuqaha, A. Blockchain for AI: Review and Open Research Challenges. *IEEE Access* **2019**, *7*, 10127–10149. [CrossRef]
125. Loreti, P.; Bracciale, L.; Raso, E.; Bianchi, G.; Sanseverino, E.R.; Gallo, P. Privacy and Transparency in Blockchain-Based Smart Grid Operations. *IEEE Access* **2023**, *11*, 120666–120679. [CrossRef]
126. Lee, G.H.; Shin, S.Y. Federated Learning on Clinical Benchmark Data: Performance Assessment. *J. Med. Internet Res.* **2020**, *22*, e20891. [CrossRef] [PubMed]
127. Norta, A.; Hawthorne, D.; Engel, S.L. A Privacy-Protecting Data-Exchange Wallet with Ownership-and Monetization Capabilities. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8. [CrossRef]
128. Butt, A.; Junejo, A.Z.; Ghulamani, S.; Mahdi, G.; Shah, A.; Khan, D. Deploying Blockchains to Simplify AI Algorithm Auditing. In Proceedings of the 2023 IEEE 8th International Conference on Engineering Technologies and Applied Sciences (ICETAS), Karachi, Pakistan, 13–14 October 2023; pp. 1–6. [CrossRef]
129. Chaffer, T.J.; von Goins, C., II; Cotlage, D.; Okusanya, B.; Goldston, J. Decentralized Governance of AI Agents. 2024. Available online: https://www.researchgate.net/profile/Tomer-Chaffer/publication/387350593_Decentralized_Governance_of_AI_Agents/links/67918baf75d4ab477e580447/Decentralized-Governance-of-AI-Agents.pdf (accessed on 14 April 2025).
130. Singh, N.; Dayama, P.; Pandit, V. Zero Knowledge Proofs towards Verifiable Decentralized AI Pipelines. In *Financial Cryptography and Data Security*; Springer International Publishing: Cham, Switzerland, 2022; pp. 248–275. [CrossRef]
131. Balta, D.; Sellami, M.; Kuhn, P.; Schöpp, U.; Buchinger, M.; Baracaldo, N.; Anwar, A.; Ludwig, H.; Sinn, M.; Purcell, M.; et al. Accountable Federated Machine Learning in Government: Engineering and Management Insights. In *Electronic Participation: Proceedings of the 13th IFIP WG 8.5 International Conference, ePart 2021, Granada, Spain, 7–9 September 2021*; Springer International Publishing: Cham, Switzerland, 2021; Volume 13, pp. 125–138. [CrossRef]
132. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–19. [CrossRef]
133. Tatar, U.; Gokce, Y.; Nussbaum, B. Law versus technology: Blockchain, GDPR, and tough tradeoffs. *Comput. Law Secur. Rev.* **2020**, *38*, 105454. [CrossRef]
134. Xing, Z.; Zhang, Z.; Li, M.; Liu, J.; Zhu, L.; Russello, G.; Asghar, M. Zero-Knowledge Proof-based Practical Federated Learning on Blockchain. *arXiv* **2023**, arXiv:2304.05590. [CrossRef]

135. Ogungbemi, O. Smart Contracts Management: The Interplay of Data Privacy and Blockchain for Secure and Efficient Real Estate Transactions. *J. Eng. Res. Rep.* **2024**, *26*, 278–300. [CrossRef]
136. Keshavarzkalhori, G.; Pérez-Solá, C.; Navarro-Arribas, G.; Herrera-Joancomartí, J.; Yajam, H. Federify: A Verifiable Federated Learning Scheme Based on zkSNARKs and Blockchain. *IEEE Access* **2024**, *12*, 3240–3255. [CrossRef]
137. Xuan, T.R.; Ness, S. Integration of Blockchain and AI: Exploring application in the digital business. *J. Eng. Res. Rep.* **2023**, *25*, 20–39. Available online: https://www.academia.edu/download/106010959/1898.pdf (accessed on 28 April 2025). [CrossRef]
138. Khan, A.A. Reconceptualizing Policing for Cybercrime: Perspectives from Singapore. *Laws* **2024**, *13*, 44. [CrossRef]
139. Ranchordas, S. Experimental regulations for AI: Sandboxes for morals and mores. *Morals Mach.* **2021**, *1*, 86–100. [CrossRef]
140. Buocz, T.; Pfotenhauer, S.; Eisenberger, I. Regulatory sandboxes in the AI Act: Reconciling innovation and safety? *Law Innov. Technol.* **2023**, *15*, 357–389. [CrossRef]
141. Morgan, D. Anticipatory regulatory instruments for AI systems: A comparative study of regulatory sandbox schemes. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, Montréal, QC, Canada, 8–10 August 2023; pp. 980–981. [CrossRef]
142. Yang, J.; Yue, Z.; Yuan, Y. Noise-aware sparse Gaussian processes and application to reliable industrial machinery health monitoring. *IEEE Trans. Ind. Inform.* **2022**, *19*, 5995–6005. [CrossRef]
143. Mishra, A.; Gangisetti, G.; Khazanchi, D. Integrating edge-AI in structural health monitoring domain. *arXiv* **2023**, arXiv:2304.03718. [CrossRef]
144. Paquet-Clouston, M.; Romiti, M.; Haslhofer, B.; Charvat, T. Spams meet cryptocurrencies: Sextortion in the bitcoin ecosystem. In Proceedings of the 1st ACM Conference on Advances in Financial Technologies, Zürich, Switzerland, 21–23 October 2019; pp. 76–88. [CrossRef]
145. Feichtinger, R.; Fritsch, R.; Vonlanthen, Y.; Wattenhofer, R. The hidden shortcomings of (D)AOs–An empirical study of on-chain governance. In *International Conference on Financial Cryptography and Data Security*; Springer: Cham, Switzerland, 2023; pp. 165–185. Available online: https://link.springer.com/chapter/10.1007/978-3-031-48806-1_11 (accessed on 28 April 2025).
146. Yeung, K. Regulation by blockchain: The emerging battle for supremacy between the code of law and code as law. *Mod. Law Rev.* **2019**, *82*, 207–239. [CrossRef]
147. Zhang, P.; Ding, S.; Zhao, Q. Exploiting blockchain to make AI trustworthy: A software development lifecycle view. *ACM Comput. Surv.* **2024**, *56*, 1–31. [CrossRef]
148. Chenna, S. AI and Blockchain: Towards Trustworthy and Secure Intelligent Systems. *SSRN* **2023**. [CrossRef]
149. Mylrea, M.; Robinson, N. Artificial Intelligence (AI) trust framework and maturity model: Applying an entropy lens to improve security, privacy, and ethical AI. *Entropy* **2023**, *25*, 1429. [CrossRef] [PubMed]
150. AlShamsi, M.; Salloum, S.A.; Alshurideh, M.; Abdallah, S. Artificial intelligence and blockchain for transparency in governance. In *Artificial Intelligence for Sustainable Development: Theory, Practice and Future Applications*; Springer International Publishing: Cham, Switzerland, 2020; pp. 219–230.
151. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*; Profile Books: London, UK, 2019.
152. Caarten, A.B.; van Heugten, L.; Merkle, O. The intersection of corruption and gender-based violence: Examining the gendered experiences of sextortion during migration to South Africa. *Afr. J. Reprod. Health* **2022**, *26*, 45–54. Available online: https://www.ajol.info/index.php/ajrh/article/view/229679 (accessed on 28 April 2025).
153. Mumporeze, N.; Han-Jin, E.; Nduhura, D. Let's spend a night together; I will increase your salary: An analysis of sextortion phenomenon in Rwandan society. *J. Sex. Aggress.* **2021**, *27*, 120–137. [CrossRef]
154. Patil, S.; Desai, D. AI Enabled Blockchain solution for the Indian Judicial System. In Proceedings of the 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 1–3 March 2023; pp. 1–6. [CrossRef]
155. Ahmad, L.; Khanji, S.; Iqbal, F.; Kamoun, F. Blockchain-based chain of custody: Towards real-time tamper-proof evidence management. In Proceedings of the 15th International Conference on Availability, Reliability and Security, Virtual Event, 25–28 August 2020; pp. 1–8. [CrossRef]
156. Collingridge, D. *The Social Control of Technology*; Pinter: London, UK, 1982.
157. Ahuja, R.; Chug, A.; Gupta, S.; Ahuja, P.; Kohli, S. Classification and clustering algorithms of machine learning with their applications. In *Nature-Inspired Computation in Data Mining and Machine Learning*; Patnaik, S., Yang, X.-S., Naik, B., Eds.; Springer: Cham, Switzerland, 2020; pp. 225–248. Available online: https://link.springer.com/chapter/10.1007/978-3-030-28553-1_11 (accessed on 28 April 2025).
158. White, J.; Hays, S.; Fu, Q.; Spencer-Smith, J.; Schmidt, D.C. ChatGPT prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. In *Generative AI for Effective Software Development*; Springer Nature: Cham, Switzerland, 2024; pp. 71–108. Available online: https://link.springer.com/chapter/10.1007/978-3-031-55642-5_4 (accessed on 28 April 2025).

159. Lu, Y.; Liu, S.; Zhang, Q.; Xie, Z. Rtllm: An open-source benchmark for design RTL generation with large language model. In Proceedings of the 2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC), Incheon, Republic of Korea, 22–25 January 2024; pp. 722–727. [CrossRef]
160. Nazir, A.; Chakravarthy, T.K.; Cecchini, D.A.; Khajuria, R.; Sharma, P.; Mirik, A.T.; Kocaman, V.; Talby, D. LangTest: A comprehensive evaluation library for custom LLM and NLP models. *Softw. Impacts* **2024**, *19*, 100619. [CrossRef]