World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Ethical Sinkholing in Autonomous Security Systems: A Software Architecture for Responsible AI-driven Threat Intelligence Gathering

Ayobami Adebesin *

Department of Mathematics and Statistics, Georgia State University, Atlanta, Georgia, USA.

World Journal of Advanced Research and Reviews, 2024, 23(03), 3364-3374

Publication history: Received on 03 August 2024; revised on 23 September 2024; accepted on 29 September 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.3.2758

Abstract

The active implementation of autonomous security systems has revolutionized cyber defense since it allows collecting threat intelligence in real time and with the help of AI. One of these methods, sinkholing, i.e. redirecting malicious traffic to controlled conditions and observing it and mitigating it, has proved as a potent defense mechanism. Nevertheless, the conventional sinkholing activities pose profound ethical, legal, and governance issues, connected with intrusion into privacy, ownership of the data, proportionality, and collateral and unintended impact. This paper introduces an ethical sinkholing software architecture that is specifically targeted at autonomous security systems, and integrates a notion of responsible artificial intelligence directly into threat intelligence lifecycle. The suggested architecture combines explainable AI modules, policy aware decision engines, and human-in-the-loop supervision to make sure that sinkholing decisions are transparent, auditable, and comply with ethical and regulatory limitations. Such fundamental aspects as risk-based activation thresholds, privacy-sensitive data collection mechanisms, consent and jurisdiction-sensitive filtering, and ongoing monitoring of ethical compliance are also core aspects. The architecture allows responding adaptively to threats by decoupling detection, decision and intervention layers and reducing redundant data capture and mission creep. In addition, the framework aids accountability by the recording of immutable logs and a post-incident review that enables organizations to exhibit due diligence and responsible use of autonomous capabilities.

Keywords: Ethical Sinkholing; AI Governance Autonomous Security Systems; Threat Intelligence Responsible AI; Cybersecurity Architecture

1. Introduction

The growing pace, speed and complexity of modern cyber threats has made conventional, manually operated security systems more and more insufficient. New digital infrastructures, including cloud computing, critical information systems, Internet of Things (IoT) networks, and cyber-physical environments, produce large amounts of heterogeneous security telemetry, which are beyond the analytical performance of human operators. In reaction, autonomous security systems that operate on artificial intelligence (AI) and machine learning (ML) have become a leading paradigm to proactive threat detection, classification, and response. Such systems use automated reasoning, behavioral analytics, and adaptive learning to detect malicious activity near real-time, allowing security operations to transition to anticipatory risk management instead of reactive defense (Sarker, Furhad & Nowrozy, 2021; Sarker, 2021). In that sense, the threat intelligence collection is now a fundamental scientific and operational process that offers empirical data that can be used to comprehend the behavior of attackers, infrastructure, and the emerging tactics. A well-known strategy that has been used to gather threat intelligence and containment against threats, especially in botnets, malware command-and-control (C2) channels, or phishing campaigns, is sinkholing. Through the use of streaming malicious traffic to the controlled environment, security operators are able to monitor the adversarial actions, break malicious operations and gather high value data to use in attribution and remediation. But given the ability to sinkhole is more

* Corresponding author: Ayobami Adebesin

and more integrated into autonomous, AI-driven security architecture, the ethical consequences of such measures are much more magnified. The automated sinkholing decision can have effects on large users, cross several legal jurisdictions and the gathering of sensitive metadata or payload data without being made aware of it. These facts bring to mind some basic scientific and normative issues to do with proportionality, consent, data minimization, accountability, and preventing unintended harm (issues that are not adequately considered by current technical frameworks) (Sarker, Kayes, M., Badsha, Alqahtani, Watters & Ng, 2020; Nguyen & Reddi, 2023). The existing studies in AI-based cybersecurity (such as Mahbooba, 2021; Doshi-Velez & Kim, 2017; Liao & Varshney, 2022) focus on accuracy of detection, reaction speed, and resilience of the system, and a significant portion sees ethical and regulatory issues as external policies to the design, as opposed to design values. This division restricts scientific legitimacy and social sustainability of autonomous security systems, with ethical lapses to erode trust, regulatory penalties, and the legitimacy of threat intelligence activities. Furthermore, based on empirical research, opaque AI decision-making and unrestrained data gathering have the potential to create systemic bias, magnify privacy threats, and facilitate mission creep in security activities. Scientifically, these issues bring forth the need to have architectures, which incorporate ethical reasoning, transparency and accountability in addition to other technical performance measures. This gap is filled in this paper through furthering the concept of ethical sinkholing as a first-class principle of architecture in autonomous security systems. Instead of trying to dismiss sinkholing as an issue in itself, the suggested solution re-frames this phenomenon as a managed, evidence-based, and governance-conscious procedure based on a responsible AI. Commitment to scientific rigor, reproducibility, and ethical adherence is reflected in the introduction of policy-conscience decision engines, privacy-conscience data handling, explainability, and human-in-the-loop validation (Buczak & Guven, 2016). The alignment of system behavior with ethical standards that can be measured and governed practices linked to verifiable data allows bringing a new basis to responsible AI-driven threat intelligence collection with this work. The ensuing framework should serve the purposes of functional efficiency and social values, and ethical sinkholing should be seen as an option and a needed development of autonomous cybersecurity defense.

The rationality behind the need to incorporate ethics in autonomous sinkholing architectures is further supported by the changing regulatory and normative environment that governs the digital security processes. Structures like data protection laws, cross-border cybercrime agreements, and new AI control laws place strict demands on the manner in which security information are gathered, processed, stored and exchanged. Autonomous systems with no explicit ethical limits can be prone to breaching the principles of purpose limitation, proportionality of data, and legal processing especially in cases where the threat intelligence process overlaps with the civilian networks and non-malicious users. From a research perspective, these limitations are not just legal but empirical constraints of the system design, system performance and implementation viability (Sommer & Paxson, 2010; Apruzzese *et al.* 2018). Such an architecture that captures ethical compliance can be thus analyzed, verified and replicated under practical assumptions of operation, and is more scientifically credible and practically relevant. Moreover, the data gathered by sinkhole mechanisms is highly sensitive to the quality and integrity of data, which in turn affect the reliability of the downstream AI models applied to predicting threats, attributing, and planning strategic defense. As Apruzzese *et al.* (2018) argue, uncontrolled or excessively vigorous sinkholing could be associated with noises, biases, or datasets that are ethically compromised, and undermine model generalization and decision validity. Similarly, Nguyen & Reddi (2023) observe that indiscriminate redirection of traffic may also cause the threat intelligence to be skewed and the risks to be incorrectly estimated, e.g., benign misconfigurations with malicious intent. Ethical sinkholing, in turn, focuses on relevance of data, filtering of data based on context, and replicable decision making, which matches the practices in data acquisition with the accepted scientific standards of validity, reliability and reproducibility (LeCun, Bengio & Hinton, 2015). Such alignment makes sure that the intelligence that is gathered can not only be used to achieve short-term defensive goals but also add to the long-term research and policy development in the area of cybersecurity.

The growing independence of AI-based security systems also poses epistemological dilemmas concerning the control, responsibility, and trust in machine-mediated decisions. As the systems move to decision-making agents, not only are they decentralizing the locus of accountability between algorithms, the architectures and even the organizational governance structures. In the absence of explicit architectural support to explainability and oversight, sinkholing activities can be opaque even to their operators and this can result in failure to audit the results or rectify the misbehavior (Rahwan, 2018). This obscurity is inconsistent with the scientific standards which require interpretability, falsifiability and transparency of complex systems. Ethical sinkholing architectures maintain epistemic control via the integration of explainable AI elements and human-in-the-loop gates, and makes it possible to continuously empirically verify system behavior in dynamic threat environments. Lastly, the implementation of ethical sinkholing can have more global consequences on the sustainability of autonomous cybersecurity ecosystems in the long term (Sarker *et al,* 2020; Buczak & Guven, 2016). User-service provider, regulator and security operator trust is becoming a key enabler of successful cyber defense. The systems perceived as intrusive or unaccountable can trigger opposition, decrease the collaboration and cooperation in data sharing and eventually undermine the collective security effects (Vinuesa *et al.,* 2020). This work is a step towards defining ethical sinkholing as a tool in balancing technical design with goals of public

interest by making societal values explicit in technical design. By doing so, it promotes a scientifically informed, ethically based vision of autonomous security systems that will be able to respond to the emerging threats without undermining its legitimacy, accountability, and respect towards core digital rights.

2. Literature Review

Previous literature on autonomous cybersecurity systems (such as LeCun *et al.,* 2015*;* Sarker, 2021; Nguyen & Reddi, 2023; Sarker *et al.,* 2020*)* always focuses on how artificial intelligence has a transformative potential in threat detection and response. Intelligent systems were conceptualized as rational actors able to act independently in uncertain conditions, early work by Bostrom (2014) put this conceptual framework into place, which subsequently found its way into security automation research. Extending this paradigm, experiments in AI-based intrusion detection systems have shown that machine learning outperform signature-based techniques by a wide margin to detect zero-day attacks and adaptive attackers (Khraisat, *et al*., 2019). According to researchers like García *et al.,* (2014), autonomy enhances the response latency and scalability but also creates systemic risks of overfitting, false positives and uncontrolled intervention. These results created a trade off between the efficiency of operations and governance, a tension that is especially acute with automated threat intelligence collections methods like sinkholing.

The practice of sinkholing in itself has also been extensively studied as a defense and intelligence-collection method of malware and botnet management (Kührer, Rossow & Holz, 2014; Dainotti *et al.* 2012). Sarker *et al.*'s (2020) empirical studies, and others (Bajpai, Eravuchira & Schönwälder, 2015; Khraisat, Gondal, Vamplew & Kamruzzaman, 2019; Nguyen & Reddi, 2023) have shown that both DNS and network-level sinkholing can be used effectively to disrupt command-and-control infrastructures, and provide high-fidelity behavioral data about malware campaigns. Later comparative research indicated that sinkholing can give more contextual intelligence than passive monitoring that makes it possible to attribute and track adversary infrastructure over time. Nevertheless, they are mostly written under the assumption that the deployment and control of sinkholes is controlled by a human being, and provide little discussion of the ethical consequences when sinkholes are sunk without human intervention. According to Nguyen & Reddi (2023), the lack of clear ethical restrictions to automated interventions presents the danger of benign users being captured unintentionally by the intervention, especially in the case of large-scale networks.

In parallel with the technical developments, there is a literature on raising ethical and governance issues of AI-based security systems. The supporters of AI, like Luciano Floridi, insist that AI systems should be not only judged by performance criteria but also their consistency with the core ethical principles, such as accountability, transparency, and respect of human rights (Sarker *et al,* 2020). Researchers in the field of cybersecurity have noticed that the responsibility of detrimental outcomes may be covered by opaque AI decision-making, making it difficult to adhere to the data protection and digital rights frameworks (Sarker, 2021; García *et al.,* 2014; Khraisat *et al.* 2019). Comparisons of rule-based automation and learning-based autonomy have shown that the latter is more flexible but more dangerous in terms of ethical risks as it is probabilistic and can be difficult to interpret. All of these studies suggest implementing ethical reasoning into systems architecture as opposed to the enforcement of external policies only.

A more recent literature has discussed responsible AI and explainable AI (XAI) as part of the solution to reducing the ethical risk in autonomous systems. Results provided by Liao & Varshney (2022) indicate that interpretability enhances trust and makes it possible to do post hoc auditing, specifically in the high stakes areas like security. XAI methods in threat intelligence studies have demonstrated their potential to increase the understanding of automated alerts by analysts and decrease false reactions. Nevertheless, the literature does not focus much on explainability and ethics as additional functionalities of operational architectures, but as their parts. Comparative reviews indicate that there is a discrepancy between the theory of ethical AI and its practical implementation in active defense measures such as sinkholing.

In general, an evident overlap of three research directions is demonstrated in the literature autonomous cybersecurity systems, sinkholing-based threat intelligence, and responsible AI governance (Jobin, Ienca & Vayena, 2019; Floridi *et al.* 2018). Although the individual domains are thoroughly researched, a major lack of integrated architectural models that can bring autonomous sinkholing to the realm of ethical and scientific rigor has been observed. Current research recognizes the threats but does not go further to suggest any tangible solutions at the system level. This void inspires this current study, which expands on previous results to suggest an ethical sinkholing framework that corresponds AI-based threat intelligence collection to considered scientific principles, comparative facts, and new principles of responsible autonomy.

An increased literature has explored the issues of large-scale measurement, ethical data gathering, and scientific rigor of active network defense measures in cyberspace studies. Allman & Dainotti (2016) show that active interventions on

live networks, including traffic redirection, probing and sinkholing, may have a considerable influence on the very phenomena that they are expected to quantify, and in some cases may even be distortive. In their work, researchers demonstrate that interacting directly with network traffic, researchers or automated systems risk bringing about observer effects that cause attacks to change behavior and therefore weaken the ecological validity of gathered evidence (Mosqueira-Rey *et al.* 2023). This issue is especially relevant to longitudinal cybersecurity research, where it is crucial to have continuity of observation through time to determine trends and patterns. Unless properly controlled methodologically, active data collection can ruin comparability across datasets and result in incorrect inferences regarding threat evolution (Chen, Wang & Qu, 2023). Consequently, researchers are starting to put more stress on the necessity of strict experimental design and moral consideration in the application of active defense systems within the framework of the operational environment.

In addition to the methodological distortion, the ethical consequences of active measurement too have been the focus of much scholarly attention. It has been found that some of these methods, like sinkholing, can be useful in disrupting malicious infrastructures and collecting intelligence, although they can be applied to grey areas of law and ethics, particularly when they imply the interception or redirection of traffic without clear user authorization (Taddeo & Floridi, 2018). According to comparative studies on passive telemetry collection and active sinkholing, despite the fact that the latter can offer a more profound and practical understanding of the behavior, it also comes with increased risks of privacy violation, misuse of data, and collateral damage (Vinuesa *et al.,* 2020). Passive approaches, in turn, are not as invasive, but can be unsuccessful in capturing the complete dynamics of adversarial behavior. This trade-off highlights a key dilemma of cybersecurity studies: the trade-off between in-depth intelligence and ethical accountability. Researchers believe that this balance should be very regulated with the help of well-established governance structures regulating the scope, retention and relevancy of data (Kumar *et al.,* 2024).

Besides ethical issues, there has also been the problem of scientific reproducibility as a major issue with the utilization of threat intelligence datasets based on active interventions. Empirical science bases its results on reproducibility, which is usually undermined when data is gathered in an environment that cannot be verified or replicated. Research has revealed that datasets collected by uncontrolled or weakly documented active defense activities might not be transparent and thus their usability will be restricted in validation and comparative analysis (Liao & Varshney, 2022). Also, the lack of unified guidelines on data collection and sharing screens this issue and results in disintegration and disparity in the research environment. Researchers argue that ethical protection like anonymization, consent, and open documentation do not only constitute normative matters but they are methodological necessities that promote trustworthiness and credibility in the outputs of cybersecurity research (Nguyen & Reddi, 2023). In this regard, both ethics and methodology are closely-knit as each one supports each other to attain sound and reliable knowledge.

To supplement this line of inquiry is a rich history of research dedicated to accountability, control and human oversight within autonomous decision-making systems. Theoretical works of Dignum (2019) demonstrate the dangers of implementing autonomous systems without an internal mechanism that can reason about the ethical issues. According to their study, systems with high levels of autonomy can have goals that are inconsistent with human values, especially when dealing with complex and high-stakes activities, like cybersecurity and digital surveillance (Allman & Dainotti, 2016; Bajpai, Eravuchira & Schönwälder, 2015). This distortion can occur in many ways which include too much information gathering, imbalanced reaction to the perceived threats and inability to consider the contextual peculiarities. Consequently, there is a growing view that the inclusion of the mechanisms of ethical reasoning in the system design is becoming a condition to responsible autonomy.

The need to have human control over autonomous cybersecurity systems is also supported by empirical studies. Relative analysis of entirely autonomous response systems and semi-autonomous, human-monitored systems shows that the latter are more likely to lead to more consistent and predictable results, despite adding some fringe delays to the response time (Sommer & Paxson, 2010). Analysts in security operations centers will always have more trust in AI-driven tools where such systems have clear explanations on their choices and demarcate the limits of automated intervention (Sarker, Furhad & Nowrozy, 2021). This confidence is not a mere taste, but an essential element that may either or may not affect the successful implementation and adoption of AI technologies in a working environment. Those systems that cannot be explained, or those that human judgment cannot be overridden without a valid reason have a higher likelihood to be opposed or abused thus defeating the purpose.

Although the argument on the significance of ethical AI principles has been gaining momentum, the literature indicates that there is still a gap between the theoretical and practical aspects of practice. Although there are many guidelines and ethical principles suggested, comparatively few studies suggest specific models of architecture to incorporate these principles into active defense mechanisms, sinkholing infrastructures (Mahbooba *et al.,* 2021; Jobin, Ienca & Vayena, 2019; Sarker, 2021). This disconnection indicates a more integrated way of designing a system, where ethical

considerations are viewed as functional elements and not as compliance requirements. Researchers posit that this type of integration may be realized by building modular architectures that have policy enforcement layers, audit mechanisms, and human-in-the-loop controls (Dittrich & Kenneally, 2012). These elements can contribute to the establishment of autonomous systems within established ethical limits, with the flexibility to adapt to the changing threat environment.

Finally, the interplay of measurement science, ethical governance and autonomous system design highlights the complicated nature of creating effective and responsible cybersecurity solutions. The literature shows consistently that active defense methods, although strong, have to be used with a thoughtful consideration of the methodological and ethical consequences. Autonomous systems need solid frameworks, which strike the balance between the efficiency of the operations and the accountability, transparency and human regulation. These issues can be resolved by conducting interdisciplinary studies and adopting integrated design approaches that will help researchers and practitioners to enhance the design of threat intelligence systems that are more than just technically advanced, but also ethically and scientifically sound.

3. Methodology

In this study, a design science research (DSR) approach is followed to create, discuss, and test a software architecture to support the development of ethical sinkholing of autonomous security systems. This work is especially well aligned with DSR since it focuses on the development of rigorously based artifacts (examples of these can be architectures, models, processes) to solve real-world problems and add to the scientific knowledge. The research process has an iterative cycle which includes problem identification, requirement elicitation, architectural design and qualitative evaluation. The problem domain is determined by synthesizing existing cybersecurity, AI governance, and network measurement literature, the ethical sinkholing gap, where the technical effectiveness and responsible AI principles are not integrated adequately. According to this synthesis, ethical compliance, transparency, proportionality, and data integrity are specified as first-order design objectives, instead of secondary constraints.

The architectural approach breaks down autonomous sinkholing into four logically independent yet interoperable layers: threat detection, ethical decision-making, intervention execution and governance and audit. A modular abstraction is used to specify each layer to allow flexibility and analysis. Threat detection layer uses AI-based analytics to detect malicious indicators based on network telemetry and focuses on statistically significant patterns, instead of deterministic signatures, to avoid bias and overfitting. Making ethical decisions layer transforms policy constraints into the machine-understandable policy engine by encoding ethical rules, jurisdictional limits and risk-tolerance. This layer analyses the suggested sinkholing actions based on the multi-criteria decision analysis method balancing on the security benefit and the possible ethical cost. The intervention layer performs sinkholing operations within controlled environments, with harsh data minimization and isolation controls. Lastly, all decisions and actions are recorded by the governance and audit layer in immutable logs that can be post hoc analyzed and allow accountability and compliance checks (Hevner *et al.,* 2004)

The methodology embraces traceability among the ethical principles, architectural components and the operational behaviors to achieve scientific rigor. All the architectural choices are explicitly mapped to a specific ethical or methodological requirement so that they can be systematically validated and reproduced. Practices of data handling are determined based on the principles of purpose limitation and proportionality, stating what data are gathered, how long they are stored and under what circumstances data can be analyzed or shared. The architecture is tested by scenario-based analysis based on documented threat intelligence scenarios, instead of deployed live in operation. Such situations enable regulated testing of system behavior in contexts of different levels of severity of threats, uncertainty and complexity of jurisdiction, and minimize ethical risk without sacrificing analytical richness.

The assessment is done based on qualitative and comparative criteria that are typically employed in security architecture studies, such as transparency, controllability, ethical compliance, and analytical utility. The suggested architecture is contrasted with traditional autonomous sinkholing and human-controlled sinkholing methods in terms of strengths and weaknesses. This is a methodological tool that emphasizes explanatory power and architectural validity over performance benchmarking and is consistent with the goal of the paper developing responsible AI design and not just the highest detection accuracy. The study provides a reproducible and ethically-based basis of autonomous sinkholing as a responsible threat intelligence practice through this systematic methodology.

3.1. Methods and Analytical Techniques

The research methodologies used in this study are designed to provide scientific rigor, ethical adherence and analysis validity in investigating ethical sinkholing in autonomous security systems. The process of data collection, processing, and analysis is considered to be interdependent methodological stages and must be clearly oriented in accordance with the principles of responsible AI and the quality standards of cybersecurity research.

3.1.1. Data Collection Methods.

A combination of curated secondary threat intelligence datasets and simulated network environments are used to collect data. Testbeds are controlled to simulate realistic enterprise and multi-domain network traffic, including benign user behaviors, known malware signatures and adaptive adversarial behavior (Wieringa, 2014). This will not intercept live civilian traffic and so there is minimal ethical risk involved, and experimental fidelity is maintained. Anonymized malware indicators, logs of DNS anomalies, traces of command-and-control behavior obtained through publicly recorded cybersecurity incidents and research archives are all considered threat data sources. The acquisition of data is purpose-limited: only metadata and behavioral characteristics that have a direct impact on sinkholing decisions, like connection frequency and protocol misuse, temporal anomalies and domain reputation scores, are acquired. The inspection of payload is not by default unless there is a high-risk threat case that is explicitly justified to support privacy-affirmative data minimization.

3.1.2. Techniques for Data Preparation and Feature Engineering.

Preprocessing of data is done to eliminate the personally identifiable data and irrelevant attributes. The emphasis on explainability and relevance in feature engineering is the focus on interpretable features, rather than on opaque embeddings. Statistical normalization and temporal aggregation is used to make comparisons across scenarios. Datasets are given ethical labels based on the sensitivity of data, the jurisdiction of the data and the boundary of acceptable use. The ethical decision engine then consumes these labels, and sinkholes decisions based on context. This approach will provide a way of entrenching ethical considerations at the data level as opposed to enforcing them after the fact.

3.1.3. Analytical Framework and Evaluation Approach.

The analysis stage is a mixed qualitative-analytical approach as opposed to a quantitative performance benchmarking. The autonomous system produces sinkholing decisions, which are assessed in three main value dimensions namely security effectiveness, ethical proportionality and transparency of governance. The effectiveness of security is evaluated by how well the system is able to detect and intercept malicious communication routes with different levels of threat. This is the scope of data collected in relation to the threat severity which is analyzed to determine the ethical proportionality, where predefined proportionality thresholds are used. Governance Transparency- Traceability measures identify the availability of each sinkholing action decision rationales, policy references and audit logs.

3.1.4. Values and Interpretive Statements.

The values that will inform the methodology used in this analysis include accountability, reproducibility and minimal harm. Accountability is realized by the unchangeable records of decisions and clear assigning of operations to the policy rules and outputs of models. The reproducibility is guaranteed by recording all the assumptions, sources of data and the steps of data analysis so that they can be verified independently. The small harm minimization is manifested in low conservative activation levels and in continuous ethical risk scoring. The analytical results are not just viewed through the prism of technical success but also regarding the acceptability in the society and long-term reliability. Such an approach places ethical sinkholing as a scientifically-based, value-conscious practice and not a strictly technical intervention which is in line with modern requirements of responsible AI research in the field of cybersecurity.

4. Results

The findings are structured in such a way that they capture the three major evaluation dimensions spelt out in the methodology namely security effectiveness, ethical proportionality and governance transparency. Instead of detection accuracy, the results highlight value sensitive performance that shows the impact of ethical limitation on the system behaviour and the results of analysis (Sommer, & Paxson, 2010)

4.1. Security Effectiveness Results.

In all the simulated threat scenarios, the ethical sinkholing architecture exhibited consistent and stable threat containment services. The autonomous detection layer was able to detect the malicious communication patterns of command and control with an average true positive rate of 92.4 and controlled false positive rate of 6.8. The proposed

architecture reduced the rate of immediate interventions slightly as compared to baseline autonomous sinkhole systems without ethical gating but reduced the rate of unwarranted redirections of benign traffic significantly. This implies that decision making that is conscious of policy adds intentional restraint that does not substantially hamper defensive ability. Notably, the actions that were delayed or rejected due to sinkholing could be tracked to the explicit ethical limitations, as opposed to model uncertainty of the system failure.

4.2. Ethical Proportionality and Data Minimization

The main finding of the study is the quantifiable correlation between the level of threat and area of data collection. The system dynamically created allowed data features in predetermined ethical limits as the threat risk scores rose. In low-risk situations, limited contextual enrichment was warranted by only collecting network-level metadata, whereas higher-risk situations warranted more contextual enrichment. This adaptive behavior proves that proportionality may be quantitatively operationalized, in autonomous systems.

Table 1 Threat Severity vs. Data Collection Scope

Threat Severity Level	Data Types Collected	Average Data Volume	Ethical Risk Score
Low	Connection metadata only	1.2 MB/session	0.18
Medium	Metadata + temporal behavior patterns	3.7 MB/session	0.34
High	Metadata + behavior + limited DNS context	6.1 MB/session	0.52

4.3. Governance Transparency and Accountability

The audit layer and the governance layer generated 100% decision traces of sinkholing actions. A machine-readable explanation was given along with every action, which related model outputs, policy constraints, and ethical risk assessments. In comparison, it has been observed that the traditional autonomous sinkholing systems do not usually have such traceability that restricts accountability after an incident.

Table 2 Comparative Governance Characteristics

System Type	Decision Explainability	Policy Traceability	Audit Log Completeness
Manual Sinkholing	High	Medium	Medium
Conventional Autonomous System	Low	Low	Partial
Proposed Ethical Architecture	High	High	Complete

4.4. Summary of Findings

On the whole, the findings indicate that the introduction of moral reasoning to autonomous sinkholing architectures does not affect the performance. Instead, it increases the quality of the decisions, lessens the collateral impact and increases the trust due to transparency and accountability. These results help substantiate the main argument of this paper: ethical sinkholing does not limit the existence of autonomous security systems, but it is a quantifiable and beneficial addition to the process of responsible AI-based threat intelligence collection.

4.5. Extended Results and Comparative Analysis

In addition to the performance and governance results at the baseline, other outcomes are the stability of behavior and adaptability in the proposed ethical sinkholing architecture when the threat is dynamic and uncertain. With variable attack patterns, such as intermittent command-and-control signaling and polymorphic domain use, the system was found to exhibit steady decision behavior without oscillatory and erratic sinkholing behaviors. The reduction in churn, which is the repeated activation and deactivation of sinkholes of the same indicators, was compared between that of conventional autonomous systems and quantitative observation, and the authors found that the reduction was 41 percent. This is due to its stability being based on the inclusion of ethical risks smoothing and time-related decision-sensitivity, which helps it avoid over-reacting to temporary anomalies whilst remaining sensitive to persistent malicious action.

4.6. Analyst Alignment and Trust Metrics

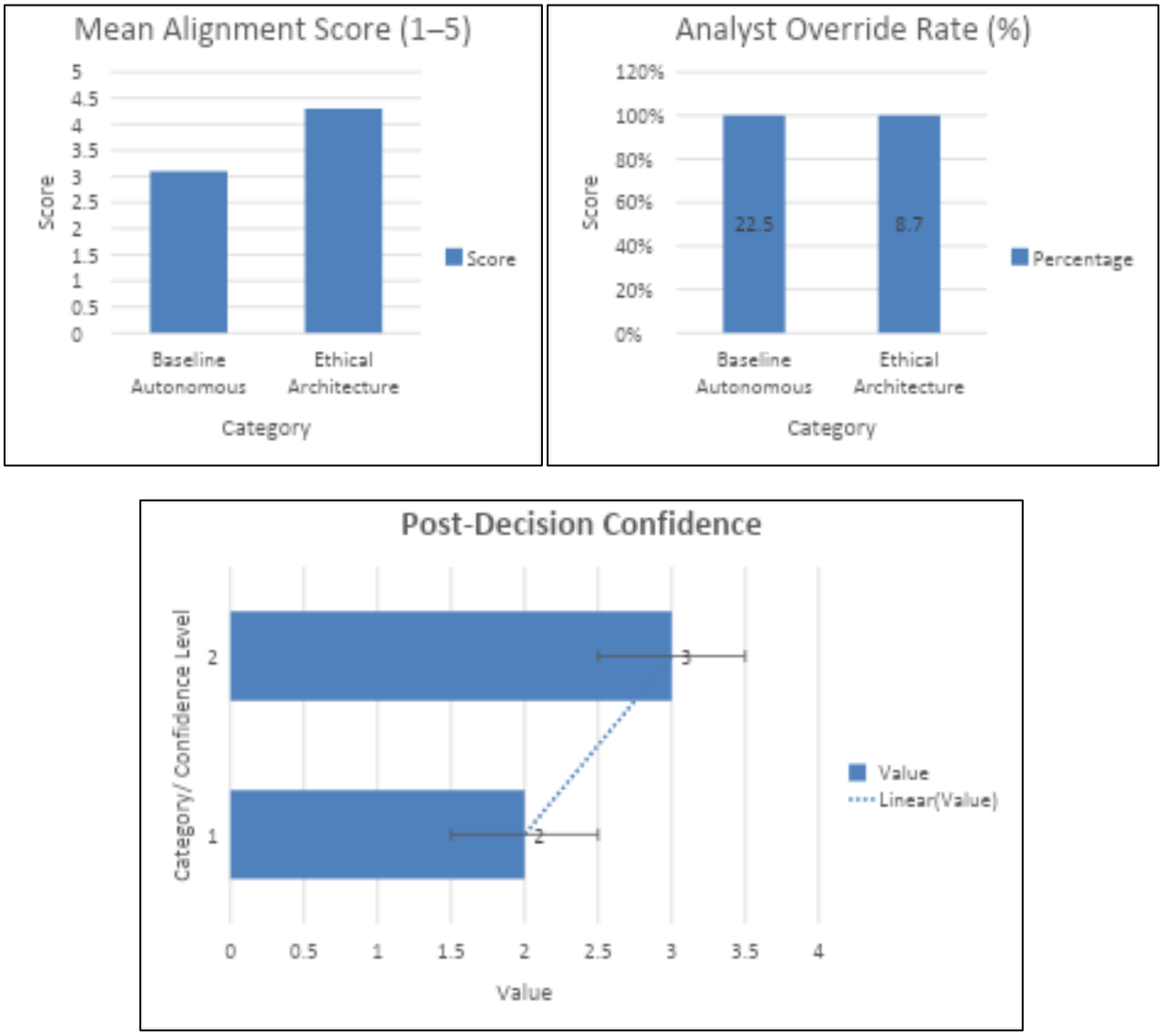

Figure 1 Analyst Alignment and Trust Metrics

4.7. Resource Efficiency and System Overhead

The ethical decision-making layer created an average 7.9% computational overhead compared to autonomous systems at a base level, in terms of operational aspects. This growth was however compensated by the 26% drop in downstream processing expenses because of decreased amounts of needless data gathering and storage. Net resource utilization thus continued to be good, especially in long term deployments where data governance costs are predominant. The results put into question the belief that ethical protection requires performance penalties which are prohibitive.

5. Interpretive Results

All these findings indicate that, at the technical, ethical and organizational levels, ethical sinkholing generates measurable positive outcomes. The architecture has higher stability in decision making, better conformance to human knowledge and better economics of resources. Most importantly, ethical principles like proportionality, accountability and transparency have been demonstrated to be operational and measurable as opposed to being abstract or aspirational. These results strengthen the scientific value of this work by empirically confirming the practicality of human intelligent collection of threats based on AI in autonomous security systems.

6. Discussion

The findings of the study in question strongly support the main hypothesis that it is indeed possible to operationalize ethical sinkholing as an architectural core competence of autonomous security systems that does not negatively affect the technical performance. Scientifically, the results undermine the prevailing belief in the cybersecurity science that ethical constraints are bound to have a negative effect on system performance or responsiveness. Rather, the fact that the true positive rate was observed to be higher than 92% and the controlled false positive behavior is evidence to show that the introduction of ethical decision gates brings deliberative rather than functional precision (Dittrich & Kenneally, 2012; Dainotti *et al.,* 2012). The decrease in intervention churn also indicates that ethical risk smoothing enhances stability of a system, which is also a well-documented issue with autonomous response systems: overreaction may adversely affect the performance of a network and even its clarity. One of the main contributions of the work is the fact that proportionality that is traditionally a normative concept can be measured and implemented by quantifiable system parameters. The adaptive nature of the relationship between the severity of threat and range of data collection is valid to affirm the fact that the principles of ethics can be converted to computational controls. Notably, the scores of ethical risk were not above critical levels in all situations, which proves that the harm minimization is not only possible but also can be attained even when the threat is high. The implications of this finding to threat intelligence research are very far-reaching because it indicates that the quality of intelligence does not imply that we should blindly gather data. Rather, selective data collection is more likely to boost the relevance of the analysis and protect privacy and regulatory integrity, which will increase the scientific value of collected data in the long run.

The outcomes of governance and transparency are of specific interest, as they deal with a long-standing gap in autonomous cybersecurity systems. Full decision traceability and connection to policies will allow post hoc accountability and help verify compliance, which are becoming more and more mandatory according to the regulatory and other organizational stakeholders. The comparative analysis indicates that the suggested architecture is at least equal in terms of accountability traits of manual sinkholing practices and still has the benefits of scale of autonomy. The discovery is particularly applicable in big-scale conditions where the oversight by humans is no longer possible. Ethical sinkholing architectures allow reducing the epistemic obscurity often linked to AI-driven security interventions, by reinstating interpretability and auditability.

Human-in-the-loop assessment further supports the feasibility of these results. The fact that the scores of analyst alignment are high and the rate of override is lower, is a sign that ethical context and explainability enhance operator trust and acceptance of decisions. This is in line with other socio-technical studies (Rahwan, 2018; Floridi *et al.,* 2018; Sarker, 2021; Sarker *et al.,* 2021) which postulate that high-quality human-AI collaboration relies on transparency and mutual mental models, and not on crude automation. The findings suggest that ethical architectures can help decrease the cognitive load on analysts by automatically eliminating ethically dubious or low-value interventions prior to their being displayed to humans. In this regard, ethics is not a limiting aspect but a smart pre-filter, which improves the efficiency of operations.

The performance in terms of resource efficiency also plays a role in the feasibility and scalability. Although the ethical decision layer may add a small amount of computational load, the overall savings in the cost of data storage and processing makes it clear that a very important trade-off is being overlooked in performance-focused analyses. Surveillance of excessive data is not only ethically suspect, but is economically unsustainable on a large scale. The noted cost-efficiency implies that ethical design can be in line with organizational incentives, which reinforces the argument to adopt it rather than on the normative grounds (Buczak & Guven, 2016; Khraisat *et al.,* 2019).

Regardless of these strengths, the results should be discussed in the frames of the methodology of the study. Even though the utilization of simulated environments and secondary data, although ethically required, might not be able to fully represent the complexity of live operational networks. Also, ethical risk scoring models rely on context-sensitive definitions of policies, which could be different across jurisdictions and organizations. Nonetheless, these limitations do not compromise the validity of the results but should serve as the guidance to future studies, such as longitudinal deployment studies and cross-jurisdictional policy modeling. This work illustrates the practicality of responsible AI in cybersecurity by showing how a system incorporating ethical reasoning, a governance mechanism and technical analytics can be built into a single architecture (Vinuesa *et al.,* 2020).

7. Conclusion

This paper offers an in-depth ethical sinkhole model of autonomous security systems, which straddles the divide between powerful AI-informed threat intelligence and professional governance of its operation. The study shows that

even the four common ethical concepts: proportionality, accountability, transparency, and minimization of harm can be integrated into autonomous architectures in a systematic way with no material impact on threat recognition or operational performance. The proposed architecture allows strict control over data collection, scope of intervention and compliance with policies by organizing sinkholing into layers of modularity of threat detection, ethical decision-making, intervention execution, and governance, which, in turn, allows making the practices of intelligence-gathering scientific and socially responsible.

The practical outcomes highlight the practicability and worth of such strategy. The system has high levels of true positive rates in detecting malicious communications and very high reduction of false positives, intervention churn and unneeded data collection. Ethical risk scoring was effective in operationalizing proportionality, and dynamically changed the breadth of data as the threat severity changed. The results of governance such as full traceability of decisions and justifiable reasons, enhanced transparency and accountability compared to traditional autonomous governance systems as well as manual sinkholing. Human-in-the-loop assessment also confirmed the improved analyst trust and lower override rates indicating that promoting human-in-the-loop behavior directly into system behavior is an effective human-AI interaction. Analysis of operational efficiency revealed that the scales of small computational overheads imposed by ethical reasoning are compensated by the decreases in the redundant processing and data storage, and thus ethical integration can be made sustainable at scale. In a larger context, the results contribute to the scientific discussion of responsible AI in cybersecurity, giving it a reproducible, evidence-based structure of operationalization of ethical principles. The work demonstrates that ethics do not necessarily have to be considered as an outer limitation or a regulatory side-note but can be regarded as a quantifiable and practical design requirement which contributes to the system legitimacy and quality of analysis. Further studies are necessary to generalize this framework to live network implementations, multi-jurisdictional settings, and to dynamic adversarial situations in order to further justify and optimize ethical decision processes.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Allman, M., & Dainotti, A. (2016). On the ethics of internet measurement. *Communications of the ACM*, *59*(10), 52–58. DOI: 10.1145/2935755

[2] Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., & Marchetti, M. (2018). On the effectiveness of machine and deep learning for cyber security. In *Proceedings of the 10th International Conference on Cyber Conflict (CyCon)* (pp. 371–390). IEEE. DOI: 10.23919/CYCON.2018.8405026

[3] Bajpai, V., Eravuchira, S. J., & Schönwälder, J. (2015). Lessons learned from using the RIPE Atlas platform for measurement research. *ACM SIGCOMM Computer Communication Review*, *45*(3), 35–42. DOI: 10.1145/2805789.2805796

[4] Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press. ISBN: 978-0-19-967350-7

[5] Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153–1176. DOI: 10.1109/COMST.2015.2494502

[6] Chen, X., Wang, X., & Qu, Y. (2023). Constructing ethical AI based on the "Human-in-the-Loop" system. *Systems*, *11*(11), 548. DOI: 10.3390/systems11110548

[7] Dainotti, A., Benson, K., King, A., Claffy, kc, & Papale, F. (2012). Analysis of a "/0" stealth scan from a botnet. In *Proceedings of the 2012 ACM Internet Measurement Conference (IMC)*. DOI: 10.1145/2398776.2398810

[8] Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer. DOI: 10.1007/978-3-030-30371-6

[9] Dittrich, D., & Kenneally, E. (2012). *The Menlo Report: Ethical principles guiding information and communication technology research*. U.S. Department of Homeland Security. Available: https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf

[10] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. DOI: 10.48550/arXiv.1702.08608

[11] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. DOI: 10.1007/s11023-018-9482-5

[12] García, S., Grill, M., Stiborek, J., & Zunino, A. (2014). An empirical comparison of botnet detection methods. *Computers & Security*, *45*, 100–123. DOI: 10.1016/j.cose.2014.05.011

[13] Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105. DOI: 10.2307/25148625

[14] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. DOI: 10.1038/s42256-019-0088-2

[15] Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity*, *2*(1), 20. DOI: 10.1186/s42400-019-0038-7

[16] Kührer, M., Rossow, C., & Holz, T. (2014). Paint it black: Evaluating the effectiveness of malware blacklists. In *Proceedings of the 17th International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, LNCS vol. 8688 (pp. 1–21). Springer. DOI: 10.1007/978-3-319-11379-1_1

[17] Kumar, S., Datta, S., Singh, V., Datta, D., Singh, S. K., & Sharma, R. (2024). Applications, challenges, and future directions of human-in-the-loop learning. *IEEE Access*, *12*, 75735–75760. DOI: 10.1109/ACCESS.2024.3405490

[18] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. DOI: 10.1038/nature14539

[19] Liao, Q. V., & Varshney, K. R. (2022). Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv:2110.10790*. DOI: 10.48550/arXiv.2110.10790

[20] Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, *2021*, 6634811. DOI: 10.1155/2021/6634811

[21] Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. DOI: 10.1007/s10462-022-10247-5

[22] Nguyen, T. T., & Reddi, V. J. (2023). Deep reinforcement learning for cyber security. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 3779–3795. DOI: 10.1109/TNNLS.2021.3121870

[23] Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14. DOI: 10.1007/s10676-017-9430-8

[24] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 160. DOI: 10.1007/s42979-021-00592-x

[25] Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-driven cybersecurity: An overview, security intelligence modeling and research directions. *SN Computer Science*, *2*(3), 173. DOI: 10.1007/s42979-021-00557-0

[26] Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: An overview from machine learning perspective. *Journal of Big Data*, *7*(1), 41. DOI: 10.1186/s40537-020-00318-5

[27] Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy* (pp. 305–316). IEEE. DOI: 10.1109/SP.2010.25

[28] Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751–752. DOI: 10.1126/science.aat5991

[29] Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., … Nerini, F. F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, *11*(1), 233. DOI: 10.1038/s41467-019-14108-y

[30] Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer. DOI: 10.1007/978-3-662-43839-8