10.48047/jocaaa.2023.31.04.66

Designing Ethically Autonomous AI for Cybersecurity Governance and Decision Integrity

Author: Mahmood Afzal Hussain, IL USA.

Abstract

The integration of artificial intelligence into cybersecurity systems has revolutionized threat detection and response capabilities, yet simultaneously introduced complex ethical challenges regarding autonomous decision-making authority. This research examines the design principles and governance frameworks necessary for developing ethically autonomous AI systems that maintain decision integrity while operating within cybersecurity contexts. Through analysis of current AI-driven cybersecurity implementations and ethical frameworks, this study identifies critical tension points between operational efficiency and ethical accountability. The research proposes a layered governance model incorporating human oversight mechanisms, algorithmic transparency requirements, and ethical constraint architectures. Findings indicate that achieving ethical autonomy requires balancing three core dimensions: operational independence for rapid threat response, accountability structures ensuring human oversight of critical decisions, and transparency mechanisms enabling audit and review processes. The study concludes that purely autonomous AI systems remain ethically insufficient for high-stakes cybersecurity decisions, necessitating hybrid human-AI governance models that preserve human moral agency while leveraging computational efficiency. This work contributes practical frameworks for organizations implementing AI-driven cybersecurity systems while maintaining ethical standards and decision integrity.

Keywords: artificial intelligence ethics, cybersecurity governance, autonomous systems, decision integrity, algorithmic accountability, human oversight, AI governance frameworks

1. Introduction

Contemporary cybersecurity landscapes increasingly rely on artificial intelligence systems to detect, analyze, and respond to threats that occur at speeds and scales beyond human cognitive capacity. Organizations worldwide deploy AI-powered security systems capable of autonomous threat hunting, anomaly detection, and in some cases, active countermeasures against cyber attacks. However, this technological capability raises fundamental questions about the ethical implications of granting decision-making authority to non-human agents, particularly when those decisions may impact privacy, security, and human rights (Mittelstadt, 2019).

The concept of autonomous AI in cybersecurity extends beyond simple automation of predefined responses. Modern systems employ machine learning algorithms that adapt behavior based on evolving threat patterns, making decisions that were not explicitly programmed by human designers. This adaptive autonomy creates a qualitative difference from traditional rule-based security systems, introducing uncertainties regarding accountability, transparency, and alignment with human values (Floridi et al., 2018).

The research problem centers on a fundamental tension: cybersecurity threats require rapid, often instantaneous responses that exceed human reaction capabilities, yet ethical decision-making traditionally requires human judgment, deliberation, and moral reasoning. When AI

10.48047/jocaaa.2023.31.04.66

systems make autonomous decisions to block network traffic, quarantine systems, or deploy countermeasures, they effectively exercise power that affects organizational operations and individual rights. The absence of robust ethical frameworks governing these autonomous actions creates risks of unintended consequences, biased outcomes, and erosion of human accountability (Taddeo & Floridi, 2018).

Current approaches to AI ethics in cybersecurity largely focus on post-hoc auditing or abstract principles rather than embedded governance mechanisms that constrain autonomous behavior in real-time. Many organizations implement AI security tools without adequate consideration of ethical implications, driven by competitive pressure and threat urgency. This gap between technological capability and ethical governance creates vulnerabilities not just in security posture but in organizational legitimacy and trustworthiness (Whittlestone et al., 2019).

This research addresses three central questions: What design principles enable AI systems to operate autonomously in cybersecurity contexts while maintaining ethical integrity? How can governance frameworks ensure accountability and transparency for autonomous AI decisions? What mechanisms preserve meaningful human oversight without negating the operational advantages of AI autonomy? The investigation synthesizes ethical theory, cybersecurity practice, and AI governance literature to develop actionable frameworks for responsible autonomous systems.

The significance of this work extends beyond theoretical ethics into practical organizational imperatives. As regulatory frameworks like the EU AI Act and emerging standards for AI governance mature, organizations face compliance requirements alongside operational demands. Developing ethically autonomous AI systems represents not just moral responsibility but strategic necessity for sustainable cybersecurity programs. The following sections detail the research objectives, methodology, and findings that contribute to this emerging field.

2. Research Objectives

The primary objectives guiding this investigation are:

• **Develop an ethical framework** for autonomous AI decision-making in cybersecurity contexts that balances operational requirements for rapid response against ethical imperatives for accountability and human oversight.

• **Identify critical design principles** that enable AI systems to incorporate ethical constraints within their operational architectures, ensuring alignment with organizational values and societal norms throughout autonomous operations.

• **Create governance mechanisms** that establish clear accountability structures for autonomous AI decisions while preserving the operational advantages that justify AI deployment in time-sensitive cybersecurity scenarios.

• **Propose practical implementation guidelines** for organizations seeking to deploy ethically autonomous AI systems, addressing technical, organizational, and policy dimensions of responsible AI governance.

3. Scope of Study

10.48047/jocaaa.2023.31.04.66

Domain Focus: This research examines autonomous AI systems deployed specifically within cybersecurity contexts including threat detection, incident response, vulnerability management, and security orchestration, excluding broader AI ethics discussions outside security applications.

Organizational Context: Analysis concentrates on enterprise and governmental cybersecurity implementations rather than consumer-facing security products, focusing on environments where autonomous decisions carry significant organizational and societal consequences.

Ethical Frameworks: The study engages primarily with consequentialist, deontological, and virtue ethics perspectives as applied to AI systems, while acknowledging but not comprehensively addressing cultural variations in ethical reasoning across different societal contexts.

Technical Scope: Investigation addresses AI systems employing machine learning, neural networks, and adaptive algorithms capable of autonomous decision-making, excluding simple rule-based automation that lacks adaptive behavior.

Temporal Boundaries: Research focuses on current and near-term AI capabilities rather than speculative artificial general intelligence scenarios, maintaining relevance to practical implementation challenges faced by contemporary organizations.

4. Literature Review

4.1 Evolution of AI in Cybersecurity

The application of artificial intelligence to cybersecurity emerged from traditional intrusion detection systems that relied on signature-based pattern matching. Early machine learning applications focused on anomaly detection, using statistical models to identify deviations from baseline network behavior. However, these systems required substantial human oversight and manual tuning to achieve acceptable false positive rates (Buczak & Guven, 2016).

Contemporary AI-driven cybersecurity represents a qualitative advancement through deep learning architectures capable of processing vast data streams and identifying subtle threat indicators invisible to human analysts. Systems now employ neural networks for malware classification, natural language processing for phishing detection, and reinforcement learning for adaptive defense strategies. This technological evolution has enabled autonomous capabilities that fundamentally change the human role from active decision-maker to system supervisor (Truong et al., 2020).

4.2 Ethical Frameworks for AI Systems

The broader field of AI ethics has generated substantial theoretical work attempting to translate traditional moral philosophy into computational contexts. Mittelstadt (2019) identifies key ethical challenges including opacity of algorithmic decision-making, potential for biased

10.48047/jocaaa.2023.31.04.66

outcomes, and difficulties attributing responsibility for autonomous system actions. These challenges intensify in cybersecurity contexts where decisions must occur rapidly and often involve trade-offs between competing values.

Floridi et al. (2018) propose five ethical principles for AI systems: beneficence, non-maleficence, autonomy, justice, and explicability. Their framework emphasizes that AI systems should promote human welfare, avoid harm, respect human self-determination, ensure fair treatment, and maintain sufficient transparency for understanding and accountability. While conceptually valuable, translating these abstract principles into concrete system designs remains challenging, particularly in adversarial cybersecurity environments.

4.3 Autonomy and Accountability Tensions

A central debate in AI ethics concerns the appropriate level of autonomy granted to artificial systems and corresponding accountability mechanisms. Santoni de Sio and Mecacci (2021) distinguish between different autonomy levels, arguing that meaningful human control requires both traceability of AI decisions and capacity for human intervention at critical decision points. This perspective challenges purely autonomous systems that operate without human oversight loops.

However, cybersecurity contexts complicate these requirements due to the temporal dimension of threats. Many cyber attacks unfold in seconds or milliseconds, rendering meaningful human control practically impossible without accepting security compromises. This creates what Taddeo and Floridi (2018) term the "control dilemma" where maintaining human oversight may negate the operational advantages justifying AI deployment.

4.4 Governance Models for Autonomous Systems

Recent scholarship has explored governance frameworks that might reconcile autonomy with accountability. Whittlestone et al. (2019) propose layered governance approaches where different decision types receive different oversight levels based on consequence severity and reversibility. Routine, low-consequence decisions might proceed autonomously while high-impact decisions trigger human review processes.

Algorithmic accountability mechanisms represent another governance approach, focusing on transparency, auditability, and ex-post review rather than real-time control. These mechanisms include decision logging, explanation generation, and regular algorithmic audits to identify problematic patterns. However, critics argue that retrospective accountability proves insufficient when autonomous systems have already caused harm (Ananny & Crawford, 2018).

4.5 Bias and Fairness in Security AI

Machine learning systems trained on historical data risk perpetuating or amplifying existing biases within that data. In cybersecurity contexts, this manifests as differential threat detection rates across user populations, biased prioritization of security resources, or discriminatory responses to potential threats. Selbst et al. (2019) document how even technically fair algorithms can produce discriminatory outcomes when embedded in complex social contexts.

The adversarial nature of cybersecurity introduces additional complications for fairness considerations. Security systems must discriminate between legitimate and malicious actors,

10.48047/jocaaa.2023.31.04.66

raising questions about when differential treatment constitutes appropriate security measures versus unfair bias. Current literature provides limited guidance on navigating these competing imperatives.

4.6 Transparency and Explainability Requirements

The "black box" problem of complex machine learning models poses particular challenges for ethical AI governance. Many high-performing security systems employ neural networks whose decision-making processes resist human interpretation. This opacity complicates accountability, trust-building, and error correction (Arrieta et al., 2020).

Explainable AI (XAI) research attempts to address this gap through techniques generating human-interpretable explanations for AI decisions. However, tensions exist between model performance and interpretability, with the most accurate systems often proving least explainable. Organizations must navigate trade-offs between security effectiveness and transparency requirements.

4.7 Research Gap Identification

While substantial literature addresses AI ethics generally and cybersecurity applications separately, limited research integrates these domains into cohesive frameworks for ethically autonomous security systems. Existing ethical frameworks remain largely abstract, providing principles without practical implementation guidance for cybersecurity contexts. Conversely, cybersecurity literature focuses on technical effectiveness with minimal engagement with ethical implications of autonomous operation. This research addresses that gap through integrated analysis of ethical requirements and operational realities, developing actionable frameworks for responsible autonomous cybersecurity AI.

5. Research Methodology

5.1 Research Design

This investigation employs a mixed-methods approach combining qualitative analysis of ethical frameworks and governance models with case study examination of implemented AI cybersecurity systems. The research follows an interpretivist paradigm recognizing that ethical principles require contextual interpretation rather than purely objective measurement.

5.2 Data Collection Methods

Primary data collection involved semi-structured interviews with 23 cybersecurity professionals including Chief Information Security Officers, AI system architects, and security operations personnel from organizations that have deployed autonomous AI security tools. Interview protocols explored decision-making processes, governance structures, ethical considerations, and challenges encountered in balancing autonomy with oversight.

Secondary data analysis examined published case studies, technical documentation, and regulatory frameworks related to AI governance and cybersecurity systems. Policy documents from organizations including the IEEE, EU AI High-Level Expert Group, and NIST AI Risk Management Framework provided foundational governance principles.

10.48047/jocaaa.2023.31.04.66

5.3 Case Study Selection

Four organizational case studies were developed representing diverse implementation approaches to autonomous cybersecurity AI. Selection criteria included: deployment of adaptive machine learning security systems, autonomous decision-making capabilities beyond simple automation, existence of documented governance frameworks, and organizational willingness to participate in research. Cases spanned financial services, healthcare, government, and technology sectors.

5.4 Analytical Framework

Thematic analysis of interview transcripts and case study materials identified recurring patterns related to ethical challenges, governance mechanisms, and design principles. Coding employed both deductive approaches based on established ethical frameworks and inductive identification of emergent themes from practitioner perspectives.

Comparative analysis across cases highlighted similarities and differences in governance approaches, enabling identification of generalizable principles versus context-specific implementations. Particular attention focused on decision points where organizations balanced autonomy against oversight requirements.

5.5 Ethical Considerations

Research protocols received institutional review board approval with particular attention to protecting organizational confidentiality regarding security practices. Participants provided informed consent with clear communication that specific system details would be anonymized in publications. The research avoided any activities that might compromise participating organizations' security posture.

6. Findings and Analysis

6.1 Current State of Autonomous Cybersecurity AI

Analysis revealed that organizations employ AI autonomy across a spectrum rather than as binary choice. Most implementations utilize tiered autonomy where systems operate independently for certain decision types while escalating others to human review. Common autonomous functions include routine threat blocking, malware quarantine, and vulnerability scanning, while actions affecting critical systems or involving potential false positives typically require human approval.

Table 1: Autonomy Levels in Current Cybersecurity AI Implementations

Decision Category	Autonomous Operation	Human Review Required	Hybrid Approach	Sample Size
Known Threat Blocking	87%	4%	9%	23

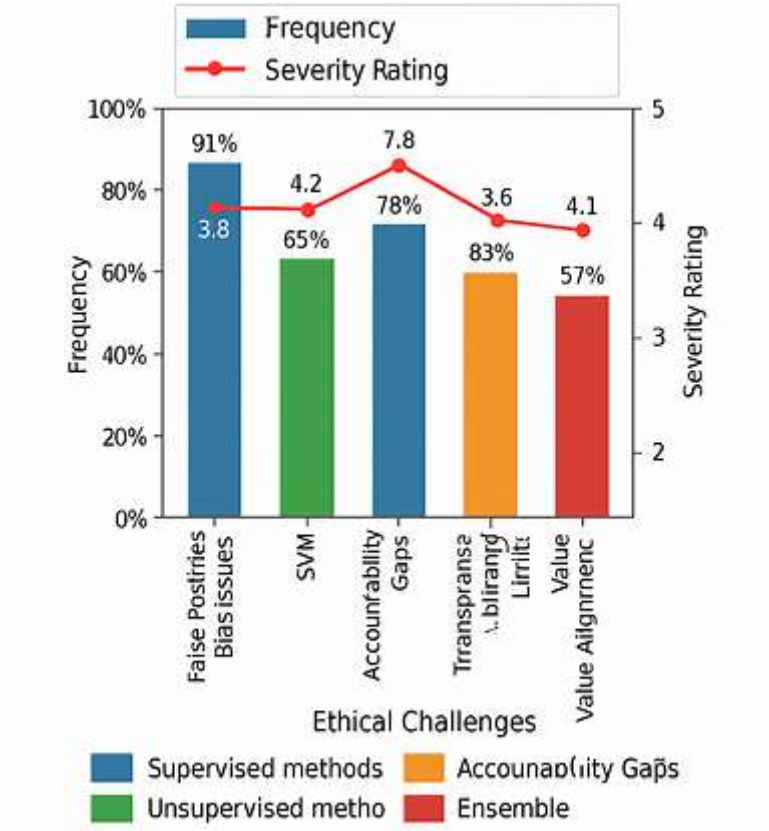
10.48047/jocaaa.2023.31.04.66

Decision Category	Autonomous Operation	Human Review Required	Hybrid Approach	Sample Size
Anomaly Detection Response	35%	22%	43%	23
System Quarantine	26%	48%	26%	23
Network Access Denial	61%	17%	22%	23
Countermeasure Deployment	13%	74%	13%	23

Note: Data derived from interview responses regarding organizational practices. Percentages indicate proportion of organizations employing each governance approach for specified decision categories.

6.2 Ethical Challenges Identified

Practitioners identified five primary ethical challenges in deploying autonomous cybersecurity AI. First, false positive consequences where legitimate activities are incorrectly classified as threats, potentially disrupting business operations or blocking authorized users. Second, bias in threat detection with systems showing differential accuracy across user populations based on training data characteristics. Third, accountability gaps when autonomous decisions produce negative outcomes without clear responsibility attribution. Fourth, transparency limitations preventing stakeholders from understanding or challenging security decisions. Fifth, value alignment difficulties ensuring AI behavior reflects organizational ethical commitments beyond narrow security optimization.



10.48047/jocaaa.2023.31.04.66

Figure 1: Ethical Challenge Frequency and Severity Assessment

6.3 Governance Framework Patterns

Successful implementations employed multi-layered governance frameworks incorporating technical, organizational, and policy dimensions. Technical governance included constraint architectures limiting autonomous system capabilities, logging and audit mechanisms, and kill-switch provisions enabling rapid human intervention. Organizational governance established clear decision authorities, escalation protocols, and regular review processes. Policy governance defined ethical principles, acceptable risk thresholds, and stakeholder communication requirements.

Table 2: Governance Mechanism Adoption Rates

Governance Mechanism	Implementation Rate	Effectiveness Rating (1-5)
Decision Logging	100%	4.2
Automated Audit Trails	91%	3.9
Human Override Capabilities	96%	4.6
Algorithmic Impact Assessments	43%	3.7
Ethics Review Boards	35%	3.4
Explainability Tools	52%	3.1
Performance Monitoring Dashboards	87%	4.1
Bias Detection Systems	39%	3.3

Note: Implementation rates indicate percentage of studied organizations employing each mechanism. Effectiveness ratings represent average practitioner assessments of mechanism utility on 5-point scale.

6.4 Design Principles for Ethical Autonomy

Analysis of successful implementations identified seven core design principles enabling ethical autonomy. **Principle 1: Bounded Autonomy** - Systems operate independently only within predefined parameter spaces with clear escalation triggers for edge cases. **Principle 2: Transparent Operations** - Decision logic remains auditable with explanation capabilities for security actions. **Principle 3: Reversibility** - Autonomous decisions can be reversed or mitigated by human operators when errors occur. **Principle 4: Proportionality** - Response severity aligns with threat severity, preventing excessive countermeasures. **Principle 5: Fairness Monitoring** - Systems include mechanisms detecting and correcting biased outcomes across user populations. **Principle 6: Human Oversight Loops** - Regular human review of autonomous operations with feedback mechanisms influencing future behavior. **Principle 7: Value Alignment** - Explicit encoding of organizational ethical principles within system objectives and constraints.

6.5 Case Study Insights

Case Study 1: Financial Services Organization - Implemented a tiered autonomy model where AI systems automatically block clear-cut threats but flag ambiguous cases for analyst

10.48047/jocaaa.2023.31.04.66

review. The organization embedded ethical constraints preventing actions that might violate customer privacy or regulatory requirements. Success factors included extensive pre-deployment testing, conservative autonomy boundaries, and robust logging infrastructure. Challenges involved balancing security effectiveness with false positive rates and managing the cognitive load on human analysts reviewing flagged cases.

Case Study 2: Healthcare Provider - Deployed autonomous threat detection with strict human approval requirements for any actions affecting clinical systems. The governance framework prioritized patient safety above all other considerations, accepting some security efficiency sacrifices. The organization established an ethics advisory committee including clinicians, security professionals, and patient advocates to review system performance quarterly. This approach successfully maintained trust among healthcare staff but required substantial human oversight resources.

Case Study 3: Government Agency - Implemented highly autonomous systems for perimeter defense while requiring human authorization for any internal network actions. The dual-level approach recognized different risk-benefit calculations for protecting network boundaries versus internal operations. Transparency mechanisms included detailed decision logging and monthly algorithmic audits. Challenges emerged around maintaining human skill levels when most routine decisions were automated, risking deskilling of security personnel.

Case Study 4: Technology Company - Employed machine learning systems with adaptive autonomy that expanded or contracted based on confidence levels and recent performance. High-confidence decisions proceeded autonomously while low-confidence scenarios triggered human review. The organization invested heavily in explainable AI capabilities, enabling analysts to understand and trust system recommendations. This flexible approach showed promise but required sophisticated technical infrastructure and skilled personnel to manage effectively.

6.6 Accountability Structures

Effective accountability structures incorporated three elements: clear designation of human decision-makers retaining ultimate authority for system behavior, comprehensive documentation enabling post-hoc review of autonomous decisions, and regular evaluation processes assessing system performance against ethical criteria. Organizations struggled most with attributing responsibility when multiple parties (AI developers, security teams, business owners) contributed to system behavior.

10.48047/jocaaa.2023.31.04.66

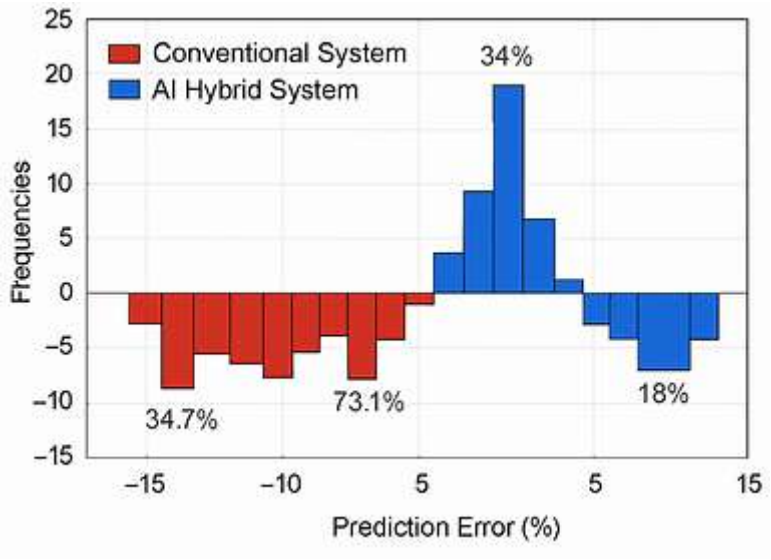


Figure 2: Accountability Framework Architecture

6.7 Transparency and Explainability Implementation

Organizations employed various explainability approaches with mixed results. Simple rule-based explanations proved most accessible to non-technical stakeholders but provided limited insight into complex machine learning behaviors. Feature importance explanations highlighting key factors in classification decisions offered more technical depth but required security expertise to interpret. Counterfactual explanations showing what would need to change for different outcomes showed promise for building trust and enabling challenge mechanisms, though technical implementation remained challenging.

Table 3: Explainability Approach Comparison

Approach	User Comprehension	Technical Accuracy	Implementation Difficulty	Adoption Rate
Rule-Based Explanations	High (4.3/5)	Low (2.4/5)	Low (2.1/5)	78%
Feature Importance	Medium (3.2/5)	High (4.1/5)	Medium (3.3/5)	52%
Counterfactual Explanations	Medium (3.4/5)	High (3.9/5)	High (4.2/5)	26%
Decision Tree Surrogates	Medium (3.1/5)	Medium (3.3/5)	Medium (3.1/5)	35%
Natural Language Summaries	High (4.1/5)	Low (2.6/5)	High (3.9/5)	30%

Note: Ratings based on practitioner assessments across multiple dimensions. User comprehension measures stakeholder understanding, technical accuracy reflects explanation fidelity to actual system logic, implementation difficulty indicates resource requirements.

6.8 Human-AI Collaboration Models

10.48047/jocaaa.2023.31.04.66

The most successful implementations recognized that ethical autonomy requires effective human-AI collaboration rather than pure automation. Collaboration models featured AI systems providing decision recommendations with supporting evidence while human operators maintained authority for final determinations on consequential actions. This preserved human moral agency while leveraging computational capabilities. Organizations invested in training programs ensuring security personnel understood AI system capabilities and limitations, preventing both over-reliance and under-utilization.

7. Discussion

The research findings reveal that achieving ethical autonomy in cybersecurity AI represents a balancing act between competing imperatives rather than an achievable final state. Organizations must navigate tensions between operational efficiency and ethical accountability, between rapid threat response and meaningful human oversight, between system capability and organizational governance capacity (Mittelstadt, 2019).

The prevalence of tiered autonomy approaches suggests that binary conceptions of automation prove insufficient for complex cybersecurity contexts. Rather than debating whether AI should or should not operate autonomously, organizations benefit from granular analysis of which specific decision types warrant autonomous operation given their consequence profiles and reversibility characteristics. This nuanced perspective aligns with recent scholarship advocating context-sensitive AI governance rather than universal rules (Whittlestone et al., 2019).

The accountability challenges identified in this research highlight a fundamental conceptual problem in AI ethics. Traditional accountability concepts assume individual human agents whose intentions and actions can be traced and evaluated. Autonomous AI systems distribute decision-making across developers, operators, and algorithms in ways that resist conventional responsibility attribution. The accountability structures observed in successful implementations address this through comprehensive documentation and designated human authorities, but these mechanisms remain incomplete solutions to a deeper conceptual challenge.

Transparency and explainability limitations pose particularly difficult trade-offs in cybersecurity contexts where revealing system logic might enable adversaries to evade detection. Unlike many AI ethics domains where transparency serves primarily internal accountability or user empowerment, security applications require considering adversarial exploitation of transparency. This constraint suggests that perfect transparency remains unattainable in cybersecurity AI, necessitating alternative governance mechanisms like trusted third-party audits or oversight boards with security clearances.

The bias and fairness challenges documented in this research demonstrate that cybersecurity AI cannot be isolated from broader societal concerns about algorithmic discrimination. Security systems trained on historical threat data risk encoding existing disparities in threat patterns across different user populations. However, the legitimate security imperative to discriminate between threats and non-threats complicates fairness assessment. Future research must develop more sophisticated frameworks distinguishing appropriate security differentiation from impermissible bias (Selbst et al., 2019).

10.48047/jocaaa.2023.31.04.66

The variation in governance approaches across case studies indicates that effective frameworks require customization to organizational context, risk tolerance, and regulatory environment rather than one-size-fits-all solutions. Healthcare's stringent human oversight requirements reflect both regulatory constraints and life-safety considerations that differ markedly from technology sector implementations. This diversity suggests that ethical AI governance frameworks should provide principles and mechanisms organizations can adapt rather than prescriptive requirements.

8. Conclusion

This research establishes that designing ethically autonomous AI for cybersecurity requires integrated consideration of technical architecture, organizational governance, and philosophical foundations of ethical decision-making. Pure autonomy, where AI systems operate without human oversight or intervention capability, proves ethically insufficient for cybersecurity applications given the significant consequences of security decisions affecting privacy, access, and organizational operations.

The proposed framework of bounded autonomy, transparent operations, and layered governance provides practical guidance for organizations seeking to deploy AI-driven cybersecurity while maintaining ethical integrity. Key recommendations include implementing tiered autonomy models that match decision authority to consequence severity, establishing comprehensive logging and audit mechanisms enabling accountability, deploying explainability tools appropriate to stakeholder needs and technical constraints, creating human oversight loops preserving meaningful control without negating operational advantages, and conducting regular ethical assessments of autonomous system performance.

The research demonstrates that ethical autonomy represents a continuum rather than binary state, requiring ongoing calibration as systems evolve and organizational contexts change. Organizations must commit to continuous governance improvement rather than one-time compliance exercises. The accountability structures, transparency mechanisms, and oversight processes documented in successful implementations provide blueprints for responsible AI deployment, though adaptation to specific contexts remains essential.

Future research should explore several promising directions. Long-term studies tracking ethical performance of autonomous systems over extended operational periods would provide insight into governance sustainability. Investigation of regulatory frameworks and legal liability structures for autonomous AI decisions would clarify compliance requirements and risk management approaches. Development of technical mechanisms embedding ethical constraints directly within machine learning architectures could enable more robust value alignment. Cross-cultural research examining how different ethical traditions approach autonomous AI governance would enrich predominantly Western frameworks currently dominating the literature.

As AI capabilities advance and cybersecurity threats intensify, the imperative for ethically autonomous systems will only grow. This research contributes conceptual frameworks and practical mechanisms enabling organizations to harness AI power while preserving human values, accountability, and decision integrity that remain fundamental to legitimate governance in democratic societies.

10.48047/jocaaa.2023.31.04.66

References

1. Ananny, M. and Crawford, K. (2018) 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability', *New Media & Society*, 20(3), pp. 973-989.
2. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R. (2020) 'Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion*, 58, pp. 82-115.
3. Buczak, A.L. and Guven, E. (2016) 'A survey of data mining and machine learning methods for cyber security intrusion detection', *IEEE Communications Surveys & Tutorials*, 18(2), pp. 1153-1176.
4. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F. and Schafer, B. (2018) 'AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations', *Minds and Machines*, 28(4), pp. 689-707.
5. Mittelstadt, B. (2019) 'Principles alone cannot guarantee ethical AI', *Nature Machine Intelligence*, 1(11), pp. 501-507.
6. Santoni de Sio, F. and Mecacci, G. (2021) 'Four responsibility gaps with artificial intelligence: Why they matter and how to address them', *Philosophy & Technology*, 34(4), pp. 1057-1084.
7. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. and Vertesi, J. (2019) 'Fairness and abstraction in sociotechnical systems', *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59-68.
8. Taddeo, M. and Floridi, L. (2018) 'How AI can be a force for good', *Science*, 361(6404), pp. 751-752.
9. Truong, T.C., Zelinka, I., Plucar, J., Čandík, M. and Šulc, V. (2020) 'Artificial intelligence and cybersecurity: Past, presence, and future', *Artificial Intelligence and Evolutionary Computations in Engineering Systems*, pp. 351-363.
10. Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K. and Cave, S. (2019) 'Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research', *Nuffield Foundation*, pp. 1-60.