# Ethical and Legal Considerations in AI-Based Cybersecurity Solutions

Pou Vanthana

Bachelor of Information Technology Engineering (ITE), Department of Information Technology Engineering, RUPP, Phnom Penh, Cambodia

**Cybersecurity Law and Policy Review**

*Volume 1, Issue 1 | 2025 | https://clprj.com/journal/index*

**Abstract**

Artificial intelligence has emerged as the defining technology of contemporary cybersecurity practice, with AI-powered tools now deployed across the full spectrum of defensive and offensive security functions: threat detection and anomaly identification, malware analysis, vulnerability discovery, intrusion prevention, identity and access management, threat intelligence synthesis, and automated incident response. The integration of AI into cybersecurity operations has generated genuine and substantial defensive benefits — enabling detection at speeds and scales that human analysts cannot achieve, identifying patterns invisible to conventional rule-based systems, and enabling proactive defence against evolving threat actor tactics. Yet AI-based cybersecurity solutions simultaneously introduce a set of ethical and legal challenges whose significance has not been adequately addressed by the scholarly literature or by emerging AI governance frameworks.

**Keywords:** *artificial intelligence; cybersecurity; AI ethics; algorithmic bias; explainability; autonomous cyber operations; privacy; liability; EU AI Act; GDPR; machine learning; threat detection; responsible AI; international law*

## 1. INTRODUCTION

The deployment of artificial intelligence in cybersecurity is no longer a prospective development to be considered by forward-looking regulators — it is a present and pervasive reality that is reshaping both the capabilities and the legal and ethical landscape of the security domain with a speed that governance frameworks have struggled to match. Across the public and private sectors, AI-powered security tools are already making — or substantially influencing — decisions of enormous consequence: determining whether a network connection is malicious and should be blocked; classifying whether a user's behaviour indicates account compromise; prioritising which vulnerabilities in a complex system require urgent remediation; generating threat intelligence assessments that inform national security decision-making; and, in the most advanced deployments, autonomously executing responses to identified threats without human authorisation of individual actions.

The security benefits of AI deployment in cybersecurity are genuine and significant. The volume of security events generated by modern enterprise networks — billions of log entries, millions of alerts, and thousands of potential indicators of compromise per day in large organisations — is simply beyond the processing capacity of human security operations teams working with conventional rule-based tools. AI systems trained on threat



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

intelligence data can identify patterns indicative of sophisticated attack techniques that evade signature-based detection, correlate signals across disparate data sources that human analysts would not associate, and respond to identified threats in milliseconds rather than the hours or days that manual processes require. In the context of a threat landscape characterised by the professionalisation of ransomware-as-a-service, the deployment of AI by offensive actors to automate attack discovery and execution, and the increasing sophistication of nation-state cyber operations, the defensive application of AI is not merely beneficial but increasingly necessary.

Yet AI-based cybersecurity solutions are not ethically or legally neutral instruments, and the urgency of the defensive imperative that drives their adoption does not suspend the ethical and legal framework within which they must operate. The opacity of AI decision-making — the fundamental difficulty, in many AI architectures, of understanding why a particular output was generated — creates accountability gaps that are ethically problematic in any high-stakes decision context and legally significant in contexts where affected parties have rights to explanation, contestation, and redress. The training data on which AI security systems are built reflects the historical biases of the security domain — including biases that associate certain types of network behaviour, certain geographic IP ranges, or certain user characteristics with threat activity in ways that may produce discriminatory outcomes. The deployment of AI in offensive cyber operations — whether by state actors conducting military cyber operations or by private sector actors conducting 'hack-back' defensive operations — raises profound questions about legal responsibility, proportionality, and the applicability of the laws of armed conflict to machine-executed attacks. And the mass surveillance of network behaviour and user activity that AI-powered security requires is in fundamental tension with privacy rights that data protection law and human rights instruments seek to protect.

## 2. ARTIFICIAL INTELLIGENCE IN CYBERSECURITY: CAPABILITIES, ARCHITECTURES, AND LIMITATIONS

### *2.1 The AI-Powered Security Lifecycle*

AI technologies are now deployed across virtually every phase of the cybersecurity lifecycle, from pre-attack vulnerability management through active threat detection and incident response to post-incident forensics and threat intelligence enrichment. In the vulnerability management and pre-attack phase, AI systems are used to prioritise the remediation of identified vulnerabilities based on predicted exploitability and organisational risk exposure, to analyse code for security weaknesses through automated static and dynamic analysis, and to simulate attack paths through enterprise networks to identify the most significant security gaps. The application of large language models (LLMs) to code security analysis — exemplified by tools including GitHub Copilot's security features and commercial code analysis platforms — has substantially expanded the scale and accessibility of automated vulnerability detection.

In the threat detection and monitoring domain — arguably the most mature area of AI application in cybersecurity — machine learning models are deployed in security information and event management (SIEM) systems, endpoint detection and response



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

(EDR) platforms, network detection and response (NDR) tools, and user and entity behaviour analytics (UEBA) systems to identify anomalies, classify potential threats, and correlate indicators of compromise across large and heterogeneous datasets. The supervised learning models that underpin many commercial security products are trained on labelled datasets of known malicious and benign activity; unsupervised learning and anomaly detection approaches identify deviations from established behavioural baselines without requiring pre-labelled training data; and reinforcement learning approaches are beginning to be applied to adaptive threat response systems that improve their responses over time through experience.

### *2.2 Offensive AI and the Dual-Use Challenge*

The same AI capabilities that power defensive security applications are simultaneously being deployed by threat actors to enhance the sophistication, scale, and effectiveness of offensive operations — a dual-use dynamic that fundamentally complicates the governance of AI in the security domain. AI-powered phishing and social engineering tools — including LLM-based systems capable of generating highly personalised, contextually accurate, grammatically flawless phishing communications at industrial scale — have substantially raised the floor of attacker capability in the social engineering domain. The generative AI-enabled production of synthetic audio and video ('deepfakes') of targeted individuals — including corporate executives, government officials, and trusted personal contacts — enables impersonation attacks of a quality and scale that were not achievable with pre-AI technology.

On the technical attack surface, AI is being applied to the automated discovery of software vulnerabilities, the generation of working exploits for identified vulnerabilities, the evasion of AI-powered security defences through adversarial machine learning techniques, and the optimisation of malware to evade specific detection architectures. The application of reinforcement learning to penetration testing

### *2.3 Limitations and Failure Modes*

AI-based cybersecurity systems exhibit characteristic failure modes that are legally and ethically significant and that are not always adequately disclosed by vendors or understood by deploying organisations. False positive rates — the generation of security alerts for benign activity — impose significant operational costs in AI-powered security operations and, where alerts trigger automated responses including network isolation or account lockout, can cause direct harm to legitimate users and business operations. False negative rates — the failure to detect genuine malicious activity — may be more serious in their security consequences but are less visible operationally, creating an over-reliance on AI detection systems whose limitations are not adequately understood.

Distribution shift — the degradation of AI model performance when deployed on data that differs from the training distribution — is a fundamental limitation of machine learning systems that has particular significance in the cybersecurity domain, where the threat landscape evolves rapidly and the tactics, techniques, and procedures of threat actors change continuously in response to defensive improvements. An AI security model trained on data from one organisational environment may perform poorly when deployed in a different environment with different network architectures, user behaviours, and threat profiles. The temporal dimension of distribution shift — in which models trained on



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

historical data become increasingly stale as the current threat landscape diverges from historical patterns — creates an ongoing model maintenance obligation that many deploying organisations underestimate.

## 3. OPACITY, EXPLAINABILITY, AND ACCOUNTABILITY IN AI CYBERSECURITY

### *3.1 The Black Box Problem in Security Contexts*

The opacity of AI decision-making — the inability to understand, in human-interpretable terms, why a particular AI system produced a particular output — is a general challenge of machine learning that assumes heightened legal and ethical significance in the cybersecurity context. When an AI security system classifies a user's behaviour as indicative of account compromise and triggers an automated response that locks the user out of critical systems, terminates their network access, or initiates an investigation that affects their employment, the affected individual has a legitimate interest in understanding the basis for that decision. When an AI threat intelligence system generates an attribution assessment that informs a government's response to a perceived cyberattack — potentially including diplomatic, economic, or military countermeasures — the opacity of the underlying analytical process raises profound accountability concerns.

The legal significance of AI opacity in cybersecurity is most directly engaged by the GDPR's Article 22, which establishes the right of individuals not to be subject to decisions based solely on automated processing that produce legal or similarly significant effects, and Article 13-15's requirements for meaningful information about the logic involved in automated decision-making. The applicability of Article 22 to AI-driven security decisions — including automated account suspension, network access revocation, or fraud flagging — has been the subject of regulatory guidance from the European Data Protection Board (EDPB) and national supervisory authorities, which have confirmed that such decisions can constitute 'legal or similarly significant effects' within the meaning of Article 22 where they substantially affect the individual's ability to exercise rights, access services, or pursue legitimate activities.

### *3.2 Explainable AI in Security Operations*

The development of explainable AI (XAI) techniques — approaches that generate human-interpretable explanations of AI system outputs — represents a significant and growing research and development investment in the AI security domain, driven by both the governance requirements of emerging AI regulation and the operational need of security analysts to understand and act on AI-generated alerts effectively. Techniques including LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), attention mechanisms in transformer architectures, and rule extraction from trained models are being applied to generate post-hoc explanations of AI security system outputs that enable analysts to understand which features of an observed event contributed to an alert classification.

The legal and operational adequacy of current XAI techniques in the cybersecurity context is, however, significantly limited. The explanations generated by post-hoc interpretability



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

methods are approximations of the underlying model's decision process — not expositions of the true computational basis for the decision — and may be misleading in ways that are difficult for non-expert users to detect. The trade-off between model complexity (which generally correlates with predictive performance) and interpretability (which is easier to achieve in simpler models) means that the most capable AI security systems are frequently the least interpretable. And the explanations provided by XAI techniques may satisfy a formal compliance requirement for 'meaningful information about the logic involved' without providing the substantive understanding that would enable an affected individual to genuinely contest an AI-driven security decision.

### *3.3 Human Oversight and the Accountability Framework*

The legal and ethical imperative of meaningful human oversight of AI security decisions — particularly those with significant consequences for individuals — is widely affirmed in AI governance frameworks including the EU AI Act, the OECD AI Principles, and the US Executive Order on AI. The implementation of this imperative in AI security operations is, however, significantly complicated by the volume and velocity of AI-generated security outputs, which may number in the thousands per day in large security operations centres and which are generated at speeds that make genuine human review of individual decisions impractical without substantially reducing the operational benefits of AI deployment.

## 4. BIAS, FAIRNESS, AND DISCRIMINATION IN AI CYBERSECURITY SYSTEMS

### *4.1 Sources of Bias in AI Security Systems*

AI security systems are susceptible to bias from multiple sources, with potentially significant consequences for the fairness and legality of security decisions applied differentially across demographic, organisational, and geographic groups. Training data bias — the most extensively studied source of algorithmic bias — arises where the datasets on which AI security models are trained do not adequately represent the full population of entities whose behaviour the system will assess, or where historical labelling of security incidents reflects the biases of the human analysts who generated the labels. In the cybersecurity domain, training data is frequently sourced from large enterprise environments with predominantly Western, English-speaking user populations, creating models that may systematically misclassify the behaviour of users from different cultural, linguistic, or organisational contexts as anomalous or suspicious.

The geographic bias embedded in many AI threat detection systems is particularly documented and consequential. IP reputation and geolocation-based threat intelligence — which forms part of the feature set of many AI security systems — assigns elevated risk scores to traffic from certain geographic regions based on historical association with malicious activity. The logical circularity of this approach — in which regions with weaker cybersecurity infrastructure attract disproportionate use by criminal actors, generating training data that leads AI systems to flag legitimate traffic from those regions as suspicious, which reduces legitimate user access to services and concentrations of observed legitimate activity — is well understood analytically but poorly addressed in current AI security practice. The practical consequence is that users in developing countries, or users



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

whose legitimate internet activity routes through infrastructure in regions with elevated threat intelligence scores, may face systematic disadvantage in AI-mediated security decisions — an outcome that raises both fairness concerns and, where it amounts to differential treatment on grounds protected by applicable law, potential discrimination liability.

### *4.2 Legal Frameworks for Algorithmic Fairness*

The legal framework applicable to algorithmic bias in AI security systems spans data protection law, equality and non-discrimination law, and the specific sector-specific obligations that apply to regulated industries deploying AI security tools. Under the GDPR, the principle of fairness — one of the six data processing principles established by Article 5(1)(a) — requires that personal data be processed in a manner that is fair to the data subject, including in ways that do not produce unjust, discriminatory, or harmful effects. The EDPB's Guidelines on Automated Individual Decision-Making and Profiling (2018) explicitly state that fairness requires that AI systems do not produce decisions that discriminate against individuals on the basis of protected characteristics or that produce differential treatment not justified by legitimate grounds.

The EU AI Act's prohibited AI practices include the use of AI systems that deploy subliminal techniques or that exploit vulnerabilities of groups of persons to materially distort their behaviour in a way likely to cause psychological or physical harm, and AI systems that evaluate or classify natural persons based on social behaviour or personal characteristics in ways that lead to unjustifiable detrimental treatment. The Act's high-risk category requirements include obligations for testing AI systems for accuracy across defined groups of persons to ensure absence of discriminatory outcomes — requirements that, if applied to AI security systems used in employment and workplace monitoring contexts, would impose meaningful non-discrimination obligations on deploying organisations.

## 5. AUTONOMOUS AI IN CYBER OPERATIONS: LEGAL AND ETHICAL DIMENSIONS

### *5.1 The Spectrum of Autonomy in AI Cyber Operations*

The deployment of AI in cyber operations encompasses a spectrum of autonomy that ranges from AI-assisted human decision-making — in which AI systems provide analysis and recommendations but humans retain decision authority — through semi-autonomous operations in which AI executes predefined response actions subject to human override, to fully autonomous operations in which AI systems identify, decide, and execute offensive or defensive actions without human involvement in individual decisions. The legal and ethical implications of AI in cyber operations vary substantially across this spectrum, with the most profound challenges arising at the semi-autonomous and fully autonomous ends where the causal connection between human decision-making and operational consequences is most attenuated.

Fully autonomous offensive AI cyber capabilities — sometimes called 'lethal autonomous cyber weapons' by analogy with the better-studied literature on lethal autonomous weapons



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

systems in kinetic military operations — raise the starkest legal and ethical questions. An autonomous AI system that can, without human authorisation of individual attacks, identify targets, develop exploits, and execute cyber operations against those targets removes human judgement from consequential decisions that international law, military doctrine, and basic ethical principle require to be subject to human responsibility. The difficulty of ensuring that an autonomous AI cyber system can reliably distinguish legitimate military objectives from protected civilian infrastructure — a distinction that requires contextual judgement of a kind that current AI systems cannot reliably exercise — is a fundamental challenge that existing autonomous weapons frameworks have not yet adequately addressed in the cyber domain.

### *5.2 International Humanitarian Law and Autonomous AI Cyber Operations*

International humanitarian law (IHL) — the law of armed conflict — applies to cyber operations conducted in the context of armed conflict and imposes constraints on autonomous AI cyber systems used in such contexts that are analogous to, but distinct from, those applicable to autonomous weapons systems in the kinetic domain. The Tallinn Manual 2.0's analysis of IHL's applicability to cyber operations confirms that the core IHL principles of distinction (between combatants and civilians, and between military objectives and protected civilian infrastructure), proportionality (in the assessment of expected civilian harm relative to anticipated military advantage), and precaution (in the selection and use of means and methods of warfare to minimise civilian harm) apply to cyber operations in armed conflict.

The application of these principles to autonomous AI cyber systems creates significant challenges. The principle of distinction requires that targeting decisions — including decisions about which systems to attack — be based on a judgment about the military status of the target. An autonomous AI system that identifies and attacks systems based on network traffic patterns, IP reputation, or behavioural indicators associated with military activity faces the fundamental challenge that the same systems may simultaneously serve civilian functions — and that the AI system's pattern-matching approach to target identification may not be capable of reliably making the contextual judgements that IHL requires. The principle of proportionality requires an assessment of expected civilian harm that is inherently subjective and contextual — a kind of assessment that current AI systems are not designed to perform and that legal scholarship has generally agreed requires human judgement.

### *5.3 Private Sector Autonomous Cyber Operations and the Law*

The deployment of autonomous or semi-autonomous AI in private sector cyber operations — including the growing category of 'active cyber defence' measures that go beyond passive monitoring to include automated countermeasures against identified attackers — raises a distinct set of legal questions that domestic law, rather than international humanitarian law, primarily governs. In most jurisdictions, the Computer Fraud and Abuse Act (in the United States), the Computer Misuse Act (in the United Kingdom), and equivalent domestic computer crime statutes criminalise unauthorised access to computer systems without exception for defensive purposes — meaning that an automated AI system that responds to a cyberattack by accessing and disrupting the attacking infrastructure may violate domestic criminal law regardless of the legitimacy of its defensive purpose.



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

## 6. PRIVACY, SURVEILLANCE, AND AI-POWERED SECURITY MONITORING

### *6.1 The Mass Surveillance Dimension of AI Security*

AI-powered cybersecurity systems are fundamentally surveillance systems. Their effectiveness depends on the continuous, comprehensive collection and analysis of data about network activity, user behaviour, device status, application usage, communication patterns, and an enormous range of other observable indicators of system state and user activity. The same data collection that enables an AI security system to identify an anomalous login attempt or detect a data exfiltration pattern is, from a privacy law perspective, the monitoring of individuals' activities at a granularity and scale that would have been unimaginable with pre-AI technology — and that raises fundamental questions about the adequacy of the legal frameworks that nominally regulate employee monitoring, network surveillance, and the collection of personal data for security purposes.

The tension between AI security monitoring and data protection law is particularly acute in the employment context, where the GDPR, the EU's proposed Platform Work Directive's monitoring provisions, and domestic employment law in many jurisdictions impose significant constraints on the monitoring of employee activity. GDPR Article 6's requirement that personal data processing be based on a lawful basis — including legitimate interest, which requires that the controller's interests in security monitoring not be overridden by the interests or fundamental rights of employees — creates a framework that requires employers to justify the specific data collection practices of AI security systems, to limit collection to what is necessary for the security purpose, and to provide employees with meaningful information about the monitoring to which they are subject.

### *6.2 State Surveillance and AI Security Tools*

The deployment of AI security tools by state actors — including government cybersecurity agencies, law enforcement authorities, and intelligence services — raises privacy concerns of particular gravity, given the coercive power that states possess and the fundamental human rights at stake when state surveillance is conducted without adequate legal authorisation or oversight. The post-Snowden debate about the scale and adequacy of oversight for government mass surveillance programmes has been substantially supplemented in the AI era by the development of AI-powered capabilities for network traffic analysis, bulk communications data processing, facial recognition and biometric identification, and social graph analysis that substantially exceed the surveillance capabilities available to pre-AI intelligence programmes.

The legal framework governing state use of AI security surveillance tools is established primarily by domestic intelligence and law enforcement legislation — which varies widely in the adequacy of its authorisation procedures, oversight mechanisms, and individual rights protections — supplemented by European Convention on Human Rights and International Covenant on Civil and Political Rights obligations that apply in relevant jurisdictions. The European Court of Human Rights' jurisprudence on Article 8 (right to private life) — most significantly articulated in Big Brother Watch v. United Kingdom (2021), which found the UK's bulk communications data interception regime to violate Article 8 in the absence of adequate safeguards — provides an authoritative framework for



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

assessing the compliance of state AI surveillance programmes with human rights law. The Court's criteria for lawful surveillance — foreseeability, necessity, proportionality, and adequate independent oversight — provide normative anchors for the governance of AI-powered state security surveillance that domestic legislation must implement.

## 7. LIABILITY FOR AI-DRIVEN SECURITY DECISIONS

### *7.1 The Liability Gap in AI Cybersecurity*

The deployment of AI in cybersecurity decision-making creates a liability gap that existing legal frameworks are inadequately equipped to address. When an AI security system makes a decision — whether to classify a transaction as fraudulent, to suspend an employee's network access, to block a legitimate business communication as spam, or to generate a threat attribution that informs a consequential governmental response — and that decision proves incorrect and causes harm, the question of who bears legal responsibility for the harm is typically far less clear than in cases of equivalent harm caused by human decision-making. The combination of the AI system's opacity (making it difficult to identify the specific failure that caused the incorrect decision), the distribution of responsibility across the AI developer, the deploying organisation, and the human operators who may or may not have had meaningful authority to override the AI decision, and the absence of legal rules specifically addressing AI liability, creates a landscape in which harmed parties frequently find it difficult to identify a legally responsible defendant and to establish the elements of a cause of action.

The EU AI Liability Directive, proposed in September 2022 alongside the AI Act as part of the EU's AI governance package, seeks to address the liability gap by establishing a rebuttable presumption of causation in AI liability claims — making it easier for claimants to establish that the AI system's failure caused their harm — and by creating a disclosure obligation requiring AI developers and deployers to provide claimants with access to relevant evidence about the AI system's operation. The proposed directive applies a non-contractual fault-based liability framework to claims arising from AI system failures, but does not create strict liability — meaning that defendants can avoid liability by demonstrating that they complied with applicable standards of care.

### *7.2 Vendor Liability for AI Security Product Failures*

The liability of AI security system vendors for failures of their products is currently constrained by contractual limitation of liability provisions that are standard in commercial software agreements, by the inherent difficulty of establishing that a specific product failure — rather than a configuration error by the deploying organisation — caused a particular security incident, and by the absence of product liability frameworks that clearly apply to software. The forthcoming EU AI Act's requirements for high-risk AI systems — including conformity assessments, technical documentation, and post-market monitoring obligations — create a framework of regulatory compliance obligations for AI security vendors that, if implemented and enforced effectively, will substantially improve the quality and safety of AI security products. However, the regulatory compliance framework does not in itself create private law liability — a defendant who has complied with the AI



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

Act's requirements can still cause harm for which it should be held accountable, and the AI Act's compliance requirements do not substitute for a clear private law liability framework.

## 8. INTERNATIONAL LEGAL DIMENSIONS OF AI-ENABLED CYBER OPERATIONS

### *8.1 AI and State Responsibility in Cyberspace*

The deployment of AI in state cyber operations — including offensive operations conducted by national intelligence and military cyber commands — engages the framework of state responsibility under international law in ways that the increasing autonomy of AI-enabled operations makes both more important and more challenging to apply. The Articles on State Responsibility (ARSIWA) attribute the conduct of state organs and entities exercising elements of governmental authority to the state, creating state responsibility for internationally wrongful AI cyber operations conducted by or attributable to state actors. The due diligence principle — which requires states to take reasonable steps to prevent their territory from being used as a base for cyber operations against other states — applies to AI-enabled operations launched from state-controlled infrastructure as it does to conventional cyber operations.

The increasing autonomy of AI in state cyber operations creates specific challenges for the attribution of internationally wrongful acts. Where an autonomous AI system conducts a cyber operation against foreign critical infrastructure without specific authorisation by a human decision-maker — executing a pre-programmed mission in response to conditions it identifies autonomously — the causal chain between human decision-making and the specific harmful conduct is attenuated in ways that complicate the attribution of responsibility. The legal framework must nevertheless maintain the principle that states bear responsibility for the autonomous operations of AI systems they develop, deploy, and control, even where individual operational decisions are made by the AI system rather than by human operators — otherwise the development of autonomous AI cyber capabilities would enable states to conduct internationally wrongful cyber operations with reduced accountability.

### *8.2 Norms for AI in Cyber Operations*

The international community's efforts to develop norms for responsible state behaviour in cyberspace — reflected in the consensus reports of successive UN Groups of Governmental Experts (GGEs) and the work of the UN Open-Ended Working Group (OEWG) — have not yet specifically addressed the deployment of AI in state cyber operations. The existing voluntary norms — including the norms against attacking critical infrastructure, against conducting cyber operations that damage the computer emergency response capabilities of other states, and requiring states to prevent their territory from being used for internationally wrongful cyber acts — are formulated in terms that do not distinguish between AI-enabled and conventionally executed cyber operations, and their applicability to autonomous AI cyber operations depends on interpretive analysis rather than explicit textual coverage.



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

The development of AI-specific norms for state cyber operations — addressing questions including the permissible degree of autonomy in offensive AI cyber operations, the minimum human oversight requirements for consequential AI-enabled targeting decisions, the obligation to conduct pre-deployment legal reviews of AI cyber capabilities, and the transparency obligations of states regarding the AI technologies deployed in their cyber operations — represents an important agenda for international cybersecurity law and diplomacy that has not yet received the sustained multilateral engagement it deserves. The establishment of a dedicated multilateral working group specifically addressing AI in cyber operations, building on the existing GGE and OEWG frameworks, would provide an institutional vehicle for this norm development.

## 9. A NORMATIVE FRAMEWORK FOR THE GOVERNANCE OF AI IN CYBERSECURITY

### *9.1 Principles for Ethical AI in Cybersecurity*

The governance of AI in cybersecurity should be grounded in a set of principles that take seriously both the defensive imperative that drives AI adoption and the ethical and legal constraints within which that adoption must occur. The following principles, drawn from the analysis in the preceding parts, provide a normative foundation for AI cybersecurity governance. First, the principle of proportionate transparency: AI security systems should provide transparency appropriate to the significance of the decisions they support, with more consequential decisions requiring greater explainability and human oversight. Second, the principle of fairness by design: AI security systems should be designed and evaluated to identify and mitigate systematic bias, with pre-deployment fairness testing and ongoing monitoring for discriminatory effects. Third, the principle of meaningful human control: human oversight of AI security decisions should be substantive rather than formal, with human reviewers having the expertise, information, and authority to genuinely evaluate and override AI recommendations.

Fourth, the principle of data minimisation: AI security systems should collect and process only the personal data that is necessary and proportionate to the defined security purpose, with privacy impact assessments conducted before deployment and privacy-enhancing techniques applied where consistent with security objectives. Fifth, the principle of accountability: responsibility for the consequences of AI security decisions should be clearly allocated — between AI vendors and deploying organisations — in a manner that creates genuine incentives for security, fairness, and compliance with applicable law.

### *9.2 Regulatory Measures*

Translating these principles into regulatory measures requires action across multiple levels of governance. At the product level, the EU AI Act's high-risk classification should be applied clearly and comprehensively to AI security systems used in high-stakes contexts including critical infrastructure protection, employment monitoring, and national security operations, ensuring that these systems are subject to the Act's most demanding conformity assessment, bias testing, transparency, and documentation requirements. The AI Act's requirements should be supplemented by sector-specific cybersecurity AI standards developed by ENISA and national cybersecurity agencies, addressing the specific technical



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

and operational characteristics of AI security systems in ways that the AI Act's horizontal framework cannot fully capture.

At the liability level, the revised EU Product Liability Directive and the proposed AI Liability Directive should be implemented in a manner that closes the accountability gap for AI security system failures, creating genuine financial incentives for AI vendors to develop products that are safe, fair, and effective. The development of mandatory AI security product testing and certification schemes — analogous to the certification requirements for other safety-critical software — would provide an independent quality assurance framework that complements the liability regime. At the international level, states should engage multilaterally in the development of AI-specific cyber operation norms through the existing UN cybersecurity governance framework, and should incorporate legal review requirements for AI cyber capabilities into their national military and intelligence legal review processes.

### *9.3 Organisational Governance*

Organisations deploying AI in cybersecurity operations should implement internal governance frameworks that operationalise the principles articulated above. This includes: the conduct of AI impact assessments before deployment of AI security systems with significant implications for individuals, documenting the identified risks and the measures implemented to mitigate them; the establishment of AI ethics review processes for AI security tools that consider fairness, privacy, and accountability alongside the technical security requirements they serve; the maintenance of human oversight mechanisms that are genuinely effective rather than nominally compliant; the development of clear internal accountability assignments for AI security system decisions and their consequences; and the implementation of continuous monitoring programmes that assess AI security system performance for accuracy, bias, and fairness on an ongoing basis.

## 10. CONCLUSION

The deployment of AI in cybersecurity represents one of the most consequential intersections of technological capability and legal and ethical obligation in contemporary governance. The defensive benefits of AI in cybersecurity are real and growing — enabling detection, response, and protection at speeds and scales that human-only security operations cannot match, and providing the capacity to defend against an adversarial AI-enabled threat landscape that makes AI adoption for defenders increasingly necessary rather than merely advantageous. But the ethical and legal challenges that AI cybersecurity solutions introduce are equally real and equally growing — in their scope, their severity, and the inadequacy of the current legal frameworks to address them.

The opacity of AI decision-making, the bias embedded in AI security systems trained on historically skewed data, the accountability gaps created by increasing operational autonomy, the surveillance implications of AI-powered security monitoring, the unresolved liability framework for AI security failures, and the international legal questions raised by AI-enabled cyber operations together constitute a governance challenge whose complexity and urgency match the defensive challenge that drives AI adoption. The normative framework advanced in this article — grounded in principles of proportionate



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

transparency, fairness by design, meaningful human control, data minimisation, accountability, human dignity, and international responsibility — provides a foundation for regulatory, liability, and organisational governance frameworks that take both dimensions of the challenge seriously.

## REFERENCES

Antitrust Division, US Department of Justice, 'Antitrust Guidance for Human Resource Professionals' (DOJ, 2016).

Barocas, S. and Hardt, M., 'Fairness in Machine Learning' (NIPS 2017 Tutorial, 2017).

Big Brother Watch and Others v. United Kingdom, Application Nos. 58170/13, 62322/14 and 24960/15, Grand Chamber Judgment of 25 May 2021.

Brundage, M. et al., 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation' (Future of Humanity Institute, 2018).

Chesney, R. and Citron, D., 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107(6) California Law Review 1753.

Crawford, K. and Calo, R., 'There is a Blind Spot in AI Research' (2016) 538 Nature 311.

Doshi-Velez, F. and Kim, B., 'Towards a Rigorous Science of Interpretable Machine Learning' (2017) arXiv:1702.08608.

European Commission, 'Proposal for a Regulation on Harmonised Rules on Artificial Intelligence (AI Act)', COM(2021) 206 final.

European Commission, 'Proposal for a Directive on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive)', COM(2022) 496 final.

European Data Protection Board (EDPB), 'Guidelines 05/2020 on Consent under Regulation 2016/679' (EDPB, 2020).

European Data Protection Board (EDPB), 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679' (EDPB, 2018).

European Union AI Act, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024.

European Union Revised Product Liability Directive, Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024.

Floridi, L. et al., 'An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations' (2018) 28 Minds and Machines 689.

Geiger, R. and Lehr, D., 'Pseudonymization as a Privacy-Enhancing Technology' (2021) 12 European Journal of Law and Technology 1.

General Data Protection Regulation, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016.

Goodfellow, I.J. et al., 'Explaining and Harnessing Adversarial Examples' (2015) Proceedings of the 3rd International Conference on Learning Representations.

Hadlington, L., 'Human Factors in Cybersecurity: Examining the Link between Internet Addiction, Impulsivity, Attitudes Towards Cybersecurity, and Risky Cybersecurity Behaviours' (2017) 3(3) Heliyon e00346.

Huang, L. et al., 'Adversarial Machine Learning' (2011) Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, 43.



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index

International Committee of the Red Cross (ICRC), 'Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons' (ICRC, 2016).

Schmitt, M.N. (ed.), 'Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations' (2nd ed., Cambridge University Press, 2017).

Taddeo, M., 'The Debate on the Moral Responsibilities of Online Service Providers' (2016) 22(3) Science and Engineering Ethics 1575.

Taddeo, M. and Floridi, L., 'The Debate on the Moral Responsibilities of Online Service Providers' (2016) 22 Science and Engineering Ethics 1575.

United Nations Group of Governmental Experts, 'Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security', A/76/135 (2021).

United States, 'Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence' (30 October 2023), 88 Fed. Reg. 75191.

Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies (1996, as amended).

Whittaker, M. et al., 'AI Now Report 2018' (AI Now Institute, NYU, 2018).

Wieringa, M., 'What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability' (2020) Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 1.

Zeiler, M.D. and Fergus, R., 'Visualizing and Understanding Convolutional Networks' (2014) Proceedings of the 13th European Conference on Computer Vision, 818.



© 2026 Cybersecurity Law and Policy Review | https://clprj.com/journal/index