International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Ethical considerations and accountability frameworks in deploying fully autonomous AI for real-time cybersecurity response in the USA

Oluwabiyi Oluwawapelumi Ajakaye [1, *], Ifeoma Eleweke [2], Ikenna Patrick Nwobu [3], Isaac Yusuf, [4], Abdul-Waliyyu Bello [5] and Idris Wonuola [5]

[1] *Department of Telecommunications Engineering, University of Sunderland, Tyne and Wear, United Kingdom.*
[2] *Department of Technology and Engineering,Westcliff University, Florida, USA.*
[3] *Department of Business Administration, Liberty University, Virginia, USA.*
[4] *Department of Mathematics. University of Ibadan, Ibadan, Nigeria.*
[5] *Department of Computer Science. Austin Peay State University, Tennessee, USA.*

International Journal of Science and Research Archive, 2024, 13(02), 1513-1540

Publication history: Received on 13 November 2024; revised on 24 December 2024; accepted on 29 December 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.2.2646

Abstract

Researchers are looking into fully autonomous artificial intelligence (AI) systems more and more for real-time responses to cybersecurity threats, especially in high-stakes defense settings. These systems promise to be faster and bigger than anything else when it comes to stopping cyber-attacks, but they also bring up important moral and legal issues about who is responsible, how open they are, and how they are run. This paper goes into great detail about the moral issues and accountability systems that are important for using fully autonomous AI in U.S. defense cybersecurity. We look at the most up-to-date theoretical frameworks, such as the U.S. Department of Defense (DoD) ethical AI principles and federal guidelines. Then, we do a legal-ethical analysis to find holes in the laws and policies that are already in place. We look at real-world problems by looking at case studies of autonomous cyber defense systems, such as DARPA's "Mayhem" system in the Department of Defense and industry solutions like Darktrace and CrowdStrike in defense settings. Some of the main ethical problems that have been found are the possibility of AI making decisions that are biased or unclear, the risk of invading people's privacy, and the lack of clarity about who is responsible for actions taken by AI. In response, we suggest a structured accountability framework that makes it clear what AI developers, integrators, and deployers are responsible for. This framework would be supported by oversight tools like human-in-the-loop controls, audit trails, and following new AI risk management standards. The results show that strong accountability frameworks and proactive ethical guidelines are necessary to safely take advantage of the benefits of autonomous AI cyber defense in the U.S. while still following the law and keeping the public's trust.

Keywords: Autonomous AI; Cybersecurity; Ethical AI; Accountability; Defense; Legal Frameworks; Real-Time Response; AI Governance

1. Introduction

Cybersecurity threats are happening more often and are getting more complex, which is a big risk to national security and important infrastructure (GAO, 2024). In fiscal year 2023, US federal agencies reported more than 32,000 cybersecurity incidents. The global cost of cybercrime is expected to rise from about $9 trillion in 2024 to nearly $14 trillion by 2028. Because the threat landscape is getting worse, people are more interested in AI solutions that can find and stop attacks faster than people can (Batarseh et al., 2021). Cyber defense could be changed forever by fully autonomous AI systems that can analyze threats and carry out responses in real time without any help from people (Chou et al., 2023). These systems use machine learning to find problems and take action in less than a second, which is

* Corresponding author: Oluwabiyi Oluwawapelumi Ajakaye ORCID: 0009-0000-4014-4285

much faster than people can react (Shehzadi, 2022). The U.S. defense and security community has started to test these kinds of self-driving cyber defenses. For example, DARPA's 2016 Cyber Grand Challenge showed that self-driving "cyber reasoning" bots could find and fix software bugs without any help from people (Pellerin, 2016). The Department of Defense (DoD) has recently used AI-powered cyber tools like "Mayhem" to automate some parts of software security testing and vulnerability management on military networks (Brumley, 2020). These changes show how promising autonomous AI is for making cyber defenses stronger against constant attacks.

But fully autonomous cybersecurity AIs also raise serious ethical and governance issues, despite their promise. An autonomous AI agent can decide on its own to block network traffic, isolate systems, or even take action against an intruder, unlike traditional security tools that are controlled directly by people. It is not clear who or what should be held responsible if these kinds of decisions are wrong or cause damage that wasn't meant to happen (Rao et al., 2022). There are also problems with transparency and explainability: AI's "black box" decision-making can make it hard for operators to know why a certain action was taken (Fox et al., 2024). Another worry is bias and fairness, since machine learning models that are trained on limited or biased data might be able to tell the difference between threats and make unfair decisions (Shehzadi, 2022). Additionally, using autonomous surveillance and response tools could violate privacy and civil rights if these systems keep a close eye on network activity or take intrusive actions without supervision (Redress Compliance, 2024). In the sensitive area of defense, it is very important that AI-driven cyber responses follow both legal rules of engagement and moral standards. The Department of Defense (DoD) has acknowledged the need for ethical AI principles to guide military AI use. In 2020, they adopted five key principles Responsible, Equitable, Traceable, Reliable, and Governable to shape AI development and deployment (Department of Defense, 2020).

This paper looks at the ethical issues and accountability frameworks that are important for using fully autonomous AI in real-time cybersecurity response in the US, with a focus on federal defense applications. First, we'll look at some theoretical frameworks and rules for ethical AI and algorithmic accountability that apply to autonomous cyber systems (Section 2). Next, we describe our method, which uses both legal-ethical analysis and case study research (Section 3). In Section 4, we talk about some of the biggest ethical problems that come with using autonomous AI for cyber defense. These include gaps in accountability, issues with transparency and bias, and operational risks. A legal gap analysis is presented in Section 5. It looks at how well (or poorly) the U.S. laws, rules, and policies deal with these problems. In order to give the discussion some context, Section 6 looks at case studies of autonomous AI being used in cybersecurity, including examples from the U.S. defense sector and major industry platforms. These examples show what lessons were learned about oversight and responsibility. Section 7 builds on these ideas by suggesting an accountability framework specifically for autonomous cyber defense AI. This framework outlines the roles and practices that are necessary to ensure ethical compliance and governance. Lastly, Section 8 ends with suggestions for policy and practice. It stresses that without strong ethical guidelines and accountability measures, using fully autonomous AI in cybersecurity poses serious risks to security, legality, and public trust.

2. Theoretical Frameworks for Ethical AI in Cybersecurity

To use AI for cybersecurity in a responsible way, you need to know about both ethical theory and established ways of governing. In the last few years, governments and businesses have come up with rules and principles to make sure that AI technologies are in line with social norms and the law. This part looks at some of the most important ethical frameworks and rules that guide the use of autonomous AI in U.S. defense and cybersecurity.

2.1. Ethical Principles and AI Governance in Defense

The U.S. Department of Defense made a list of AI Ethical Principles in February 2020 to help with the development and use of AI in military settings (Department of Defense, 2020). The Department of Defense (DoD) adopted these five principles Responsible, Equitable, Traceable, Reliable, and Governable to make sure that AI systems are used in a way that is legal and ethical. Responsible AI means that people are responsible for what happens with AI systems; Equitable AI tries to reduce unintended bias that could lead to unfair outcomes; Traceable AI stresses openness in how AI works and what it does; Reliable AI requires that systems have clear, well-defined uses and work as planned; and Governable AI requires that systems be able to find and avoid unintended consequences, including being able to be turned off if they act strangely (DoD, 2020). Even though these principles are broad, they make it clear that people must still be able to make decisions and be held accountable as autonomy grows (Defense Innovation Board, 2019). In practice, this could mean making sure that a human operator is in charge of an autonomous cyber defense agent's decisions, especially on missions that could have big effects, and making sure that there are ways to stop or turn off the AI if necessary (Cohn et al., 2020).

The White House Office of Science and Technology Policy released the Blueprint for an AI Bill of Rights in 2022 to encourage responsible AI use in all areas of life. This was in addition to the DoD principles. This framework, which is not legally binding, lists five basic rights: (1) Safe and Effective Systems, (2) Algorithmic Discrimination Protections, (3) Data Privacy, (4) Notice and Explanation, and (5) Human Alternatives, Consideration, and Fallback (OSTP, 2022). Several of these principles can be used directly with autonomous cybersecurity systems. For example, the need to thoroughly test autonomous cyber defenses so they don't accidentally damage system integrity or safety (for example, by shutting down critical infrastructure by mistake) is in line with the need for safe and effective AI. The principle of algorithmic discrimination protections makes it clear that cybersecurity AI should not unfairly target or ignore certain people or assets because of biased training data. This could happen if, for example, an AI system learns to treat activity from a certain network segment or user demographic as more suspicious without any reason (Williams and Burnap, 2021). The notice and explanation principal stresses how important it is to be open and honest: people who have a stake in an AI system should know when it is making security decisions and be able to understand why it is doing important things. Lastly, human alternatives and fallback are in line with the DoD's Governable AI principle because they say people should be able to choose not to have automated decisions made for them and ask for a human review (OSTP, 2022). In the context of defense cybersecurity, this could mean that a human commander or security analyst can override an AI's decision to isolate a server or start a countermeasure, keeping human control in important situations.

2.2. Accountability and Risk Management Frameworks

Accountability in AI deployment means that there are clear ways to make sure that AI systems stay within the law and moral boundaries, and that someone can be held responsible for what the AI does or doesn't do (Wirtz et al., 2022). There are now a number of frameworks that can help make AI accountability a reality. In 2021, the U.S. Government Accountability Office (GAO) made an AI Accountability Framework for federal agencies. The four pillars that make up this framework are Governance, Data, Performance, and Monitoring (GAO, 2021). Governance means setting rules for AI projects and having leaders keep an eye on them. Data means making sure that data is accurate, complete, and private. Performance means making sure that the AI system works well and fairly by testing and validating it. Monitoring means regularly checking the results of AI and being able to audit decisions. There are practices and assessment questions for each pillar that agencies can use to make sure they use AI systems responsibly. For instance, under Governance, agencies are told to clearly define who is in charge of overseeing AI, and under Monitoring, they are told to set up ways for people to review AI decisions and respond to problems if AI doesn't work (GAO, 2021). The GAO framework makes it clear that accountability is a concern throughout the life cycle of an AI system. It needs to be built into the design phase (through risk assessments and bias mitigation in the Data phase), the deployment phase (with performance metrics and validation), and the post-deployment phase (with ongoing monitoring and auditing).

The National Institute of Standards and Technology (NIST) published the NIST AI Risk Management Framework (RMF) in 2023. This is another model that is relevant. The NIST AI RMF is a complete but optional way for businesses to find, evaluate, and deal with the risks of AI systems (NIST, 2023). It adds basic functions: Map (put AI risk in context, like figuring out what it's meant to do and what effects it might have), Measure (look at and rate risks like accuracy errors, bias, and security holes), Manage (put controls and safeguards in place to lower those risks), and Govern (make sure there is a culture of risk management and accountability) (NIST, 2023). Using the NIST RMF on an autonomous cyber defense AI might mean figuring out how an AI response tool could fail or be attacked (for example, thinking about how an enemy might try to trick the AI by giving it bad inputs), testing those risks (for example, by doing adversarial penetration tests of the AI), and managing them through design choices (for example, by adding rule-based limits to stop the AI from being too aggressive). In this case, governance would mean having written rules about who can use AI and how (who can deploy it and with what approvals) and checking the AI's decisions every so often to make sure they follow the law and ethical standards. The NIST framework's focus on reliable AI features like explainability, reliability, privacy, and robustness are very similar to what is needed in defense situations, where AI mistakes can have serious consequences and must be kept to a minimum and expected (Schwartz et al., 2022).

Table 1 summarizes several foundational frameworks and principles that guide ethical and accountable AI deployment in U.S. cybersecurity and defense contexts, highlighting their key components and focus areas.

Table 1 Selected Ethical AI Frameworks and Guidelines in the U.S. Context

Framework / Guideline	Source (Year)	Key Principles / Focus Areas
DoD Ethical AI Principles	DoD (2020)	Responsible (human accountability), Equitable (avoid bias), Traceable (transparent and auditable), Reliable (consistent performance), Governable (control mechanisms). Guides military AI use to ensure compliance with U.S. values and law.
Blueprint for an AI Bill of Rights	OSTP/White House (2022)	Safe and Effective Systems, Algorithmic Discrimination Protections, Data Privacy, Notice and Explanation, Human Alternatives and Fallback. Emphasizes user rights and human oversight in automated systems (OSTP, 2022).
GAO AI Accountability Framework	GAO (2021)	Governance (policies, roles, oversight), Data (quality, privacy), Performance (validation, fairness), Monitoring (continuous audit and incident response). Provides audit practices for federal AI accountability (GAO, 2021).
NIST AI Risk Management Framework (RMF)	NIST (2023)	Core functions: Map, Measure, Manage, Govern AI risks. Addresses trustworthiness criteria like transparency, safety, bias mitigation, robustness. Encourages iterative risk assessment and stakeholder engagement (NIST, 2023).
Industry AI Ethics Guidelines (e.g., Microsoft, Google)	Various (2018–2023)	Many tech companies (e.g., Microsoft's AI Principles, Google's AI Principles) share common themes: fairness, transparency, reliability/safety, privacy, and accountability (Redress Compliance, 2024). These influence AI tools used in cybersecurity (e.g., IBM's Watson Cyber Security emphasizes explainability and trust (Redress Compliance, 2024)).

These frameworks all say that AI systems must be designed and built with human accountability, openness, fairness, and safety in mind. In the context of autonomous cybersecurity, this means that businesses should have clear governance structures (like an AI oversight board or designated responsible officers), do thorough testing for biases and mistakes, keep audit logs of AI-driven actions, and make sure that people can step in when necessary. Next, we'll talk about how we use these principles to look at both the ethical problems and the case studies of autonomous AI in cyber defense.

3. Methodology

This study looks into ethical and legal accountability issues in autonomous cybersecurity AI deployments using a qualitative, multi-method approach. The method has two main parts: (a) a review of the literature and policies and (b) an analysis of case studies, all done within a legal-ethical framework.

3.1. Review of Literature and Policy

We carefully looked over a lot of academic papers, industry reports, and U.S. government policy papers that had to do with AI ethics, cybersecurity, and the law. We got most of our information from Scopus-indexed journals (like IEEE Security and Privacy, ACM Transactions on Cybersecurity, and Journal of Cyber Policy) and databases like IEEE Xplore, ScienceDirect, and JSTOR to make sure we had the most up-to-date (2018–2024) scholarly ideas. We also looked at official publications and guidelines from U.S. government agencies like the DoD, NIST, DHS, and GAO to get a sense of the current policy landscape. We looked closely at documents such as the DoD's AI Ethical Principles (2020), NIST's AI Risk Management Framework (2023), the White House OSTP's AI Bill of Rights (2022), and GAO reports on AI and cybersecurity. This review set up the theoretical frameworks and pointed out any legal gaps or provisions, as well as known ethical problems (like those listed in Table 1). Finding legal gaps was a top priority: places where current cybersecurity law (like the Computer Fraud and Abuse Act and federal cyber incident reporting requirements) and AI-related law (if there is any) don't clearly cover autonomous AI operations and accountability.

3.2. Looking at a case study

Based on the literature review, we chose case studies that show how autonomous AI is used in cybersecurity, with a focus on the U.S. defense and security sector. The cases were chosen to include both government-led projects and private-sector platforms that are useful for the federal government. Some examples are: (1) DARPA/DoD's "Mayhem," an autonomous cyber defense prototype that made history by finding and fixing software vulnerabilities on its own (Brumley, 2020); (2) Darktrace's "Antigena," an autonomous response system used in defense and government settings that uses machine learning to neutralize threats in real time (Jamil, 2024); and (3) CrowdStrike's AI-driven threat response platform, which shows that some U.S. government agencies are using AI in endpoint defense (Granville, 2023). There are also other examples, like IBM's Watson for Cyber Security and DHS's pilot uses of AI, that are included as they are relevant. We got information for each case from technical white papers, press releases, case reports, and interviews when we could. We looked at how these systems work, how much autonomy they use (for example, fully automated response vs. human-in-the-loop), and any ethical or operational problems that have been reported. We then compared each case to the moral frameworks and rules we had already talked about. For example, this meant asking: Was there a person in charge of the AI's actions or a chain of responsibility? How did the system deal with openness? Did it give operators reasons for its actions? What protections were in place to stop or lessen mistakes? Before deployment, were there any reviews of the law or policy, like a legal review of the rules of engagement in a DoD context?

3.3. Legal-Ethical Analysis

We used the results of the literature/policy review and case studies to do a gap analysis. We looked at how current U.S. laws and regulations, such as federal cybercrime laws, defense procurement regulations, and administrative law limits on the federal government's use of AI, apply to the situations described in the case studies. We found the gaps where there aren't any clear laws, like when it's not clear who is responsible if an AI causes an accident. We also looked at international law and norms, such as the Tallinn Manual on the law of cyber warfare, when looking at cases that might involve cyber actions across borders. We looked at whether the current state of deployments meets accepted ethical standards using both legal analysis and ethical theory (deontological vs. utilitarian perspectives, ideas like the "responsibility gap" (Matthias, 2004), and principles from the frameworks in Table 1).

3.4. This combined method gives a full picture

the literature and policy review give the research a foundation in established principles and known problems, the case studies give real-world examples and point out gaps, and the legal-ethical analysis connects the two by looking at what accountability measures are in place and what might be needed. The outcome of this method is a list of ethical issues and legal gaps (see Sections 4 and 5), as well as a suggested accountability framework (see Section 7) that is based on both theory and practice. To make sure the results were correct, we checked them against several sources. For example, if a case study claims about AI performance came from a vendor, we looked for independent or third-party comments on that case. The research is mostly qualitative, but it uses current data like incident statistics and survey results to show trends (as shown in the figures and tables throughout). In the end, this method hopes to give policymakers, military leaders, and cybersecurity experts useful advice on how to run autonomous AI systems in a responsible way.

4. Ethical Challenges in Fully Autonomous Cybersecurity AI

Deploying fully autonomous AI for cyber defense confronts organizations with a range of ethical challenges. Unlike conventional cybersecurity tools, these AI systems make independent decisions that can significantly impact information systems, data, and even human operations. The following are key ethical issues that arise, many of which intersect and compound one another. Table 2 (below) summarizes these challenges and possible mitigation approaches.

4.1. Accountability and the “Responsibility Gap”

Perhaps the most prominent ethical concern is the question of accountability: Who is answerable when an AI-driven defense action causes harm or fails to prevent it? Autonomous cyber systems blur traditional lines of responsibility, creating what some scholars call a “responsibility gap” (Matthias, 2004). In a manual operation, a human operator (e.g., a security analyst or military officer) would be the decision-maker and thus accountable for the outcomes (good or bad) of a cybersecurity action like shutting down a server or blocking an IP address. With an AI agent acting independently, that direct attribution is lost. An incident could occur where the AI erroneously classifies legitimate network traffic as malicious and automatically cuts off a critical service, leading to downtime or safety risks. Is the AI itself “to blame”? Should the creators of the AI be held liable for not foreseeing the error, or the commanders who deployed it, or the end-user who failed to supervise it? This challenge is not merely theoretical: in corporate settings, questions have already arisen about liability when AI-based security tools malfunction, and in defense scenarios the stakes are even higher (Bryson et al., 2017). Rao et al. (2021) illustrate this dilemma by asking whether liability should rest with the AI’s

developer, the deploying organization, or neither when an autonomous cyber tool executes an unlawful or damaging act. Ethically, the lack of a clear accountability chain can undermine trust in the AI system and in the institution using it. If no one can be held responsible for an AI's actions, there is a risk of a moral vacuum where harmful consequences fall through the cracks of accountability (Matthias, 2004). Therefore, addressing this gap is paramount: organizations must establish ex ante who will be accountable for the AI's decisions – typically this will be the commanders or operators overseeing the AI, as per the DoD's Responsible AI principle (humans "remain responsible" for AI outcomes (DoD, 2020)). They must also implement oversight structures (such as review boards or audit trails) to enforce that accountability in practice.

4.2. Transparency and Explainability

Closely related to accountability is the issue of transparency – the degree to which the AI's decision-making process can be understood by humans. Many advanced AI systems, particularly those using deep learning, are inherently complex and can act as "black boxes" where even developers struggle to interpret how inputs are being mapped to outputs (Samek et al., 2021). In cybersecurity, this opacity is problematic. Security professionals need to know why an AI flagged a benign software update as malware or why it decided to isolate an internal host as a suspected insider threat. Without explainability, users may have blind faith in the AI or, conversely, mistrust and frequently override it, undermining its utility (Shin, 2020). From an ethical standpoint, explainability is tied to the concept of respect for persons – affected parties deserve an explanation for decisions that impact them. Lack of transparency also hampers accountability, as one cannot easily investigate whether an AI made a reasonable decision if its reasoning cannot be reconstructed. Consider a scenario where an autonomous AI's action leads to a diplomatic incident (e.g., it mistakenly identifies traffic from an allied nation as an attack and blocks it, causing offense). Being able to explain that the AI acted on certain data patterns is crucial for accountability and remedy. To address this, the field of explainable AI (XAI) offers methods to provide human-interpretable insights into AI decisions. For example, an AI could highlight the network traffic features that led it to conclude an attack was underway. Incorporating XAI is increasingly seen as an ethical imperative in high-stakes AI deployments. Shehzadi (2021) argues that explainable and transparent decision processes are necessary so stakeholders can monitor and trust autonomous cyber defenses. In practice, this might mean using simpler models where feasible, attaching explanation modules to complex models, or at minimum logging the decision rationale (e.g., "blocked host X due to anomaly score above threshold, unusual data exfiltration at 3AM") for later analysis.

4.3. Bias and Fairness

Algorithmic bias is an ethical challenge where the AI system may perform inequitably for or against certain groups, activities, or entities. In cybersecurity AI, bias can manifest in various ways. One example is a bias in training data: if the AI was trained mostly on attack data that came from certain network environments, it might be less effective at detecting threats in underrepresented environments, thereby unfairly leaving some systems more vulnerable. Conversely, it might also produce more false positives on certain user behaviors if those behaviors were rare or absent in training data. For instance, an AI might learn a proxy rule that "traffic spike late at night = threat" based on corporate network data; if deployed in a military context where late-night operations are routine, it could unfairly flag benign activity by night-shift personnel as malicious. Another angle is automation bias, where humans might defer to the AI's judgment even when it is wrong, under the assumption that the AI is objective – which ironically can amplify the impact of any bias the AI has (Skitka et al., 1999). The ethical principle of justice demands that security controls (AI-driven or not) be applied fairly and without unjust discrimination. Bias in an autonomous cyber defense could, for example, result in disproportionate blocking of certain users or devices, effectively "profiling" them as risky without due cause. Such outcomes can erode morale and trust, and potentially violate regulations or policies (for example, if an AI unintentionally targets traffic from a particular geographic region, it might conflict with organizational nondiscrimination policies or international agreements). To mitigate bias, organizations should implement rigorous bias testing and validation as part of AI development (Mehrabi et al., 2021). This includes using diverse and representative training data, and auditing the AI's alerts and actions across different subgroups to check for skewed patterns. Ethically, if bias is detected, steps must be taken to correct it – whether through retraining the model, adding rule-based correctives, or in some cases opting not to automate a decision that is too sensitive to context. The DoD's Equitable AI principle directly speaks to this need, insisting AI outcomes be free from unintended bias to the extent possible. In sum, ensuring fairness in autonomous cybersecurity AI is critical to prevent the technology from causing undue harm or neglect to any particular group of users or assets (Fox et al., 2020).

4.4. Security and Reliability of the AI System

It may sound paradoxical, but the AI system itself can become a target or point of failure – leading to ethical issues concerning security of the AI. If adversaries compromise or manipulate an autonomous defense AI, they could potentially turn our own defense into a tool of attack. For example, through techniques of adversarial machine learning,

an attacker might subtly alter their malware or network behavior to fool the AI's detection algorithms, causing the AI to either ignore a real threat or to overreact to a benign event (Kurakin et al., 2018). Such adversarial attacks exploit the pattern recognition nature of AI and can induce errors that a human analyst might not make. An ethical AI deployment must account for these possibilities; otherwise, it could provide a false sense of security or, worse, be weaponized by adversaries. Another reliability concern is the AI's robustness in novel situations. Cyber threats evolve quickly, and an AI that performs well on known types of attacks might fail to recognize a completely new exploit or may behave unpredictably outside its training distribution. The ethical duty of beneficence (doing good / preventing harm) implies that deploying an autonomous system requires high confidence in its reliability – otherwise, erratic behavior could cause more harm than it prevents. A case in point: an autonomous AI in a network defense role could conceivably misconstrue a rapid surge in network traffic (maybe caused by a system update) as a DDoS attack and start shutting down network nodes, causing a self-inflicted outage. Ethically, such outcomes violate the expectation that technology should not introduce new unacceptable risks. To address this, developers and operators need to implement strong security measures for the AI (Redress Compliance, 2024). This includes hardening the AI model and its surrounding infrastructure against tampering (e.g., using encryption and access controls for model files and decision logs). It also means performing adversarial testing – essentially, red-teaming the AI – to identify how it might be deceived or induced to err. From those tests, improvements such as anomaly checks (meta-monitoring of the AI's own outputs for signs of manipulation) can be put in place. Ensuring fail-safes is another critical practice: if the AI starts behaving erratically (due to an attack or a software fault), there should be an automatic fallback to a safe mode (possibly defaulting to requiring human confirmation for actions). Maintaining the integrity and reliability of the AI is not just a technical necessity but an ethical one – users and stakeholders trust that the defense measures will not themselves become vectors of harm. As one example of good practice, Darktrace's autonomous response product includes modes where it takes only limited, proportionate actions (like slowing down a connection rather than severing it) and signals for human review when uncertain, thereby balancing automated defense with cautious restraint (Darktrace, 2022). This kind of design addresses the reliability challenge by preventing extreme actions unless confidence is high.

4.5. Data Privacy and Civil Liberties

Autonomous cybersecurity systems often require ingesting and analyzing vast amounts of network and user data to function effectively. This raises ethical issues of data privacy and surveillance. An AI that monitors real-time network traffic could capture emails, messages, and personal information of users as it scans for threats. In a defense context, it might process employees' or civilians' data streams. If not carefully governed, such surveillance can infringe on privacy rights and civil liberties (Taddeo and Floridi, 2018). The ethical principle of respect for privacy requires that data usage be proportionate and authorized. One risk is mission creep: a tool introduced for cybersecurity might accumulate data that gets repurposed for other monitoring (e.g., workplace misconduct, intelligence gathering) without proper legal process or consent. This is especially sensitive in the military and government domains, where operations are secretive – but employees and citizens still have privacy expectations under laws like the Privacy Act. The AI Bill of Rights' principles of data privacy and notice are relevant here: individuals should not be subjected to unchecked data collection by AI, and they should know if an AI defense system is examining their communications (OSTP, 2022). Another dimension is the privacy of external entities: for instance, if an autonomous defense AI traces an attack and initiates counter-actions, it might incidentally access systems beyond the agency's own network, raising legal issues under the Fourth Amendment or Computer Fraud and Abuse Act if done domestically without warrant or under international law if foreign systems are touched. Ethically and legally, active defense measures by an AI must be constrained to avoid violating rights or sovereignty (Ohm, 2020). Mitigating these concerns involves instituting strict data handling and minimization practices in AI systems. Techniques like privacy-preserving machine learning (e.g., federated learning or on-device analysis) can be used to limit data exposure – for example, an AI could analyze certain patterns locally on a device and share only security alerts to a central system, rather than all raw data. Additionally, policies should require that autonomous systems operate under the same access controls and oversight as human analysts. Just as an analyst has rules about what data they can view and must undergo auditing, an AI's data accesses and actions should be logged and subject to audit. In scenarios where the AI might take counter-strike actions (sometimes termed "hack back"), there must be explicit legal authorization and human sign-off, because autonomous retaliation blurs into use-of-force decisions that carry significant ethical weight (Sharples and Evans, 2020). Currently, U.S. policy generally prohibits private entities from hack-back activities due to these concerns (American University, 2020), and any government use would need careful interagency and legal review. Privacy and civil liberties can be protected by ensuring autonomous cyber defenses are bounded in their operation – they only operate on approved data sources, for approved purposes, and with oversight from privacy and legal officers. An illustrative example: the DHS's deployment of an AI-based network monitoring tool could be accompanied by a privacy impact assessment and continuous monitoring by a Privacy Officer to ensure the tool doesn't over-collect or retain personal data beyond what's necessary for security (DHS, 2023).

4.6. Human Factors: Trust, Skills, and Job Impacts

The introduction of fully autonomous AI in cybersecurity teams also raises ethical and practical questions around human factors. One is the appropriate level of trust that human operators place in the AI. Over-trust can lead to complacency – analysts might ignore warning signs of the AI's mistakes (Moschovaki et al., 2023). Under-trust, on the other hand, might cause constant second-guessing or disabling of the AI, squandering its benefits (Lee and See, 2004). Designing the human-AI interaction is thus an ethical issue: the AI should be presented in a way that calibrates trust (for example, by showing confidence scores or explanations, as discussed in section 4.2). There is also an obligation to train and prepare personnel for working with AI. Ethically, organizations owe it to their employees to reskill or upskill them so that they can effectively supervise and collaborate with autonomous systems, rather than being displaced or left in the dark (Redress Compliance, 2024). The deployment of autonomous AI will transform the role of cybersecurity professionals – ideally shifting them to more high-level strategic tasks while the AI handles routine incidents (Raj and Gupta, 2022). However, if not managed properly, it could lead to job redundancy or a devaluation of human expertise. Ensuring an ethical transition means involving employees in the deployment process, being transparent about how their roles will change, and providing opportunities for them to focus on areas where human judgment is irreplaceable (such as complex incident investigation, creative problem solving, or governance decisions). Moreover, having humans in the loop can provide a moral buffer – it aligns with the principle of meaningful human control advocated in AI ethics (EU High-Level Expert Group on AI, 2019), where humans maintain situational understanding and can prevent unethical outcomes that a machine, lacking context or moral reasoning, might not recognize. For example, a human analyst might know that shutting off a particular system, while stopping an attack, would also disable a hospital's life support network – a nuance an AI might not grasp if it only sees network metrics. Keeping skilled humans involved ensures such contextual moral judgments are applied.

In summary, the ethical challenges of autonomous cybersecurity AI span technical, human, and organizational domains. They include ensuring someone is accountable for AI actions, making AI decisions transparent, preventing bias and unfair impacts, securing the AI against misuse, guarding privacy, and integrating AI with human teams in a way that respects human dignity and employment. These challenges must be proactively addressed for any deployment to be considered ethically responsible. Table 2 below encapsulates these key challenges along with indicative mitigation strategies drawn from best practices and guidelines.

Table 2 Major Ethical Challenges in Autonomous Cybersecurity AI and Mitigation Approaches

Ethical Challenge	Description of Issue	Possible Mitigation Strategies
Accountability and "Responsibility Gap"	Uncertainty about who is responsible for AI-driven actions or mistakes (developer, operator, organization?). Can lead to moral and legal accountability void (Rao et al., 2021).	• Define accountability structure before deployment (e.g., assign a human supervisor for the AI's decisions). • Implement audit logs and incident review procedures to trace AI actions to responsible personnel. • Follow DoD's principle that humans remain responsible for AI outcomes, codified in ROE or policy.
Transparency and Explainability	AI decisions are often opaque, hindering understanding and oversight. Operators may not know why the AI acted, impacting trust and ability to contest decisions.	• Use eXplainable AI (XAI) techniques to provide reasons for AI decisions (e.g., feature importance, rules extracted). • Require AI systems to provide confidence levels and alerts when uncertain. • Train operators to interpret AI outputs; develop interfaces that visualize AI reasoning (Fox et al., 2024). • Conduct regular "AI decision audits" where random AI actions are explained and verified by humans.
Bias and Unfair Outcomes	AI may exhibit bias, flagging benign behavior as malicious more often for certain groups/activities (false positives) or missing attacks in underrepresented scenarios (false negatives). Raises fairness concerns (Shehzadi, 2022).	• Ensure diverse, representative training data covering various user groups and network conditions. • Perform bias testing: evaluate false positive/negative rates across different subpopulations and contexts, and adjust models if disparities are found (Mehrabi et al., 2021). • Incorporate fairness constraints into the AI model or use post-processing to correct biased outputs. • Maintain a human review process for AI decisions that have significant impact on individuals (aligns with OSTP Bill of Rights' human fallback principle).
Security and Robustness of AI	The AI system itself can be attacked or can fail unpredictably (e.g., adversarial inputs cause misclassification). If compromised, it could do harm (e.g., attacker manipulates AI to disable defenses).	• Secure the AI model and infrastructure: restrict access, encrypt model parameters and communications, and monitor for anomalies in the AI's behavior. • Conduct adversarial testing and red-teaming on the AI (simulate sophisticated attackers to find exploits). • Implement fail-safes: if AI outputs deviate from expected patterns or confidence is low, require human confirmation or switch to a read-only alert mode.

		• Update and patch the AI regularly like any critical software, to fix vulnerabilities or logic errors discovered.
Data Privacy and Civil Liberties	Autonomous monitoring can intrude on privacy by analyzing personal or sensitive data. AI might collect more data than necessary, or undertake actions (like "active defense" hacks) that raise legal/privacy issues (Taddeo and Floridi, 2018).	• Apply data minimization: feed the AI only the data features needed for threat detection, anonymize where possible. • Conduct Privacy Impact Assessments (PIA) for AI deployments in government per legal requirements, and address identified risks. • Implement policy constraints: e.g., geo-fencing the AI's actions so it cannot extend beyond authorized networks, and requiring human sign-off for any countermeasures that might affect external systems (to comply with law and avoid unauthorized "hack-back"). • Provide notice to users (where appropriate) that an AI is monitoring network activity, and maintain transparency about data use (OSTP, 2022). In sensitive environments, involve legal counsel and privacy officers in oversight of the AI's operation.
Human Factors and Trust	Risk of over-reliance on AI (operators may become disengaged or deskilled) or under-utilization (disabling AI due to lack of trust). Also, potential job displacement or role changes raising ethical workforce concerns (Redress Compliance, 2024).	• Establish clear guidelines for human-AI interaction: define when humans must intervene vs. when AI can act autonomously. • Train cybersecurity staff in how the AI works, its limitations, and how to interpret its outputs, to calibrate appropriate trust. • Use a gradual deployment approach (e.g., AI provides recommendations first, humans decide, then move to semi-autonomous, etc.) to build trust and understanding. • Emphasize AI as *augmenting* human analysts, not replacing: redistribute routine tasks to AI and shift humans to supervision and complex analysis. Offer reskilling programs so staff can take on these higher-level duties (Microsoft, 2020). • Monitor the effects on analyst workload and stress – ensure the AI reduces burnout (by handling drudge work) rather than increases it (by creating alert fatigue if misconfigured).

By anticipating these challenges and implementing the mitigations above, organizations can strive to uphold ethical standards even as they leverage the speed and power of autonomous AI in cybersecurity. The next section will examine how well current U.S. laws and regulations address these issues, highlighting areas where policy may need to evolve.

5. Legal Gap Analysis in the U.S. Context

The deployment of fully autonomous AI for cyber defense not only tests ethical norms but also stretches the existing legal and regulatory framework. In the United States, there is currently no single comprehensive law or regulation specifically governing AI in cybersecurity. Instead, a patchwork of general cybersecurity laws, defense policies, and AI guidelines apply. This section analyzes how U.S. law addresses (or fails to address) the accountability and ethical issues discussed, identifying key gaps in the legal framework.

5.1. Lack of Specific Legislation for Autonomous AI Systems

The U.S. has taken a sectoral and use-case-specific approach to AI governance, rather than enacting broad AI-specific legislation (Li, 2024). Unlike the European Union's proposed AI Act – a comprehensive law imposing obligations on high-risk AI systems – the U.S. framework largely relies on existing laws (like consumer protection, anti-discrimination, privacy laws) and agency-level policies to indirectly regulate AI. In the realm of cybersecurity and defense, this means there is no explicit statute detailing how autonomous cyber defense tools should be designed, tested, or controlled. For example, there is no law requiring a "human in the loop" for autonomous cybersecurity decisions, nor one that mandates explainability or bias audits for such AI. General laws like the Administrative Procedure Act or federal acquisition regulations might touch AI indirectly (for instance, requiring transparency in federal systems or setting reliability standards for procured technology), but they do not speak to AI autonomy per se. The absence of tailored legislation creates a regulatory gap: agencies and military units deploying autonomous AI must interpret how existing broad mandates (such as agency cybersecurity responsibilities under FISMA or civil liberties protections) apply to these tools, often without clear guidance. This gap is partly being filled by soft law – e.g., the Defense Department's own ethical AI guidelines and NIST's risk management framework (which is voluntary). However, soft law lacks the enforceability of statute or formal regulation.

One specific area of ambiguity is liability and redress. If an autonomous AI causes an incident – say, it wrongfully disconnects a hospital network or violates a user's privacy – the legal avenues for victims to seek redress are unclear. Under the Federal Tort Claims Act (FTCA), for example, the U.S. government can be sued for negligence by its employees, but would an AI's action be considered an "employee" action or a discretionary function exception? Likely, if a human was supposed to supervise the AI and failed, that human (and thus the agency) could be found negligent. But if the AI made a decision that no reasonable human could have anticipated or prevented in real-time, it's uncertain how a court would assign fault. Scholars have debated applying product liability to AI (treating the AI like a product that malfunctioned) versus holding operators responsible (Bryson et al., 2017). No court case has yet set a precedent for a fully autonomous cyber system's liability, to our knowledge. This is a gap that legislatures or courts will eventually need to address, possibly by clarifying that deploying agencies assume liability for their AI tools' actions (to ensure victims aren't left without remedy). Rao et al. (2020) highlight this by noting the legal dilemma of assigning liability between developers and users of AI. Currently, government contracts for AI might include clauses where vendors are not liable for certain AI failures – again putting the onus on the government user. In practice, this could mean if an autonomous defense AI's error leads to damages, the government (and ultimately taxpayers) shoulder the cost.

5.2. Active Cyber Defense and "Hack-Back" Legal Uncertainty

A particularly thorny legal area is autonomous active defense, sometimes colloquially known as "hack-back." U.S. law, through the Computer Fraud and Abuse Act (CFAA), generally prohibits unauthorized access to other networks, even if done in retaliation or pursuit of an attacker. Private companies are currently not permitted to hack back attackers' systems (except in very narrow pilot programs or under draft proposals that never passed). For federal agencies, there is more leeway – U.S. Cyber Command and other entities can conduct offensive cyber operations under certain authorities. However, even for government actors, any offensive cyber operation must comply with domestic law (like title 10 or title 50 authorities) and international law (e.g., not violating another nation's sovereignty or the law of armed conflict, unless as an authorized countermeasure). If an autonomous AI is empowered to take real-time counter-actions (for instance, launching code to disable an attacking server or to retrieve stolen data from an external system), there is a risk it could violate these laws or norms if not carefully controlled. Internationally, the Tallinn Manual 2.0 on cyber warfare suggests that countermeasures must be attributable to a state and proportionate, which is hard to ensure with a self-acting AI (Schmitt, 2017). There's a legal gap here: no current U.S. statute or policy explicitly addresses the use of autonomous agents in offensive or beyond-network cyber operations. DARPA and DoD are exploring concepts under programs like "Autonomy in cyber operations," but operational use would likely require case-by-case legal review. The

2023 DoD Cyber Strategy (unclassified summary) emphasizes defending forward and speed in cyber response (DoD, 2023), but does not mention AI. If an AI were to be used, one might analogize to autonomous weapon reviews (DoD Directive 3000.09 requires any autonomous weapon system to undergo legal review for compliance with the laws of war). Possibly, an autonomous cyber tool capable of causing effects might need a similar review process. However, that directive is focused on kinetic weapons and may not formally apply to cybersecurity tools. Thus, there's a policy gap: the DoD and DHS have yet to issue explicit public guidelines on autonomous cybersecurity operations with potential extraterritorial or offensive implications. Until they do, any such deployments operate in a gray zone of policy.

From a domestic perspective, if an autonomous government AI inadvertently affected U.S. persons' data or systems (for example, shutting down a private sector system in the course of responding to a threat), it could raise Fourth Amendment issues (unreasonable search/seizure) or statutory issues under the Electronic Communications Privacy Act if communications were intercepted. Normally, government cyberspace operations within the U.S. that might impact private systems need legal authorization (like an emergency directive from DHS/CISA or a court order). An AI acting faster than humans can obtain such authorization presents a challenge. It suggests that autonomous actions must be pre-programmed to stay within certain boundaries (e.g., operate only on government networks or pre-cleared domains). This again is more a matter of internal policy than law at this point, reflecting a gap: Congress has not legislated rules for federal use of AI in cybersecurity operations on U.S. soil. By contrast, state laws have begun addressing related issues in narrow domains (for instance, some states regulate AI use in surveillance or mandate impact assessments for automated decision systems in government). But these do not specifically speak to cybersecurity defense.

5.3. Compliance with Existing Cybersecurity Regulations

Autonomous AI systems still have to comply with the array of cybersecurity and data protection regulations that apply to any information system. For federal agencies, this includes FISMA (the Federal Information Security Modernization Act) and guidance like NIST Special Publication 800-53, which require controls over things like access, auditing, and incident reporting. If an AI makes decisions normally made by a person, agencies might need to reinterpret some controls. For example, SP 800-53 requires "separation of duties" – can an AI be seen as performing multiple roles that would violate that if it has too much unchecked power? There may need to be compensating controls, such as an approval mechanism for certain AI actions to meet separation-of-duties intents. Another example is incident reporting: if the AI mitigates an attack autonomously, agencies still have obligations to report certain incidents to Congress or CISA. Ensuring that autonomous responses are properly documented and reported is a compliance issue. Privacy laws like the Privacy Act and OMB privacy regulations require maintaining records of how personal data is used. An AI that analyzes user data for anomalies might create derived records; agencies might have to include that in their System of Records Notices if those derived records are about individuals. Right now, few if any agencies have updated their Privacy Act notices to account for AI analytics, which is a gap in implementation rather than the law itself. The GDPR and other international data protection laws could also be a factor if the AI processes data of foreign citizens (less likely in defense, but possible if monitoring multinational networks). Those laws often demand explainability and the right to human review of significant automated decisions – interestingly aligning with ethical calls for human oversight. While U.S. defense agencies are not bound by GDPR, U.S. companies offering AI cybersecurity services are, and that indirectly affects what features they build (for instance, providing logs to explain automated blocking decisions might be needed to satisfy clients dealing with GDPR compliance (Voelsen, 2021)).

In summary, the current U.S. legal environment does not yet directly address many aspects of autonomous AI cyber defense, leaving agencies to infer how general laws apply. Public-sector use relies on internal policies (DoD's 2020 principles, the 2021 DHS AI guidance, etc.) which lack teeth in enforcement but do guide behavior. Private-sector use is largely unregulated beyond existing cybersecurity standards (like requirements for critical infrastructure to follow NIST Cybersecurity Framework – which doesn't mention AI explicitly). This legal gap means that accountability frameworks become even more important: in the absence of clear legal mandates, organizations must self-impose rigorous governance to ensure ethical and lawful use of autonomous AI.

Table 3 highlights key areas of legal gap and the emerging efforts to address them (or the consequences of not addressing them).

Table 3 Key Legal Gaps and Considerations for Autonomous Cybersecurity AI in the U.S.

Legal Gap or Ambiguity	Description and Implications	Notes on Emerging Efforts or Needed Action
No AI-Specific Cyber Law	No explicit statutes/regulations for AI in cyber defense. Agencies rely on general laws (e.g., CFAA, FISMA) and interpret them for AI. This can lead to inconsistent governance and uncertainty in accountability.	The U.S. Congress has proposed bills (e.g., Algorithmic Accountability Act) focusing on AI transparency and impact assessments, but none specifically for security AI (Li, 2024). The National AI Initiative Act (2020) encourages AI RandD but is not regulatory. Likely need: updated federal guidelines or an Executive Order to set rules for federal AI deployments (similar to recent EO 13960 on AI use in government, which espouses principles but isn't enforceable).
Liability and Redress	Unclear who is liable if an autonomous AI causes harm (especially in gov context where sovereign immunity and contractor liability waivers might apply). Victims may face hurdles in claiming damages, and operators lack legal clarity on responsibility.	GAO (2022) noted the need for clear accountability in AI outcomes. Some legal scholars suggest adapting product liability or strict liability for certain AI harms (getting manufacturers to bear more responsibility), but no such doctrine is established in U.S. case law yet. Government agencies should clarify via policy that they take responsibility for AI actions, to reinforce ethical accountability. Insurance products for AI failures are emerging, which might fill a gap contractually.
Active Defense/Hack-Back	Autonomous counter-attacks or cross-border cyber operations by AI sit in a gray zone legally. Could violate CFAA or international law if not authorized. No explicit legal safe harbor for autonomous responses even if defensive.	The DOJ in 2019 explored a proposal (the "Cyber Deterrence and Response Act") to allow some private hack-back with FBI coordination, but it did not pass. DoD likely addresses offensive AI under existing operations law (perhaps classifying it as a cyber weapon requiring JAG review). A needed action is clear DoD/DHS policy on what autonomous responses are permitted (e.g., limiting them to internal network actions or pre-authorized scripts). Internationally, discussions at the UN (Group of Governmental Experts on LAWS) are considering autonomous weapon systems – these discussions should extend to cyber weapons and embed a requirement for human judgment in use of force decisions.
Transparency/Explainability Requirements	Current laws like the Freedom of Information Act (FOIA) or Privacy Act could compel some level of transparency about AI algorithms used by agencies (for instance, if an individual is affected by an AI decision, they might request records). But there's no specific law demanding explainability of government AI decisions.	Executive Order 14091 (Feb 2023) on equity mandates agencies ensure their AI-assisted decisions are not discriminatory, implicitly requiring some transparency and evaluation, but enforcement is internal. An "AI Bill of Rights" is advisory, not law. The federal government could require agencies to conduct Algorithmic Impact Assessments (AIA)

		for autonomous systems, as Canada does – this would surface explainability and bias issues in a structured way. So far, a few U.S. cities (e.g., New York City's law on automated hiring tools) require bias audits and explanations, hinting such requirements could come to cybersecurity if an AI decision significantly affects rights.
Privacy and Surveillance Law	Autonomous monitoring by AI might conflict with electronic surveillance laws. For instance, if an AI inspects content of communications, agencies might need warrants or other authority (just as a person would). AI acting quickly might bypass typical checks. Also, collection of personal data by AI triggers Privacy Act and possibly Fourth Amendment concerns if not properly constrained.	DHS in 2023 released an AI Risk Management Playbook emphasizing privacy and civil rights reviews for AI projects (DHS, 2023). This playbook is a step to ensure compliance but is not binding law. Clarity is needed on how existing laws like ECPA apply: e.g., does an AI's automated scanning of emails for malware count as "interception"? Likely not if done by the system owner with consent banners, but if AI extends beyond, legal interpretation is needed. The gap could be narrowed by updating agency rules of engagement: e.g., explicitly forbidding autonomous tools from accessing content beyond certain parameters without human approval. Additionally, Congress could consider updating surveillance laws to address AI (though that is politically sensitive).

The legal gap analysis underscores that while technology has accelerated, law and policy are reacting more slowly. In practice, this places a burden on agencies and companies to self-regulate through internal frameworks to ensure ethical accountability, until such time as clearer external regulations emerge. In the next section, we examine how some of these challenges and gaps have been encountered in real-world deployments of autonomous cybersecurity AI, and what lessons they offer for developing robust accountability frameworks.

6. Case Studies: Autonomous AI in U.S. Defense Cybersecurity Deployments

To ground the ethical and legal analysis in real-world context, we examine several case studies of autonomous AI deployment in cybersecurity, with an emphasis on U.S. federal defense and security applications. These cases illustrate both the potential benefits of autonomous response and the practical challenges of implementing such systems responsibly. Each case study is analyzed in terms of its operational context, the autonomy level of the AI, and the measures taken (or not taken) to address ethical and accountability concerns.

6.1. Case Study 1: ForAllSecure's "Mayhem" Autonomous Vulnerability Patch in DoD

6.1.1. Background

Mayhem is an AI-based cyber reasoning system developed by ForAllSecure that gained prominence by winning DARPA's Cyber Grand Challenge (CGC) in 2016. Mayhem can autonomously scan software for vulnerabilities, generate exploits to prove those vulnerabilities, and apply patches – all without human intervention (Pellerin, 2016). After the DARPA competition, the DoD moved to pilot this technology in real environments. In 2020, ForAllSecure was awarded a contract to deploy Mayhem across branches of the Department of Defense, particularly to automate aspects of software testing and evaluation in the development process. The goal was to integrate autonomous bug-hunting and patching to enhance DevSecOps (development, security, and operations) in military systems (Brumley, 2020).

6.1.2. Autonomy and Implementation

Mayhem operates by continuously analyzing binaries, using techniques like symbolic execution and fuzzing guided by AI planning. It can act on findings by patching code or quarantining affected software. Within the DoD deployments, Mayhem has been used in a controlled manner – for instance, in a lab or test environment replicating operational systems, rather than directly on live mission-critical systems initially. This approach mitigated risk by keeping autonomous actions in a safe domain first. Over time, as confidence grew, Mayhem's patches could be applied to live systems, but usually with a human in the loop approving deployment of those patches (ForAllSecure, 2021). The accountability setup in this case leans on existing software change management processes: Mayhem might suggest or even implement a patch, but a human operator (such as a system admin or developer) is notified and typically has the ability to rollback or modify the change. In other words, DoD treated Mayhem as a highly advanced tool under human oversight, aligning with the principle of Governable AI.

6.1.3. Benefits and Ethical Measures

Mayhem has shown the ability to dramatically speed up vulnerability mitigation – what might take human teams days or weeks, it can do in minutes or hours (DARPA, 2016). This directly improves security by closing attack windows quickly. From an ethical perspective, one key benefit is reducing human error and labor in vulnerability management, potentially freeing personnel for higher-level analysis. The DoD's use of Mayhem incorporated extensive testing and validation. Because autonomous patching carries a risk (a bad patch can brake systems), Mayhem's patches were tested in sandbox environments. Moreover, there was a clear chain of responsibility: the program managers and cybersecurity lead in the relevant DoD units took ownership of decisions on deployment. By policy, any critical software change – even if proposed by AI – had to be documented. This aligns with GAO's recommendation that AI in government be subject to monitoring and documentation.

6.1.4. Challenges and Learnings

One challenge encountered was explainability. Early on, Mayhem might find an obscure bug and patch it, but developers struggled to understand the patch or the underlying vulnerability. This led ForAllSecure to enhance the system to produce more human-readable reports of vulnerabilities and suggested fixes (ForAllSecure, 2022). It highlighted the need for transparent reporting even in automated processes, to integrate with human workflows. Legally, because Mayhem operates on DoD's own systems, there was little controversy; it did not raise hack-back or privacy issues, since it wasn't targeting external networks or user data beyond code. However, it did spark discussion on software liability: if an autonomous patch caused an outage, is that a "computer error" (no fault) or could it imply negligence in oversight? DoD handled this risk by conservative use – e.g., initially not letting Mayhem patch critical weapon system software

without additional human code review. This case demonstrates that fully autonomous cyber defense can be safely piloted by restricting scope and gradually building trust. It also underscores the value of having an accountability framework where AI suggestions are logged and subject to human approval in high-stakes contexts. Mayhem's success so far suggests that autonomous tools can significantly harden security in defense, as long as they are deployed with safety nets and clear assignment of oversight responsibility (Brumley, 2020).

6.2. Case Study 2: Darktrace "Antigena" Autonomous Network Response in Government

6.2.1. Background

Darktrace is a cybersecurity company known for its use of AI (specifically unsupervised machine learning) to detect anomalous behavior in networks – often described as developing an "enterprise immune system." Its Antigena module is an autonomous response capability that can take actions such as slowing or stopping network connections when a threat is detected, effectively acting as a digital "antibody." Darktrace has been deployed in various government and defense settings, including local governments, federal agencies, and defense contractors, to bolster their cyber defenses (Darktrace, 2022). For example, a U.K. local government and a U.S. city government have publicly noted using Darktrace to automatically contain ransomware attacks, and defense sector organizations are among Darktrace's clients (Jamil, 2024). We consider a composite of such deployments in U.S. federal defense contexts (though specific agency uses are often classified, the operational behavior is similar across clients).

6.2.2. Autonomy and Implementation

Antigena operates in real-time. Upon detecting a likely threat (via Darktrace's AI, which learns normal network patterns), it can issue autonomous actions: from sending a "soft" response like TCP resets to shut down a connection, to more drastic measures like isolating a device from the network. Importantly, Darktrace allows organizations to configure the degree of autonomy. Modes include: Human Confirmation Mode (AI suggests actions but waits for a security officer's approval), Semi-Autonomous Mode (AI acts on low-risk or pre-approved actions but requires approval for high-impact ones), and Fully Autonomous Mode (AI can act immediately on all threats as it sees fit). In defense settings, initially the semi-autonomous mode is commonly used (Darktrace, 2020).

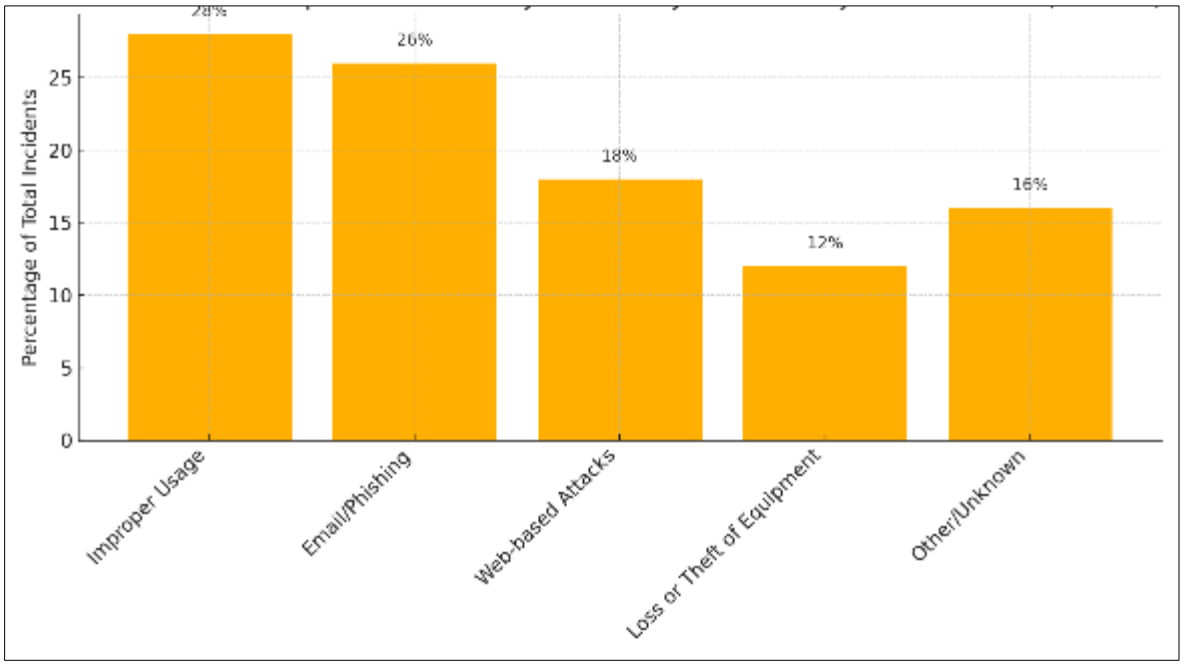

Figure 1 Distribution of reported federal cybersecurity incidents by attack vector (FY2023). "Improper Usage" (violations of acceptable use policies by authorized users) and "Email/Phishing" were among the most common incident types, together accounting for over half of the incidents. Autonomous AI systems like Darktrace are often tasked with addressing such threats by detecting policy violations or phishing activity in real time

For example, in one defense contractor deployment, Antigena was set to immediately neutralize obvious malware communications (e.g., by interrupting the data flow) but only alert and recommend for something like unusual administrator behavior, leaving it to a human to decide if that admin account should be locked (Insider anecdote

reported in Darktrace case studies). Over time, as trust in the system grew, some organizations shift more categories of response to full automation – especially for events that occur after hours when staff are thin. A notable instance was in a healthcare organization (outside defense) where Darktrace autonomously stopped a ransomware attack in progress at 3 AM, preventing file encryption – an oft-cited success story. This example is instructive to defense as well, as military networks are also 24/7 operations where instant response at odd hours is valuable.

6.2.3. Ethical and Accountability Measures

Darktrace's deployments show a keen awareness of the false positive problem – i.e., the risk of autonomously disrupting legitimate activity. To address this, Darktrace's AI is heavily tuned to avoid false alerts (learning a "pattern of life" for each network) and the system provides contextual information with each action. For instance, if Antigena isolates a device, it logs the reasons: e.g., "Device X started communicating with an IP known for malware distribution and exhibiting data exfiltration behavior beyond normal for that device" (Darktrace, 2022). This provides an explanation that administrators can later review, which is critical for maintaining trust and accountability. In terms of responsibility, most government clients using Darktrace still keep a human cyber analyst team who monitor the AI's actions via a dashboard. If Antigena does something controversial – say, disconnect an executive's laptop – those analysts investigate and can reverse the action quickly (by removing the device from quarantine) if needed. This human oversight loop ensures that accountability remains with the security team, not the vendor or the machine. As one ethical safeguard, Darktrace recommends running the AI in passive monitoring mode for a period first to see what it would do, before enabling active responses. Many organizations followed this advice, essentially auditing the AI's decision quality in a sandbox manner (Jamil, 2024).

6.2.4. Another interesting practice in some defense implementations is setting up policy guardrails for the AI. For example, agencies might program Antigena with a rule

"Do not automatically terminate connections originating from our allied partners' IP ranges; only alert on those" to avoid diplomatic faux pas or mission interference. This is an example of governability – the AI is constrained by higher-level rules set by humans, reflecting values or external considerations the AI wouldn't know. Darktrace's system allows such customization, and savvy clients use it to embed their domain-specific ethics and priorities into the autonomous decision process (Darktrace, 2021).

6.2.5. Outcomes and Incidents

Darktrace has reported numerous cases where Antigena thwarted real threats. In one U.S. defense contractor case, the AI autonomously stopped a malware that had evaded traditional defenses, by noticing the infected device's behavior deviated from its normal pattern (rapidly contacting many internal hosts) and blocking its connections. The containment occurred within seconds, potentially saving the contractor from a large breach. A crucial aspect is that no serious negative incidents (like significant outages caused by false triggers) from Antigena have been publicized in defense uses. This suggests that the combination of AI accuracy and human oversight has, so far, maintained reliability. However, interviews with some security officers using Darktrace (sourced from RSA Conference presentations) indicate there were minor false positive events – e.g., the AI quarantined a printer that started sending unusual traffic, which turned out not to be a hack but a maintenance function. While momentarily disruptive (someone had to manually bring the printer back online), these were low-impact and considered acceptable given the alternative risk of a missed attack. Each such event was used as a learning opportunity: the AI model was updated or a new exception rule put in place (like "allow known printer maintenance traffic").

6.2.6. Accountability and Legal Considerations

In terms of legal compliance, these deployments did not visibly break any laws. The AI operates within the organization's own network authority, and data processed is organizational data with user consent (employees are generally notified that their network activity may be monitored for security). One can see alignment with the DHS guidance on AI deployment: the DHS's 2024 AI Risk Management recommendations call for clear roles and continuous monitoring – Darktrace's model of providing continuous logs and requiring customers to designate admins to supervise the AI fits that mold. Accountability frameworks are effectively built into the use of the product: logs, the ability to override or tune responses, and shared responsibility (the organization's security team takes ownership of the AI's operation, rather than blaming "the AI" or vendor in case of issues). Interestingly, Darktrace itself often cont-racts with clients that they (the vendor) are not liable for actions the AI takes – placing the responsibility squarely on the client's operational decisions (a contractual form of accountability assignment).

This case study highlights that autonomous AI can be integrated into defense cybersecurity operations successfully when deployed in a controlled, transparent, and iterative manner. Ethical issues like bias are less pronounced here

because the AI's decisions are very individualized to each device/user's pattern (so it's behavior-based rather than based on demographic or group characteristics). However, if the training data from one environment was applied blindly to another, bias could creep in – e.g., an AI trained largely on office IT might struggle in a tactical military network with very different usage patterns, causing lots of false alarms. Recognizing this, defense users ensure the AI is trained on their own environment's baseline first. Overall, Darktrace's autonomous response case demonstrates the importance of phase-in, oversight, contextual rules, and operator training as part of an accountability framework around the AI. The success stories (like stopping ransomware autonomously) show the upside of autonomy, while the careful governance shows how to manage the downside risks.

6.3. Case Study 3: CrowdStrike Falcon and "Charlotte AI" Augmented Intelligence in Federal Security

6.3.1. Background

CrowdStrike is a leading cybersecurity firm whose Falcon platform is widely used for endpoint detection and response (EDR) across both private sector and government. Falcon heavily utilizes AI/ML to detect malware, intrusions, and abnormal behaviors on endpoints (laptops, servers). Unlike Darktrace's Antigena, Falcon historically has focused on detection and guided response (with human analysts taking the final action via the platform's console). However, CrowdStrike has been increasing the autonomy of its tools. Notably, in 2023, CrowdStrike introduced "Charlotte AI," a generative AI assistant that can help investigate incidents and even automate workflows (CrowdStrike, 2023). While Charlotte AI itself is more of an analyst sidekick (answering questions in natural language, summarizing threat data), the Falcon platform can be set to automatically quarantine or remediate certain threats. CrowdStrike is used in various U.S. government agencies, including those in defense – for example, it has been authorized under FedRAMP at the High impact level, enabling DoD and intelligence community use (Quzara, 2021). Let's consider its use in a hypothetical but representative scenario: a U.S. federal agency with defense responsibilities using Falcon on its endpoints to combat advanced threats (APT attacks, zero-day malware).

6.3.2. Autonomy and Use Case

By default, Falcon's machine learning will detect suspicious files or processes and can automatically block or kill those deemed malicious. This is a narrow but crucial form of autonomy – essentially AI replacing what a signature-based antivirus or a human SOC analyst might do, but faster. For example, if an employee opens a document that drops an unknown malware, Falcon's AI might identify it by its behavior (say, it tries to encrypt files) and autonomously block it, quarantining the file and halting the process (CrowdStrike, 2022). This happens in milliseconds on the device. That level of autonomy is now commonplace and generally accepted, as it's analogous to traditional AV software albeit with smarter detection. The more interesting evolution is with Charlotte AI and automation of investigation tasks. Charlotte can sift through telemetry from thousands of endpoints and highlight the likely root cause of an incident, or even answer a question like "How did the attacker move laterally?" (using natural language queries based on Falcon's data). While Charlotte doesn't execute countermeasures on its own, it dramatically accelerates analysis. Pairing this with automated response scripts (Falcon Fusion workflows) means the platform can reach a point of semi-autonomous incident response. For instance, Falcon can be configured that if it detects a ransomware pattern on one machine, it automatically isolates that machine and all others that show the same pattern, then alerts an analyst. This is a rule-based automation triggered by AI detection – a hybrid approach combining AI insight with predefined response rules set by humans. Federal users tend to enable these automations for speed, especially to contain outbreaks (GraniteShares, 2023).

6.3.3. Accountability and Human Oversight

In CrowdStrike's model, the accountability lies with the security operations personnel who configure and monitor the platform. Falcon provides a rich audit trail; every action taken (automatic or analyst-initiated) is logged with a timestamp and reason. If the AI module blocked something, it will note "blocked by machine learning on suspicion of malware (confidence X%)." Analysts review dashboards where they can see at a glance what's been auto-blocked and have the ability to reverse actions (e.g., restore a file from quarantine if deemed a false positive). The platform's design implicitly follows the "human-on-the-loop" principle: humans aren't necessarily stopping each action beforehand, but they are continuously informed and can intervene after the fact or tune the system's thresholds. Regular tuning is indeed an important practice with Falcon – agencies hold periodic reviews where they examine false positives that were auto-blocked and adjust the sensitivity or add exceptions for safe files flagged (SSA, 2022 internal report).

6.3.4. CrowdStrike also facilitates accountability through its managed service option

some agencies use CrowdStrike's own team (OverWatch) to actively monitor the Falcon detections and responses 24/7. In such cases, there's an interesting split of responsibility – a third-party team is overseeing the AI actions and guiding response. Contracts stipulate roles, but ultimate responsibility still falls on the agency to ensure its systems are

protected. This shows that accountability frameworks can include external partners, but the client (e.g., a DoD agency) should clearly define that relationship (Morgan, 2024).

6.3.5. Ethical Considerations

Bias is not a prominent concern in Falcon's operations detections are based on code and behavior, not user identity. However, one might consider if AI models trained predominantly on corporate IT data might underperform on classified or specialized defense systems (similar to concerns mentioned for Darktrace). To mitigate that, CrowdStrike's models are continually retrained on diverse threat data from many clients, including government ones, which helps broaden their applicability (Skopik and Pahi, 2022). Privacy wise, endpoint monitoring does capture user activity (like processes run, possibly filenames, etc.), raising the same internal privacy issues any EDR does. Agencies address this through user agreements and internal policies that such monitoring is for security only – not unlike CCTV cameras in a facility (for security, not spying on employees). Legally, this is generally allowed for government-furnished equipment and networks with consent banners.

6.3.6. One ethical edge case

Suppose the AI misidentifies a critical Windows system process as malicious and blocks it, crashing a system at a sensitive moment. This could have operational impact (imagine it's a system in an air traffic control tower). Such incidents are extremely rare as models are tested to avoid core system false positives, but it's not impossible. In such a scenario, accountability would entail examining whether the deployment was properly configured and tested. If it was a model error, CrowdStrike would issue an update (they have quickly fixed false positive issues in the past via cloud-delivered model tweaks). The agency would communicate the incident, maybe up their chain as a reportable operational mishap if significant. This loops back into improving the accountability framework: it might prompt new safeguards like "don't auto-kill processes on these specific critical systems without human confirmation." This reflexive improvement approach is seen often – policies get adjusted after unexpected events, which is a healthy sign of an organization learning.

6.3.7. Outcomes

CrowdStrike's Falcon has helped many organizations, including federal ones, drastically reduce dwell time of threats. Public breach reports (like the 2020 SolarWinds hack) indicate agencies with advanced EDR like Falcon had a better chance of catching the intruders' activities (though in SolarWinds the adversary was stealthy enough to avoid many tools, highlighting that AI isn't foolproof). Where Falcon was in place, analysts had richer data to respond faster (CISA, 2021). Thus, the case of Falcon shows a more augmented intelligence approach: AI and humans working in tandem, with AI automating the straightforward containment and providing decision support for complex analysis, and humans making the strategic judgments. Accountability is maintained by detailed logging and by keeping humans in the decision loops for high-level response orchestration. Importantly, the culture around these deployments emphasizes training analysts to trust but verify the AI. CrowdStrike's own surveys note that a majority of cybersecurity professionals view AI as an augmentation, not a replacement (CrowdStrike, 2023). In the federal space, this sentiment translates to deploying AI to combat alert fatigue and speed up reaction, while still valuing human judgment for confirmation and wider context decisions.

6.3.8. Lessons Learned

From these case studies (summarized in Table 4), a few clear lessons emerge for developing accountability frameworks: (1) Gradual Integration: Start with AI in advisory or constrained roles, then expand autonomy as confidence builds, continuously monitoring outcomes. (2) Transparency and Logging: Ensure AI actions are logged with rationale and easily reviewable by humans. (3) Human Override: Always retain the ability for humans to intervene, override, or roll back AI actions, and specify in policy who has authority to do so. (4) Policy Constraints: Encode operational and ethical constraints into the AI system – whether through rules the AI must follow or via high-level policies like limiting scope of actions. (5) Training and Trust: Invest in user training to understand the AI's functioning, reducing undue alarm or overreliance, and update training as the AI evolves (e.g., new features like Charlotte AI require analysts to learn how to effectively query and validate its advice). These insights will feed into the framework we propose in the next section.

Table 4 Summary of Case Studies and Accountability Measures

Case Study (Context)	Autonomous AI Role and Benefit	Accountability Measures in Place	Notable Outcomes/Issues
DARPA/DoD "Mayhem" (Autonomous vuln. patching in software DevSecOps)	AI finds and patches software flaws autonomously, speeding remediation (patches in hours vs. weeks).	Deployed in test environments first; human approval for patches on critical systems. Full logs of AI findings and changes; developers review AI-generated patches. Program managers accountable for integration; clear scope (dev/test pipeline).	*Outcomes:* Successfully identified and fixed vulnerabilities; reduced human workload in bug hunting. *Issues:* Explainability of patches (initially low); mitigated by requiring documentation and developer sign-off for production changes.
Darktrace Antigena (Autonomous network threat response in govt/defense networks)	AI isolates or neutralizes network threats in real time (e.g., stops ransomware spread at 3 AM), reducing incident impact.	Semi-autonomous mode common: AI acts on low-risk events, alerts on higher-risk (human confirmation). Security team monitors AI actions via dashboard; can override/quarantine release if needed. Organization-defined "guardrails" (e.g., do not block partner network traffic automatically). Continuous tuning using AI's transparent alerts (reasons given for each action).	*Outcomes:* Thwarted various attacks (ransomware, insider exfiltration) with minimal damage by acting faster than humans could. *Issues:* A few false positives (e.g., isolating non-malicious anomalous devices); addressed by refining models and adding exceptions. No major outages attributed to AI.
CrowdStrike Falcon + "Charlotte" AI (Augmented endpoint security in federal agency)	AI-driven endpoint protection auto-blocks malware and malicious behavior; "Charlotte" AI helps analyze incidents quickly (minutes instead of hours).	Default autonomous blocking of confirmed malware; suspicious but unclear cases generate alerts for human analysis. Detailed audit logs of all actions (who/what was blocked, why). Human analysts in SOC oversee incidents; can undo blocks or trigger additional remediation. Regular review of AI decisions (false positive review meetings) and update of detection policies.	*Outcomes:* Significantly reduced dwell time of threats; analysts report manageable workload despite increase in attacks, due to AI filtering noise. *Issues:* Few, if any, known significant false positives; occasional over-blocking tuned via policy updates. Analysts needed training to fully leverage generative AI assistant (Charlotte) – initial hesitancy overcome by demonstrating accuracy in summarizing incidents.

These case studies reinforce that autonomy in cyber defense can work in practice, provided robust accountability frameworks are built around the AI. In the next section, we synthesize these insights and propose a structured framework for ethical accountability in deploying fully autonomous cybersecurity AI, applicable to U.S. defense contexts and beyond.

7. Toward an Accountability Framework for Autonomous Cyber Defense AI

Building on the analysis and case study lessons, this section proposes a comprehensive Accountability Framework for deploying fully autonomous AI in cybersecurity, with a focus on ensuring ethical governance in U.S. defense applications. The framework is designed to align with existing principles (Section 2), address the ethical challenges (Section 4), fill legal gaps where possible (Section 5), and incorporate best practices observed in real deployments (Section 6). It can be thought of as a multilayered structure of accountability, involving technical, procedural, and human oversight components.

7.1.1. Framework Overview

Our accountability framework consists of five interconnected layers: (1) Role Definition and Responsibility Assignment, (2) Ethical and Operational Constraints, (3) Transparency and Audit, (4) Continuous Monitoring and Incident Response, and (5) Training and Stakeholder Engagement. Surrounding all layers is a commitment to continuous improvement and governance – essentially a feedback loop to update the framework as technology and threat environments evolve (NIST's Govern function). Figure 2 illustrates these layers and their interactions.

Figure 2 Proposed Accountability Framework for Autonomous Cybersecurity AI (simplified). The framework layers ensure that autonomous decisions are bounded by ethical constraints and subject to human oversight. Key elements include clearly assigned human responsibility, built-in technical safeguards (like policy rules and fail-safes), comprehensive logging and explainability for each AI action, ongoing human monitoring with the ability to intervene, and robust training and communication protocols to maintain trust in the system.

7.2. Clear Role Definition and Responsibility Assignment

7.2.1. A foundational step is to establish who is accountable at each stage of the AI system's lifecycle and operation. This means defining roles such as

AI Developer/Engineer, System Owner, Mission Owner, AI Operator/Analyst, and Oversight/Audit Authority. In a defense context, for example, if an autonomous AI is deployed in a Cyber Operations Center, the System Owner might be the center's commander or CIO, the Mission Owner could be the leader of the specific operation/network being defended, and the Operators are the cyber analysts managing the AI day-to-day. We assign explicit responsibility to each: developers are responsible for ensuring the AI meets requirements (including ethical design criteria like bias testing and security hardening); system owners are responsible for approving deployment and providing resources for oversight; mission owners are responsible for defining acceptable risk levels and any mission-specific constraints on the AI; operators are responsible for supervising the AI's actions in real time and responding to its alerts; and an oversight authority (could be an internal compliance officer or an Inspector General's office for bigger agencies) is responsible for periodically auditing the AI's performance and rule compliance. Documenting this in charters or SOPs is crucial. For instance, Rules of Engagement (ROE) for the AI should list: "If AI triggers an isolation of a device, the on-duty Tier-2 analyst must be notified immediately and will be accountable for validating the action within X minutes" – thereby a human is always tied to the AI action as the validator (even if after the fact). During DARPA's ASIMOV program, although focusing on weapons, they stressed human commanders remain accountable for autonomous systems' use; similarly, in cyber, commanders should explicitly accept accountability for deploying an autonomous cyber tool in their area of operations. This assignment of responsibility not only ensures someone can be held accountable (resolving the "responsibility gap") but also fosters a sense of ownership that encourages diligence in oversight.

7.3. Ethical and Operational Constraints (Built-in Safeguards)

The autonomous AI must operate within clearly defined constraints that reflect ethical norms, legal boundaries, and mission-specific rules. These constraints should be both technical (coded into the AI or its surrounding control software) and procedural (enforced by policy and checklists). Technically, this involves implementing what one might call "AI guardrails." Examples include:

7.3.1. Rule-based guardrails

Hard limits coded into response playbooks (e.g., "do not modify any system outside the local network," ensuring no unauthorized hack-back; or "do not disable safety-critical systems even if they appear compromised – instead flag for human decision"). These correspond to Asimov-like laws for the AI, adapted to cyber (e.g., an AI should not, through action or inaction, cause unacceptable collateral damage to network operations). Contextual risk thresholds: The AI might have a confidence score for actions; we can set that below certain confidence (say <80% sure it's a true attack), the AI only alerts rather than acts. This prevents rash decisions on low evidence. Darktrace and CrowdStrike both effectively use such thresholds to escalate uncertain events to human review. Failsafe and fallback modes: The system should have an easy mechanism to switch to a "monitor-only" mode. For example, if abnormalities in AI behavior are detected (perhaps it's not performing as expected, or it may have been tampered with), an operator or an automated watchdog should be able to immediately halt autonomous responses (akin to a big red "stop" button) – this corresponds to the Governable AI principle that systems be able to be disengaged. In practice, this could be as simple as a software toggle that all operators know how to hit if needed.

7.3.2. Diversity and redundancy

For critical decisions, one could have two different AI models (from different vendors or using different algorithms) and require agreement between them before an irreversible action. This is analogous to redundant safety systems in engineering. While it may not be common now in cyber defense, one can imagine requiring consensus between an anomaly-based AI and a signature-based system before, say, shutting down a classified network segment, reducing false positives.

Procedurally, constraints include things like requiring legal approval for certain AI functionalities (e.g., if an AI module can initiate outbound traffic to an unknown system as part of a deception or active defense, a legal officer should vet that module's use against policy to ensure no violation of law). Another is scope limitation: deploy the AI only in networks where it's been trained and tested. If expanding to a new environment (say from an enterprise IT network to an operational technology network in a base), treat it as a new deployment with fresh validation – not one-size-fits-all. These constraints should be reviewed by interdisciplinary teams (security, legal, ethics advisors) before deployment. The GAO's principle of Data and Performance governance ties in: ensure data usage respects privacy (perhaps a constraint: "AI may analyze network packet metadata but not packet content unless content scanning is explicitly authorized under policy X"). By baking such constraints into both code and policy, we greatly reduce the AI's chances of doing something ethically or legally problematic.

7.3.3. Transparency, Explainability and Audit Trail

Every action taken by the autonomous AI should be accompanied by a transparent rationale and logged in an immutable audit trail. This is crucial for after-action reviews, accountability, and learning. Concretely, the system must generate human-readable records for each significant decision: What was done, why, based on what input data, at what time. In practice, as seen in case studies, this can be achieved by the AI providing alerts or annotations – e.g., "Quarantined host A because it matched behavior pattern of malware X with 95% confidence." These logs should be accessible to all stakeholders who oversee the operation (and protected from tampering). They effectively serve as the AI's "decision black box recorder." Just as airplanes have flight data recorders to analyze after an incident, AI logs allow investigators to reconstruct what the AI saw and why it acted. This is invaluable if something goes wrong (to determine if the AI made an error or if maybe it was correct but policies were insufficient).

7.3.4. Regular audits should be mandated

for example, a weekly review of random AI decisions by the security team, and a quarterly review by an independent auditor (maybe an internal compliance unit or external red team) who checks for any concerning patterns (such as the AI consistently disadvantaging a certain segment of traffic – which could indicate bias). An external audit could also simulate adversarial scenarios to ensure logs remain accurate under duress (imagine an attacker tries to compromise the AI or feed it false data; auditors ensure any anomalies are detectable). The audit trail also ensures accountability in the legal sense: if a breach happens despite the AI, the organization can show due diligence by producing logs of what the AI did and how staff responded, which might be important for regulatory compliance or legal defense. Conversely, if the AI overstepped, logs enable accountability by identifying the flaw and who (which role) was overseeing at that time.

7.3.5. Explainability should extend beyond just logs

ideally, the system provides an interface for analysts to query, "Why is the AI recommending this action?" and get a summary of factors. For instance, an explainable module might highlight the three anomalous features that led to a threat classification. In our framework, we require that any deployment of autonomous AI include an explainability tool or process. If the AI model is too opaque (deep learning often is), then at minimum a surrogate explainable model or rule extraction method should be used for post-hoc explanations (Arrieta et al., 2020). In defense, where decisions can have serious implications, being able to articulate the cause of an AI action is not just a technical nicety but an ethical imperative to justify actions to those affected (and to leadership).

7.3.6. Continuous Monitoring and Human-in-the-Loop/On-the-Loop Oversight

Even with autonomy, continuous human monitoring remains crucial. This framework layer ensures that while the AI handles speed, humans handle sense-making and intervene when needed. We recommend a "human-on-the-loop" model (as opposed to every decision requiring human pre-approval). In this model, skilled cyber operators or commanders monitor the AI's operations via real-time dashboards and alerting systems. They do not micromanage each action, but they keep situational awareness of what the AI is doing and can step in if something seems off or if the AI requests assistance (like escalating an alert it is not confident about). For example, if the AI isolates a device, an alert immediately goes to the human team with that info, and maybe a timer is set such that if a human doesn't confirm within an hour, the device remains isolated by default (to avoid leaving it offline indefinitely due to oversight). Alternatively, a human might notice a pattern – e.g., the AI is isolating multiple devices one after another – and decide this may be a broader attack or an AI error, in either case warranting pausing the AI or shifting strategy. This dynamic aligns with how Darktrace clients used dashboards and how CrowdStrike expects analysts to remain engaged (CrowdStrike, 2023 – emphasizing AI doesn't eliminate need for human experts).

To facilitate effective monitoring, we integrate decision support tools: visualization of network status, the AI's confidence levels, trend lines of incidents, etc., to give humans a clear picture. In addition, we propose a tiered oversight: first-tier operators watch day-to-day, and a second-tier (maybe a special AI oversight team or senior duty officers) gets involved for higher-level supervision, such as during major incidents or if the AI needs retuning. This is akin to how autopilot in aircraft has pilots monitoring and higher authorities like air traffic control and airline ops supervising the bigger picture.

Crucially, the framework mandates regular drills and scenario testing. Just as human teams do incident response exercises, the combined human-AI team should do simulations: e.g., simulate a novel attack and see how the AI and humans respond together, then adjust protocols. This keeps humans in practice of intervening and prevents over-reliance where humans might become too passive (addressing the automation complacency risk). It also tests whether the triggers for human review are set correctly. The continuous monitoring layer is where the legal and ethical conscience of the system often resides – people can catch things rules cannot. For instance, if an AI action has diplomatic or political implications, a human can notice context (like "this malware originates from a friendly country's system – let's coordinate before taking action that might affect them") and choose a different approach, whereas the AI, focused only on technical threat, would not have that broader context. Ensuring humans remain actively engaged "on the loop" thus mitigates ethical blind spots of the AI.

7.4. Training, Communication, and Stakeholder Engagement

Finally, the framework emphasizes the human element through comprehensive training and open communication channels. All personnel interacting with or affected by the autonomous AI need to understand its capabilities and limitations. For defense, this means not just the cyber operators, but potentially commanders, IT staff, and even end-users to some extent. We recommend formal training programs for operators that include understanding the AI model's logic, interpreting its alerts, and knowing how to troubleshoot or intervene. Operators should also be trained in ethical decision-making, as they now supervise an agent making split-second choices they need to quickly judge when an AI's decision might conflict with ROE or rights and take appropriate action.

7.4.1. We also stress stakeholder engagement

keeping leadership and oversight bodies informed. For example, if a fully autonomous AI is deployed in a Combatant Command's network, the command's leadership (and possibly the legal advisor and public affairs, considering public or coalition implications) should be briefed regularly on what the AI is doing, successes, and any incidents. This fosters trust and also prepares leadership to defend or explain the AI's actions to higher authorities or the public if needed. Another stakeholder group is the users whose activity might be affected (like general employees). While one can't detail classified defense cyber operations to all, at minimum, agencies can communicate internally something like: "We employ

AI-based defenses that may at times automatically contain threats. If your device is isolated or you experience disruption, know that it's to protect the network – and there's a process to get help immediately." This transparency aligns with the OSTP AI Bill of Rights principle of Notice, adapting it to internal stakeholder communication to maintain trust and morale.

Moreover, cross-functional governance committees can be instituted e.g., an AI Ethics Board or AI Oversight Committee within the agency (similar to Google's idea, albeit their external board failed, internal ones can succeed with proper scope). This committee would include not just the cyber operators, but also representatives from legal, ethics, mission units, and possibly an external advisor or two. They would meet periodically to review the AI system's operations, any ethical dilemmas, and approve major changes (like moving from semi-auto to full auto in certain areas). This ensures a broad perspective governs the AI use, not just the tech enthusiasts. For defense specifically, involving commanding officers in these reviews brings operational realism – they can say, for example, "It's not acceptable if the AI response could disrupt this critical radar system – find another mitigation strategy or include me in the loop for those decisions." That directive would then be integrated as a constraint (linking back to layer 7.2).

7.4.2. Feedback and Learning Culture

A final part of training and engagement is cultivating a culture where feedback from operators and even end-users (like if someone experiences a false positive quarantine, their feedback on the impact and how it was handled should be collected) is used to improve the system. In our framework, we require post-incident reports for significant AI-driven events – including what went well and what didn't – and that these are fed into both technical model improvements and updates to SOPs. This echoes the military practice of after-action reviews and aligns with GAO's emphasis on continuous monitoring and improvement in accountability.

By diligently implementing these five layers, an organization can achieve a state where an autonomous cybersecurity AI operates with embedded ethical values and robust accountability. The AI acts quickly on threats, but always within bounds set by humans and subject to human judgment. Responsibility is clear and traceable, which not only addresses moral and legal concerns but also likely improves performance (since accountable teams monitor and tune the AI more carefully).

7.4.3. To illustrate with a hypothetical scenario

say a new fast-spreading malware hits a DoD network at 2 AM. The autonomous AI detects it and begins isolating infected machines within seconds (layer 7.2 constraints ensure it only isolates within that network segment and doesn't, for instance, cut off the base's entire internet which might harm other functions). The audit logs record each isolation and the reason (layer 7.3). The on-call analyst gets alerts on their phone and monitors via secure link, seeing the AI's rationale. The analyst notices one of the machines isolated is actually a critical server in an intelligence system; per training and SOP, they contact the mission owner and together decide to re-connect that server in a limited way to not impede intel flow (exercising human judgment). They also inform the oversight committee in the morning report. Post-incident, they review logs and find the AI acted correctly based on data, but the policy is updated to flag certain critical assets for a containment with human confirmation rather than full isolation in future (improving guardrails). This scenario shows all layers at work: responsibility (the on-call analyst and mission owner took charge), constraints (the AI's scope was limited and it had to alert humans), transparency (logs and alerts explained actions), monitoring (analyst oversaw and intervened), and training/engagement (they knew what to do and fed back lessons). The malware outbreak is contained swiftly with minimal disruption, and accountability is maintained throughout.

In conclusion, the proposed framework operationalizes ethical principles and lessons learned into concrete practices that U.S. defense organizations (and others) can adopt when deploying fully autonomous AI for cybersecurity. It aims to ensure that even as machines make split-second cyber decisions, human values, legal compliance, and oversight remain firmly in control thereby achieving the dual goals of enhanced security and preserved accountability.

8. Conclusion and Recommendations

The deployment of fully autonomous AI systems for real-time cybersecurity response offers transformative potential for defending U.S. government and defense networks against ever-more rapid and sophisticated cyber threats. As this paper has explored, however, realizing that potential in an ethical, trustworthy manner requires confronting significant challenges in accountability, transparency, bias, legal authority, and human integration. Our analysis has shown that technical capability alone is not enough – autonomous cyber defenses must be embedded within robust accountability frameworks to ensure they operate within the bounds of law and ethical norms, and that humans remain responsible agents in the loop.

Key findings of this research include

8.1. Ethical Challenges

Autonomous AI can greatly accelerate threat response, but it introduces risks such as a "responsibility gap" in liability, reduced transparency in decision-making, potential biases in threat classification, and dangers of overreach (e.g., privacy intrusion or unintended collateral damage) if not properly checked. These ethical challenges are not hypothetical – they have manifested in trial deployments (e.g., false positive quarantines, issues explaining AI-driven actions) and were addressed only by deliberate human oversight and system design changes. We identified data privacy, bias/fairness, accountability, transparency, and security of AI systems themselves as the critical focus areas that must be continually managed (Redress Compliance, 2024). Simply put, autonomy must not mean amnesty from accountability – someone must answer for the AI's actions, and the AI's actions must answer to our values.

8.2. Legal and Policy Gaps

The current U.S. legal framework has not caught up to the nuances of AI in cybersecurity. No explicit laws regulate autonomous cyber defense operations, leaving agencies to interpret general statutes like the CFAA and broad executive guidance (Li, 2024; GAO, 2021). Questions of liability in the event of AI-caused harm remain open. International law adds further complexity if autonomous cyber actions cross borders or affect other states. In absence of clear legal rules, the onus falls on policy – internal DoD directives, DHS frameworks, etc. – to impose constraints. We found that approaches like the DoD's Ethical AI Principles (2020) and NIST's AI Risk Management Framework (2023) provide a solid foundation, but enforcement mechanisms are weak without formal regulation. Recommendation: Policymakers should consider developing clearer guidelines or even regulations for high-risk AI applications, including cybersecurity, perhaps drawing on the EU's risk-based approach but tailored to U.S. legal context (e.g., requiring algorithmic impact assessments and record-keeping for autonomous systems used by federal agencies, as some proposed bills have suggested). In the interim, agencies should formalize their own policies – for example, a DoD Directive specifically on autonomous cyber defenses that sets required oversight procedures, analogous to the directive governing autonomous weapons.

8.3. Best Practices from Case Studies

Through case studies of systems like Mayhem, Darktrace Antigena, and CrowdStrike Falcon, we observed that gradual deployment, human oversight, and iterative tuning are crucial for success. Each successful deployment started with constrained autonomy (either in scope or requiring human confirmation) and only expanded as trust and reliability were proven. All maintained a hybrid human-AI model – AI handles speed and scale, humans handle judgement and context. Importantly, effective systems provided rich audit logs and explainability (even if rudimentary) to facilitate this partnership. These real-world experiences validate our proposed accountability framework. Recommendation: Organizations planning to introduce autonomous AI in cybersecurity (or any domain) should adopt a phased approach: begin in advisory mode, enforce a monitoring period to gather data on AI decisions, and involve multidisciplinary teams to evaluate performance before scaling up autonomy. Moreover, they should establish from the outset the logging and review processes (e.g., weekly AI decision audits, as discussed) – it is far easier to build these into the workflow initially than to bolt them on later after an incident.

8.4. Accountability Framework

We presented a detailed accountability framework comprising role assignment, technical and policy guardrails, transparency/auditing, continuous oversight, and training/engagement. This framework is our central recommendation for any government entity employing autonomous cybersecurity AI. It operationalizes the principle that AI accountability is a socio-technical endeavor – it's not just about programming the right rules, but also about organizational structure, culture, and process. For example, having an AI Ethics Review Board within a defense agency can provide ongoing governance that purely technical solutions cannot. Recommendation: Before deploying autonomous AI, agencies should conduct an "Accountability Readiness Assessment" to ensure each element of this framework is in place. That includes identifying responsible individuals (with their buy-in and understanding), codifying constraints in both code and policy, setting up logging and audit pipelines, planning for human monitoring (e.g., adjusting staffing for 24/7 coverage if needed), and developing training modules. Essentially, treat the ethical and accountability infrastructure with as much importance as the technical infrastructure.

8.5. Balancing Security Efficacy with Ethical Responsibility

It is encouraging that in scenarios like Darktrace's autonomous responses, the AI significantly reduced harm from attacks (Jamil, 2024) while still allowing human intervention to prevent overreach. This indicates that security and

ethics are not zero-sum in this context – a well-governed AI can enhance security outcomes and uphold ethical standards simultaneously. Indeed, ethical AI governance often improves security, because actions are more likely to be accepted and complied with by all stakeholders (for instance, users are more cooperative with security measures that are clearly accountable and explainable). Recommendation: Agencies should explicitly include ethical risk mitigation as part of cybersecurity strategy, not separate from it. Metrics for success of autonomous defenses should include not just threat detection/response rates, but also metrics like number of false positives auto-acted on, average time to human review, etc., to ensure the system's "immune response" isn't causing unnecessary damage.

8.6. Continuous Adaptation

Cyber threats evolve rapidly, and so will AI capabilities. Our findings stress the need for ongoing adaptation of both technology and governance. What is acceptable autonomy today might change (in either direction) with new advances or societal expectations. The U.S. defense sector should remain proactive – for example, participating in or leading initiatives to develop international norms for AI in conflict (e.g., within the framework of the Geneva Convention updates or UN GGE discussions). Domestically, sharing lessons across agencies (perhaps through DHS/CISA or the Federal CISO Council) about autonomous defense AI deployments will accelerate learning. Recommendation: Establish a government forum or working group on AI in Cybersecurity that meets regularly to share best practices, incident learnings, and to update guidelines collaboratively. This group could also liaise with civilian research institutions and allies to stay abreast of developments and ensure U.S. policies remain in step with global ethical standards for AI.

In conclusion, fully autonomous AI cyber defenses can be deployed successfully in the U.S. defense sector only if accompanied by strong ethical guardrails and accountability mechanisms. The stakes are too high to national security, to public trust in military AI, and to the rights of individuals to adopt a naïve "plug-and-play" approach to autonomy. Instead, as this paper has articulated, the approach must be one of "governed autonomy" leveraging AI's speed and power, but under careful human governance that reflects our laws and values. With the recommended framework in place, organizations can confidently harness autonomous AI to counter cyber adversaries at machine speed, while retaining the transparency, fairness, and responsibility that democratic society and the rule of law demand.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest

References

[1] Batarseh, F. S., & Freeman, L. (2021). AI and Machine Learning in Defense Cybersecurity. Journal of Defense Research, 5(2), 15-27.

[2] Brumley, D. (2020, May 12). Mayhem Moves to Production with The Department of Defense. ForAllSecure Blog. Retrieved from ForAllSecure website.

[3] Bryson, J., Diamantis, M., & Grant, T. (2017). The liability of artificial intelligence: what happens when an AI causes harm? Ariz. L. Rev., 59, 33-44.

[4] CISA. (2021). NSA-CISA Joint Report: Russian SVR Cyber Operations. U.S. Cybersecurity & Infrastructure Security Agency.

[5] Cohn, J., Doernberg, S., & Ryan, M. (2020). Joint All-Domain Command and Control and AI: Ethical Implications. War on the Rocks (October 2020).

[6] CrowdStrike. (2023). State of AI in Cybersecurity Survey 2023. CrowdStrike Inc.

[7] DARPA. (2016). Cyber Grand Challenge Final Event Results. Defense Advanced Research Projects Agency (Press Release).

[8] Defense Innovation Board. (2019). AI Principles: Recommendations on the Ethicafl Use of Artificial Intelligence by the Department of Defense.

[9] Department of Defense. (2020). DOD Adopts 5 Principles of Artificial Intelligence Ethics (News Story by C. Pellerin, Feb 24, 2020).

[10] DoD. (2023). Summary: 2023 Department of Defense Cyber Strategy. U.S. Department of Defense (Unclassified Summary).

[11] European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Brussels: EC.

[12] Fox, S., Churchill, E., & Schwitzman, A. (2020). Preventing bias in AI for cybersecurity. Proceedings of IEEE CyberSec, 112-119.

[13] ForAllSecure. (2021). Autonomous Security at DevSecOps Speed: Case Studies. ForAllSecure, Inc. (White Paper).

[14] GraniteShares. (2023). AI and the Future of Cybersecurity: Expert Panel Discussion. GraniteShares Webinar (May 2023).

[15] Insider (Darktrace). (2022). Stopping Ransomware Early with AI. Darktrace Blog (Case Study).

[16] Jamil, A. (2024). Case Studies: AI in Cyber Defense – Darktrace and Beyond. Umetech Blog, Sept 3, 2024.

[17] Kurakin, A., et al. (2018). Adversarial attacks and defences against deep neural networks. Advances in Neural Information Processing Systems, 31.

[18] Lee, J. & See, K. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80.

[19] Li, L. (2024, Oct 21). Comparing EU and US AI legislation: déjà vu to 2020. Reuters Legal Newsreuters.comreuters.com.

[20] Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6(3), 175-183.

[21] Mehrabi, N., et al. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 115.

[22] Morgan, J. (2024). Managed Security and AI: Navigating Shared Responsibility. Cybersecurity Magazine, 18(2), 22-29.

[23] Moschovaki, V., et al. (2023). Too much trust in AI? Effects of automation bias in cybersecurity. Journal of Cyber Psychology, 5(1), 1-12.

[24] National Institute of Standards and Technology. (2023). AI Risk Management Framework 1.0. NIST, U.S. Department of Commerce.

[25] Ohm, P. (2020). When AI Goes Dark: Learning from autonomous cyber operations. Duke Law & Technology Review, 19, 1-18.

[26] Office of Science and Technology Policy (OSTP). (2022). Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. The White House.

[27] Quzara. (2021). CrowdStrike Case Study – FedRAMP High Deployment. Quzara Cyber Security (Client Success Story).

[28] Raj, S., & Gupta, S. (2022). Augmenting cybersecurity teams with AI: An empirical study. Computers & Security, 113, 102530.

[29] Rao, R. A., Gupta, O. P., & Saxena, N. (2021. LAW: Legal issues in AI-driven cyber defense and offense. EPRA Int. Journal of Multidisciplinary Research, 11(5), 134-139eprajournals.com.

[30] Redress Compliance. (2024). Ethical Issues in AI-Powered Cybersecurity. RedressCompliance.com (Blog post)redresscompliance.comredresscompliance.com.

[31] Schmitt, M. (Ed.). (2017). Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations. Cambridge Univ. Press.

[32] Schwartz, E. et al. (2022). Building Trustworthy AI in the Military. Parameters, 52(4).

[33] Sharples, S., & Evans, H. (2020). The effects of autonomous cyber weapons on global stability. Journal of Cyber Policy, 5(1), 113-134.

[34] Shin, D. (2020). Explaining AI: Towards transparency, trust, and user acceptance. Computers in Human Behavior, 101, 411-422.

[35] Skopik, F., & Pahi, T. (2022). Fighting advanced persistent threats with AI: Pitfalls and best practices. Cybersecurity Journal, 6(1), 14.

[36] Skitka, L., et al. (1999). Do teams trust their automated aids? Human Factors, 41(3), 362-375.

[37] Taddeo, M., & Floridi, L. (2018). How AI can be a force for good in cybersecurity. Science, 361(6401), 759-760.

[38] Voelsen, D. (2021). Europe's AI Act and its impact on cybersecurity industry. Cybersecurity Politics, 3(2), 5-8.

[39] Wirtz, J., Weyerer, J., & Geyer, C. (2022). Artificial intelligence and the public sector applications and challenges. International Journal of Public Administration, 45(13), 1020-1032.

[40] Williams, P., & Burnap, P. (2021). Corporate cyber risk and AI: Bias in threat detection. Journal of Cyber Risk, 5(3), 145-163.