TRUSTED AI IN HEALTHCARE

A CISO's Framework for Balancing Innovation, Cybersecurity, and Patient Safety

A practical white paper for healthcare security, privacy, clinical, legal, compliance, and executive leaders

Prepared for healthcare organizations moving from AI experimentation to trusted AI governance

Prepared By: Kevin Royer, CISSP

July 2026

FirstLine SECURITY

Executive Summary

Artificial intelligence is no longer a future-state technology for healthcare. It is already being used by clinicians, administrators, revenue cycle teams, contact centers, researchers, security teams, and vendors. The opportunity is substantial: AI can reduce administrative burden, accelerate documentation, improve patient communications, support clinical decision-making, identify security threats, and help organizations deliver care more efficiently. But healthcare AI also changes the risk equation. In healthcare, a failure of confidentiality, integrity, availability, accuracy, or oversight can move quickly from a technical issue to a patient safety issue.

For the Chief Information Security Officer, the question is no longer whether AI should be allowed. The question is how AI can be governed, secured, monitored, and scaled without slowing responsible innovation. A blanket prohibition on AI is unlikely to succeed because employees are already bringing consumer AI tools into work. Microsoft reported in its 2024 Work Trend Index that 75% of global knowledge workers were using generative AI and that many were bringing their own AI tools to work while formal enterprise plans lagged behind adoption. [10]

Healthcare organizations need a trusted AI framework: one that enables beneficial use while creating clear boundaries for protected health information, clinical workflows, vendor review, model validation, human oversight, incident response, and ongoing monitoring. NIST describes trustworthy AI as valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. [1]

This paper translates those concepts into an operating model for healthcare leaders. It does not assume that every organization has a large AI lab or a mature data science function. Instead, it focuses on the decisions that must be made before AI reaches production: what data the tool can access, what decisions it can influence, what human review is required, what vendor commitments are enforceable, how performance will be monitored, and how the organization will respond when something goes wrong.

Trusted AI does not mean risk-free AI. It means risk-aware AI with accountable humans, protected data, monitored systems, and safety-centered governance.

75%	40%+	5 stages
global knowledge workers using generative AI	enterprises Gartner predicts may face shadow AI security or compliance incidents by 2030	from ad hoc use to trusted AI governance

Introduction: Healthcare's AI Inflection Point

Healthcare is at an AI inflection point because the technology has become accessible to nearly every workforce role, not just data science teams. Generative AI can draft clinical summaries, translate patient instructions, generate prior authorization narratives, summarize policies, support coding workflows, answer employee questions, analyze security logs, and help patients navigate care. At the same time, AI is increasingly embedded into electronic health records, medical devices, patient engagement platforms, revenue cycle systems, and third-party software.

Government and industry signals confirm that this is a governance moment, not simply a technology adoption cycle. ONC's HTI-1 rulemaking advanced transparency expectations for decision support interventions and predictive models in certified health IT. The FDA has also clarified the scope of its oversight for clinical decision support software intended for healthcare professionals, especially where clinicians must be able to independently review the basis for recommendations. [4] [5]

The World Health Organization has emphasized that AI in health can improve healthcare and medicine only if ethics and human rights are built into design, deployment, and use. CHAI has similarly called for healthcare-specific assurance standards and guidance for evaluating trustworthy AI technologies. [3] [13]

The practical implication is that healthcare AI cannot be evaluated only through a productivity lens. A tool that saves time but introduces ambiguous accountability, uncontrolled data use, or unvalidated clinical content can create more risk than value. Healthcare organizations should therefore treat AI adoption as a lifecycle management challenge that starts with intake and continues through validation, deployment, monitoring, change management, and retirement.

The CISO's AI Dilemma

The healthcare CISO is being asked to enable innovation while protecting patients, data, systems, and trust. That mandate is difficult because AI does not fit neatly into traditional security categories. It is software, but also decision support. It is data processing, but also content generation. It is a productivity tool, but also a potential clinical influence. It is vendor technology, but also something employees can access directly through consumer applications.

The dilemma has five dimensions: innovation pressure, data protection obligations, patient safety exposure, vendor opacity, and speed mismatch. Business and clinical teams want AI tools that reduce administrative burden and improve productivity. Security and privacy teams must also protect ePHI, consistent with HIPAA Security Rule risk analysis and safeguard expectations. [6]

AI tools evolve faster than policy, procurement, clinical governance, and security review cycles. A model update can change performance, output style, hallucination patterns, data handling, or integration risk. The CISO therefore cannot be only a gatekeeper. The CISO must become a trusted AI enabler who helps the organization define safe pathways for adoption.

This requires a shift in security posture. Traditional controls are still necessary, but they are not sufficient by themselves. Firewalls, identity controls, encryption, endpoint protection, and vendor questionnaires do not answer whether an AI-generated discharge instruction is clinically safe, whether a summarization model consistently omits critical facts, or whether a chatbot response could be interpreted as medical advice. The CISO must partner with clinical safety, compliance, privacy, legal, procurement, IT, and operational leaders to define risk ownership and escalation paths.

Why Patient Safety Is Now a Cybersecurity Concern

Healthcare cybersecurity has always protected data, but it now also protects care delivery. HHS connects healthcare cybersecurity performance goals to protecting patient safety, and The Joint Commission has framed cyberattacks as a patient safety concern with guidance for preserving safe care during and after a cyberattack. [7][8]

This same logic applies to AI. If ransomware can delay procedures, divert patients, disable systems, or disrupt documentation, AI failures can affect the safety and reliability of care pathways. The FDA notes that connected medical devices can be vulnerable to security breaches that may affect device safety and effectiveness. AI-enabled devices, AI-assisted decision support, and AI-integrated workflow platforms extend that concern into the algorithmic layer. [9]

AI-related patient safety risks include incorrect or fabricated clinical suggestions, biased risk scoring, erroneous summaries that omit allergies or medications, overreliance on automation, prompt injection, data poisoning, availability failures, vendor compromise, and poor auditability. Healthcare organizations should therefore expand cyber safety beyond infrastructure uptime and data confidentiality. In an AI-enabled healthcare environment, patient safety depends on secure systems, trustworthy data, reliable models, resilient workflows, and accountable human oversight.

This is especially important because AI may influence care indirectly. An administrative AI tool that drafts referral language, prioritizes call center queues, summarizes patient complaints, or suggests billing codes may shape what information reaches a clinician and when. A patient-facing AI tool may affect whether a patient seeks urgent care, follows a medication instruction, or understands next steps. Even when the AI is not formally making a clinical decision, it can influence the conditions under which clinical decisions are made.

The Rise of Shadow AI

Shadow AI is the use of AI tools without formal approval, security review, privacy review, business associate agreement evaluation, logging, or governance. It is the AI version of shadow IT, but with higher velocity and lower friction. A workforce member can paste a patient message, claims appeal, clinical note, spreadsheet, contract, or incident report into a public AI tool in seconds.

Shadow AI usually emerges for practical reasons: employees are overwhelmed, approved tools are unavailable, procurement is slow, training is limited, and consumer AI tools appear easy to use. Gartner has identified shadow AI as a top emerging risk and has warned that unauthorized AI tools can create security, privacy, compliance, and intellectual property exposure. [11][12]

The most effective shadow AI strategy combines policy, technical controls, and a credible alternative. Employees are more likely to comply when they understand which tools are approved, why PHI and confidential information cannot be entered into consumer tools, and where to go for safe AI support. Security teams should pair blocking and monitoring with approved enterprise AI environments, intake pathways for new use cases, and training that shows practical examples rather than abstract warnings. As shown in **Table 1**, shadow AI increases risk when high-impact use cases are paired with low organizational visibility.

Table 1. Shadow AI Risk Matrix

Shadow AI scenario	Likelihood	Impact	Primary risk	CISO response
PHI pasted into public chatbot	High	High	HIPAA, privacy, contractual breach	Block high-risk tools, provide approved alternatives, train workforce
AI drafts patient instructions without review	Medium	High	Patient harm, misinformation	Require human review and clinical content controls
Department buys AI SaaS outside procurement	Medium	High	Vendor risk, data exposure	Enforce intake, contract review, and AI register
AI browser extension accesses EHR or email	Medium	High	Unauthorized access, data leakage	Govern extensions and monitor with DLP
AI summarizes outdated policy documents	High	Medium	Operational error, compliance drift	Use approved knowledge sources and version control
AI-generated code used without review	Medium	High	Vulnerability injection	Secure SDLC, code scanning, peer review

Core AI Security and Governance Risks

AI risk must be organized in a way that is actionable for security, privacy, clinical, compliance, and business stakeholders. The NIST AI Risk Management Framework provides a useful foundation through its emphasis on governing, mapping, measuring, and managing AI risks.[2]

A useful taxonomy separates risks by the thing that can fail: data protection, model behavior, clinical safety, fairness, explainability, operational resilience, vendor performance, and auditability. This matters because each risk category has a different owner and a different control set. For example, data leakage may be mitigated through contract terms, DLP, access controls, and logging. Clinical reliability may require validation, workflow design, clinician review, and adverse-event reporting. Bias may require subgroup analysis and review by clinical, quality, and health equity leaders.

Table 2. AI Risk Taxonomy

Risk category	Healthcare example	Potential impact	Required governance
Privacy and PHI leakage	PHI entered into unauthorized AI tool	HIPAA violation and patient trust loss	BAA review, DLP, access controls, approved tools
Security and resilience	Prompt injection manipulates output	Unsafe workflow or data exposure	Threat modeling, red teaming, monitoring
Clinical safety	AI summary omits medication allergy	Patient harm	Human review, validation, safety escalation
Bias and fairness	Risk model underestimates acuity	Disparate care access	Bias testing and subgroup performance review
Transparency	Vendor cannot explain limitations	Poor trust and weak oversight	Model cards, documentation, source attributes
Accuracy and reliability	AI fabricates policy or clinical fact	Operational or clinical error	Grounding, validation, confidence thresholds
Vendor and supply chain	Vendor retains data or uses subcontractors	Contract, privacy, security exposure	Third-party risk review and retention controls
Auditability	No logs of prompts or outputs	Investigation gaps	Logging, retention, audit trails

A Practical AI Governance Framework for Healthcare

A practical AI governance framework should be risk-based, use-case specific, and operationally realistic. It should not require every low-risk administrative use case to go through the same review as a high-risk clinical decision support tool. At the same time, high-risk use cases should not bypass review because they are labeled as pilots, productivity tools, or workflow support.

Illustrated in **Table 3**, a healthcare AI governance operating model should connect executive oversight, clinical leadership, security, privacy, compliance, legal, and operational stakeholders into a coordinated decision-making structure for evaluating and monitoring AI use.

Table 3. Governance Operating Model

Governance layer	Accountable functions	Responsibilities
Board / Executive Oversight	CEO, CIO, CISO, CMO, Compliance, Legal	Set risk appetite, approve AI strategy, review high-risk uses
AI Governance Council	Security, Privacy, Clinical, Legal, Compliance, IT, Operations, Data, Procurement	Review use cases, maintain AI inventory, approve standards
AI Review Workstreams	Security, Privacy, Clinical Safety, Vendor Risk, Data Governance	Perform detailed assessments and define required controls
Business / Clinical Owners	Department leaders, product owners, clinical champions	Define intended use, monitor performance, ensure human oversight
Workforce Users	Clinicians, staff, analysts, administrators	Use approved tools, follow policy, validate outputs, report incidents

The framework should include an acceptable use policy, AI system inventory, risk-tiering model, security and privacy review, clinical safety review, vendor governance, monitoring and incident response, and workforce training. For enterprise healthcare AI tools, organizations should also distinguish between consumer AI use and governed enterprise AI use. OpenAI, for example, describes healthcare-oriented controls such as data residency options, audit logs, customer-managed encryption keys, and a Business Associate Agreement for ChatGPT for Healthcare. [14]

Risk tiering is the mechanism that keeps governance practical. Low-risk uses, such as internal brainstorming with no sensitive data, may require only approved tools and basic training. Moderate-risk uses, such as summarizing internal policies or drafting operational communications, should require source verification and owner review. High-risk uses, such as patient-facing messaging, clinical summarization, triage, diagnosis, treatment support, or AI connected to EHR data, should require formal review, documented controls, validation, monitoring, and executive or governance council approval.

Recommended Controls and Decision Checklist

The following controls translate AI principles into operational practices for healthcare CISOs. They are intended to support intake, procurement, pilot approval, production deployment, and periodic reassessment.

The decision checklist should be used early. Waiting until contract signature or go-live turns governance into a last-minute obstacle. A better approach is to require a lightweight AI intake before procurement, proof of concept, data sharing, or workflow design. The intake should identify intended use, users, data types, PHI involvement, clinical relevance, vendor status, integration points, human oversight, and expected benefits. From there, the governance team can right-size the review.

See **Figure 1** for a practical CISO decision checklist, and **Figure 2** for potential AI vendor review questions.

Figure 1. CISO Decision Checklist

Decision question	Required evidence
Is there a clearly defined business or clinical use case?	Use case description and accountable owner
Will the AI process PHI, confidential data, or regulated data?	Data flow diagram and data classification
Is a BAA required and in place?	Executed agreement and covered services list
Is the use case patient-facing or clinical?	Clinical safety review and approval
Can a human independently verify the output?	Human oversight plan and workflow evidence
Are prompts, outputs, and user actions logged?	Audit logging evidence and retention schedule
Is customer data used to train vendor models?	Contract terms and vendor attestation
Has the model been tested for accuracy, bias, and limitations?	Validation results and acceptance criteria
Is there a fallback process if the AI tool is unavailable?	Downtime and manual workflow plan
Is there an AI incident response process?	Playbook and escalation path

Figure 2. AI Vendor Review Questions

Domain	Questions
Data use	What customer data is processed? Is it retained? Is it used for training? Can retention be disabled?
HIPAA	Will the vendor sign a BAA? Which services are covered? Are subcontractors included?
Security	Is data encrypted in transit and at rest? Are audit logs available? Is SSO/MFA supported?
Model governance	What model is used? How often is it updated? How are changes communicated?
Validation	What healthcare-specific validation has been performed? What populations were tested?
Bias and fairness	Has subgroup performance been evaluated? How are bias concerns remediated?
Explainability	Can users understand the basis, limits, and intended use of outputs?
Incident response	How are security, privacy, model, and safety incidents reported?
Exit strategy	How can data be exported, deleted, or migrated at termination?

Figure 3. Human Oversight Model

AI use type	Human oversight requirement
Administrative drafting	User reviews before use
Policy or knowledge search	User verifies source documents
Patient communication	Qualified staff approves before sending
Clinical summarization	Clinician verifies against source record
Clinical decision support	Licensed clinician independently evaluates recommendation
Autonomous action	Prohibited unless specifically approved through high-risk governance

Human oversight should not be symbolic. It must be meaningful, documented, and built into workflow design. The FDA clinical decision support guidance is particularly relevant because it focuses on whether healthcare professionals can independently review the basis for a recommendation and avoid relying primarily on the software output. [5]

Meaningful oversight means the reviewer has the expertise, time, source material, and authority to challenge the AI output. A workflow that asks a clinician to approve hundreds of AI-generated recommendations without adequate context is not meaningful oversight. Conversely, a workflow that presents sources, limitations, confidence information, and clear escalation paths can help users apply professional judgment rather than simply accepting automation.

Roadmap: From AI Experimentation to Trusted AI

Healthcare organizations should expect AI governance to mature over time. The roadmap described in **Table 4** provides a practical sequence for moving from fragmented experimentation to trusted AI governance.

Table 4. Maturity Roadmap

Stage	Description	Key actions	Success indicators
1. Ad Hoc AI Use	Unapproved tools and limited visibility	Issue interim policy, block highest-risk tools, communicate PHI rules	Workforce understands prohibited uses
2. Controlled Experimentation	Pilots allowed with manual review	Create AI intake, AI register, risk tiers, approved sandbox	Pilots are documented and reviewed
3. Governed Adoption	Approved tools integrated into workflows	Vendor review, BAA review, logging, training, clinical safety review	Departments use approved AI pathways
4. Measured Trust	Performance, safety, and security monitored	Bias testing, model monitoring, incident playbooks, audit reporting	Leadership receives AI risk metrics
5. Trusted AI Governance	AI embedded in enterprise risk and quality programs	Board oversight, continuous assurance, lifecycle governance	AI improves care within defined risk appetite

Key maturity metrics should include the percentage of AI tools inventoried, reviewed use cases, remediated shadow AI events, workforce training completion, high-risk AI reviews, vendor AI assessments, incidents reported, clinical safety reviews, and systems with monitoring plans.

The roadmap should be treated as an organizational change program, not a one-time policy publication. Early stages focus on visibility and containment. Middle stages focus on governance repeatability and integration with procurement, IT, privacy, clinical safety, and vendor risk. Later stages focus on measurable assurance, continuous monitoring, and executive reporting. The target state is a system in which innovation teams know how to get to yes safely and risk leaders have evidence that controls are operating effectively.

Conclusion: The Future Belongs to Trusted AI

AI will become a permanent part of healthcare operations. Organizations that ignore it will face uncontrolled shadow use. Organizations that ban it outright may suppress beneficial innovation while still failing to prevent misuse. Organizations that adopt it without governance may create new privacy, security, compliance, and patient safety risks.

The better path is trusted AI: AI that is approved, explainable enough for its purpose, secure, privacy-preserving, monitored, resilient, fair, and subject to meaningful human oversight. Trusted AI does not mean risk-free AI. It means risk-aware AI. It means the organization knows where AI is being used, what data it touches, what decisions it influences, what controls are in place, who is accountable, how performance is monitored, and how incidents are handled.

For healthcare CISOs, this is a defining leadership opportunity. The CISO can help the organization move beyond fear-based AI conversations and toward a practical operating model that supports innovation while protecting patients. Cybersecurity in healthcare has already evolved from data protection to care protection. AI governance is the next step in that evolution.

The organizations that succeed will be those that make trust operational. They will have clear policies, approved tools, trained users, reviewed vendors, monitored models, accountable owners, and escalation paths for safety concerns. They will also be able to explain to patients, regulators, clinicians, and business partners how AI is being used and what safeguards are in place.

The future will not belong to healthcare organizations that adopt the most AI tools the fastest. It will belong to the organizations that earn trust.

Source List

The numbered references below correspond to citation numbers used throughout the paper.

[1] NIST AI Risk Management Framework, AI Risks and Trustworthiness.
[2] NIST AI Risk Management Framework.
[3] WHO, Ethics and governance of artificial intelligence for health.
[4] ONC, HTI-1 Decision Support Interventions and Algorithmic Transparency.
[5] FDA, Clinical Decision Support Software Guidance.
[6] HHS OCR, Guidance on HIPAA Security Rule Risk Analysis.
[7] HHS Cyber Gateway, Healthcare and Public Health Cybersecurity Performance Goals.
[8] The Joint Commission, Sentinel Event Alert 67: Preserving patient safety after a cyberattack.
[9] FDA, Cybersecurity for Medical Devices.
[10] Microsoft WorkLab, 2024 Work Trend Index: AI at Work Is Here.
[11] Gartner, Emerging Risk Deep Dive: Shadow AI.
[12] Gartner Newsroom, Critical GenAI Blind Spots.
[13] Coalition for Health AI, Blueprint for Trustworthy AI.
[14] OpenAI, Introducing OpenAI for Healthcare.

Legal Disclaimers and Copyright Protections

This white paper is provided for general informational and educational purposes only and does not constitute legal, regulatory, cybersecurity, clinical, compliance, risk management, or other professional advice. Readers should consult qualified legal counsel, compliance professionals, cybersecurity advisors, clinical leadership, or other appropriate experts before acting on any information contained in this document.

Healthcare organizations are responsible for evaluating artificial intelligence tools, cybersecurity controls, privacy obligations, patient safety considerations, vendor relationships, and governance practices based on their own operations, risk profile, contractual obligations, and applicable laws and regulations. Laws, standards, technologies, and regulatory expectations may change over time, and readers are responsible for confirming current requirements.

This white paper does not provide medical advice, clinical guidance, diagnostic recommendations, treatment recommendations, or patient-specific decision support. Healthcare professionals remain responsible for independent clinical judgment, appropriate human oversight, and compliance with applicable standards of care.

The content is provided "as is," without warranties of any kind. The authors, publishers, contributors, and affiliated organizations disclaim liability for any actions taken or not taken based on this document.

Third-party sources, frameworks, and links are referenced for informational purposes only and do not imply endorsement, sponsorship, or approval.

For permissions or licensing inquiries, contact: Kevin Royer, kevin@firstlinesecurity.net.

About the Author

Kevin Royer, CISSP, is an experienced information security executive specializing in cloud security, healthcare compliance (HIPAA), and enterprise IT infrastructure. As Senior Director of Information Security and IT at Nice Healthcare, he leads the development of scalable security programs, strengthens regulatory compliance, and oversees high-performing security and IT teams aligned to business objectives.

Kevin brings a diverse background across global enterprises, including IBM, Kaiser Permanente, American Express, and CSC, where he has led initiatives spanning risk management, security architecture, and operational resilience. He is also the founder of FirstLine Security, advising small and midsize organizations on building mature, compliance-driven IT and security programs.

In addition to his leadership role, Kevin publishes the LinkedIn newsletter The Healthcare CISO: Cybersecurity Insights for a Healthier Future, where he shares practical insights on topics such as AI governance, phishing defense, red team strategy, identity and access management, and secure system design. He also serves on the board of the ISC2 Phoenix Chapter, contributing to the growth and development of the regional cybersecurity community.