SKILLWEED HEALTHCARE SYSTEM PRESENTS

THE HEALTHCARE AI PLAYBOOK

A PRACTICAL GUIDE TO AI & ROBOTICS GOVERNANCE

FOR AUDITORS, CLINICIANS, DATA SCIENTISTS & GOVERNANCE LEADERS
ASSESSMENT PERIOD: 2026 | 53 AI SYSTEMS | 7 HIGH-RISK | 18 FDA-REGULATED

SkillWeed

CONTENTS

HOW TO USE THIS BOOK

Welcome to The Healthcare AI Playbook — your companion guide for understanding, auditing, and improving the governance of AI and robotics systems at Skillweed Healthcare System. Whether you are a clinician curious about how algorithms affect your patients, a data scientist building the next model, or an auditor evaluating compliance, this book was written for you.

WHAT'S INSIDE

This book mirrors the 10 control domains of the AI & Robotics Audit Assessment Template. Each chapter covers one domain with:

- The Big Idea — a plain-language explanation of why this domain matters
- Core Concepts — the terms and frameworks you need to know
- Real-World Examples — healthcare scenarios that bring it to life
- The Audit Lens — the exact assessment questions and what auditors look for
- Key Takeaways — what good looks like, and common pitfalls to avoid

THE MATURITY FRAMEWORK

Every domain is scored on a 0–5 scale per question, with 10 questions per domain (max 50 points each). Your maturity level tells you where you stand:

Score	Maturity Level	What It Looks Like
90–100%	⭐ Optimized	Continuous improvement; industry-leading practices; innovation mindset baked in.
70–89%	✅ Managed	Metrics-driven, well-controlled; quantitatively measured performance.
50–69%	📋 Defined	Documented standards and processes exist; consistently followed.
30–49%	🔧 Initial	Basic processes established; some documentation; inconsistent execution.
< 30%	⚠️ Ad-hoc	Unpredictable, reactive, minimal documentation. High risk zone.

TIP FOR AUDITORS

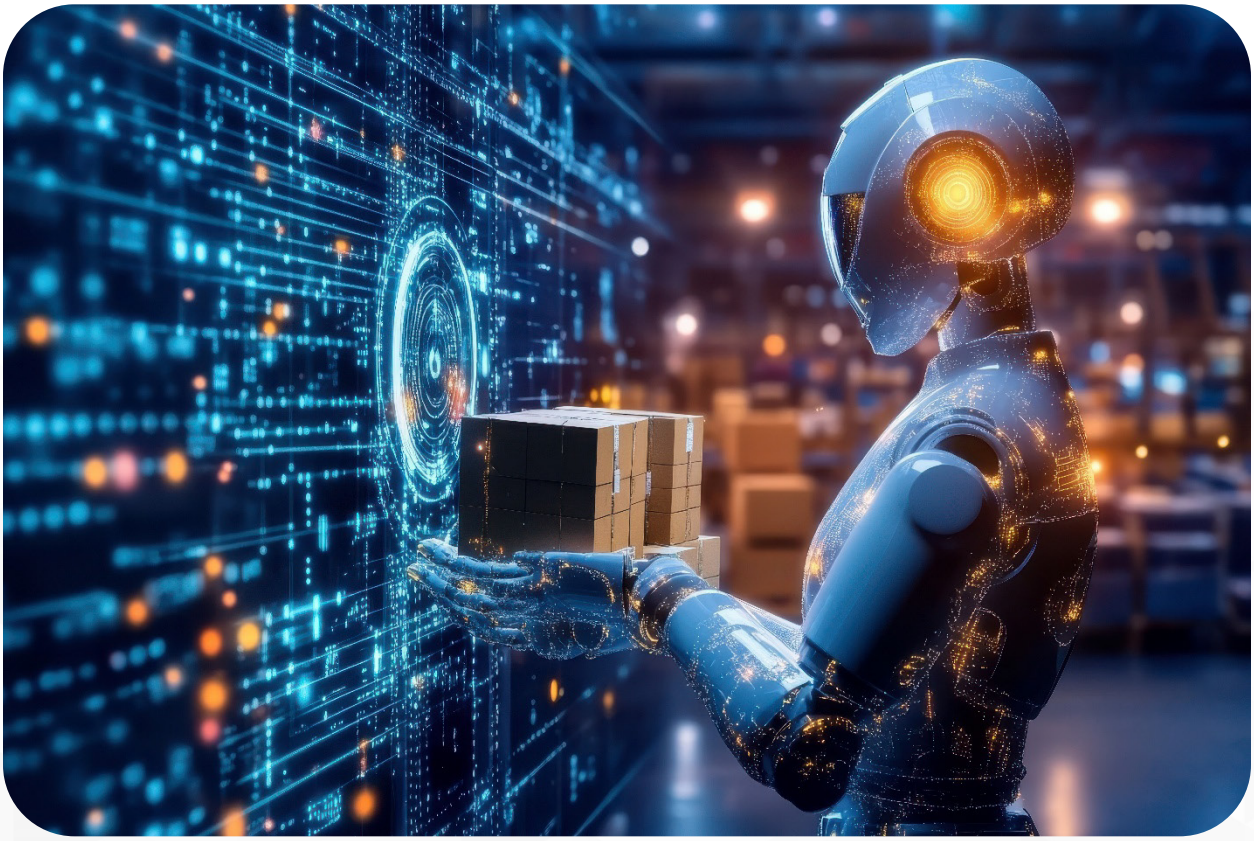
Use the assessment table at the end of each chapter to record scores during fieldwork. The Action Plan tab in the companion spreadsheet links directly to findings that need remediation.

TIP FOR LEARNERS

Read each chapter straight through once, then revisit the 'Audit Lens' section with the actual assessment template open. The combination of narrative + checklist builds lasting understanding.

CHAPTER 1:
AI SYSTEM INVENTORY & CLASSIFICATION

KNOW WHAT YOU HAVE BEFORE YOU GOVERN WHAT YOU USE



THE BIG IDEA

Imagine running a hospital pharmacy where nobody knows how many drugs are on the shelves, which ones require a prescription, or which have expired. Sounds terrifying, right? That is exactly the situation many healthcare organizations find themselves in with AI systems — and it is just as dangerous.

An AI system inventory is your master list of every algorithm, decision-support tool, and robotic system operating in your organization. Classification tells you which ones carry the highest potential for harm. Without this foundation, every other governance activity is flying blind.

HEALTHCARE REALITY CHECK

Skillweed Healthcare System currently operates 53 AI systems, 7 of which are classified as high-risk and 18 of which are regulated by the FDA. Each one touches patient care in some way — from reading radiology images to predicting sepsis onset.

CORE CONCEPTS

WHAT IS AN AI SYSTEM?

In healthcare, an AI system is any software that uses machine learning, deep learning, natural language processing, or other algorithmic techniques to analyze data and support or make decisions. This includes:

- Diagnostic imaging AI (radiology, pathology, dermatology)
- Predictive scoring tools (sepsis risk, readmission, length of stay)
- Clinical decision support (drug-drug interactions, dosing recommendations)
- Robotic surgery assistance (da Vinci-style guidance systems)
- Natural language processing (ambient documentation, prior authorization)
- Scheduling and operational optimization tools

RISK CLASSIFICATION TIERS

Not all AI is created equal. Following the NIST AI Risk Management Framework (AI RMF) and ISO 14971, systems are classified by the severity of potential harm if they fail:

Risk Level	Example Systems	Governance Intensity
🔴 **High**	Autonomous diagnostic AI, treatment recommendation engines	Full ethics review, continuous monitoring, quarterly audits
🟠 **Substantial**	Sepsis prediction, ICU early warning	Annual validation, enhanced human oversight
🟡 **Moderate**	Scheduling optimization, supply chain AI	Periodic review, standard monitoring
🟢 **Low**	Administrative chatbots, appointment reminders	Basic documentation, lightweight oversight

FDA REGULATORY STATUS

In the United States, AI systems that meet the definition of a medical device must be cleared or approved by the FDA. The two main pathways are:

- 510(k) Clearance — The system is substantially equivalent to a legally marketed predicate device.
- De Novo Classification — A novel, low-to-moderate risk device with no existing predicate.
- PMA (Pre-Market Approval) — Reserved for high-risk devices requiring clinical trial data.

Some AI tools qualify as 'clinical decision support software' (CDS) and may be exempt from device regulation under the 21st Century Cures Act — but this determination requires careful legal analysis. When in doubt, assume device status applies.

COMMON PITFALL

Many organizations acquire AI tools through departmental contracts without involving IT, Legal, or the AI Governance Committee. This 'shadow AI' problem is how you end up with untracked systems operating in clinical workflows.

REAL-WORLD EXAMPLE

Imagine Dr. Chen, a radiologist at Skillweed, who has been using an AI tool for lung nodule detection for two years. When a new compliance officer asks about it, no one can produce:

- The original vendor contract or FDA clearance letter
- Validation data showing its performance on Skillweed's patient population
- Who the designated clinical owner is
- Whether it has had any updates since deployment

This is a real scenario at many health systems. A comprehensive inventory would have captured all of this at onboarding — and a quarterly review cycle would have kept it current.

THE AUDIT LENS: ASSESSMENT QUESTIONS

The following questions are used during a formal audit of this control area. Auditors assign a score of 0–5 for each.

#	Assessment Question	Evidence to Request	Score
1	Is there a comprehensive inventory of all AI/robotics systems?	Documentation of all 53 systems with unique IDs, vendors, FDA status	___/5
2	Are AI systems classified by risk level (High, Substantial, Moderate, Low)?	Risk classification per NIST AI RMF and ISO 14971	___/5
3	Is FDA regulatory status documented for each system?	FDA 510(k), De Novo, or non-device classification	___/5

#	Assessment Question	Evidence to Request	Score
4	Are annual volumes/usage metrics tracked?	Patient encounters, procedures, scans processed per system	___/5
5	Is clinical use case clearly defined for each system?	Specific clinical application and intended use	___/5
6	Are AI system owners/responsible parties identified?	Department chairs, system administrators designated	___/5
7	Is deployment date and version history maintained?	Change log, version control, update tracking	___/5
8	Are performance metrics defined and measured?	Sensitivity, specificity, AUC, error rates, etc.	___/5
9	Is inventory updated quarterly or upon changes?	Regular review process, new system additions	___/5
10	Are decommissioned systems removed from active inventory?	End-of-life procedures, archival process	___/5

KEY TAKEAWAYS: WHAT GOOD LOOKS LIKE

A dynamic, searchable inventory updated at least quarterly. Each entry includes a unique system ID, vendor, clinical use case, FDA status, risk tier, designated owner, deployment history, and performance benchmarks. New systems trigger a governance review before go-live. Decommissioned systems are archived with data retention documentation.

Three questions every leader should be able to answer without hesitation:

1. How many AI systems are actively used in patient care right now?
2. Which ones would cause the most harm if they failed tonight?
3. Who is accountable for each one?

If any of those answers require an email and a meeting, the inventory work is not done yet.

CHAPTER 2:
AI RISK ASSESSMENT

**NOT ALL ALGORITHMS ARE EQUALLY DANGEROUS —
BUT ALL NEED TO BE EVALUATED**

THE BIG IDEA

Every clinical decision involves risk. A physician weighs the probability of harm against the probability of benefit before prescribing a medication, ordering a test, or recommending surgery. AI systems require the same disciplined thinking — but the analysis happens before the system touches a patient, and it needs to be documented, reviewed, and updated over time.

Risk assessment is the process of identifying what could go wrong with an AI system, how likely it is, how bad the consequences would be, and what you are going to do about it. Done well, it prevents disasters. Done poorly — or skipped entirely — it creates liability, harm, and headlines.

THE CORE QUESTION

For every AI system, ask: 'What is the worst clinically plausible outcome if this system fails in the worst possible way?' The answer determines how hard you work to prevent and detect that failure.

CORE CONCEPTS

THE ISO 14971 FRAMEWORK

ISO 14971 is the international standard for risk management in medical devices, and it applies directly to AI-based medical devices. Its six-step process is:

4. Risk Analysis — Identify hazards and estimate their probability and severity
5. Risk Evaluation — Compare estimated risk to acceptability criteria
6. Risk Control — Implement technical and organizational controls
7. Residual Risk Evaluation — Assess remaining risk after controls
8. Overall Risk-Benefit Analysis — Confirm benefits outweigh residual risks
9. Production and Post-Production Review — Monitor in real-world use

TYPES OF AI FAILURES IN HEALTHCARE

Understanding the failure modes helps identify risks systematically:

- False Negatives — The AI misses a finding that a clinician would have caught (e.g., missed cancer on imaging). Often the most dangerous.
- False Positives — The AI flags something incorrectly, leading to unnecessary testing, anxiety, or treatment.
- Distributional Shift — The AI was trained on data that no longer reflects your current patient population.
- Alert Fatigue — So many low-quality alerts are generated that clinicians start ignoring all alerts, including valid ones.
- Integration Failure — The AI output is not properly embedded in the workflow, leading to it being ignored or misapplied.

REAL EXAMPLE

A sepsis prediction algorithm deployed in an ED generated alerts for 45% of patients on some shifts. Nurses began ignoring the alerts entirely — including the ones that were accurate. Three patients deteriorated without timely intervention. The problem was not the algorithm's sensitivity; it was the lack of alert fatigue risk assessment at deployment.

THE RISK PRIORITY NUMBER (RPN)

A simple scoring tool adapted from Failure Mode and Effects Analysis (FMEA):

Factor	Scale	What to Consider
Severity (S)	1 (minor inconvenience) → 10 (patient death)	What is the worst clinical outcome if this failure occurs?
Occurrence (O)	1 (extremely rare) → 10 (almost certain)	How often does this failure mode actually occur?
Detectability (D)	1 (always detected) → 10 (never detected)	How quickly would a clinician catch the failure?
RPN = S × O × D	Range: 1 → 1,000	Higher RPN = prioritize for immediate risk controls

THE AUDIT LENS: ASSESSMENT QUESTIONS

#	Assessment Question	Evidence to Request	Score
1	Is risk assessment performed for all AI systems?	ISO 14971 risk management, hazard analysis	___/5
2	Are potential harms and hazards identified?	Patient safety risks, diagnostic errors, treatment delays	___/5

#	Assessment Question	Evidence to Request	Score
3	Is likelihood and severity rating documented?	Risk priority numbers, criticality assessment	___/5
4	Are risk mitigation controls implemented?	Technical and organizational controls in place	___/5
5	Is residual risk evaluated and accepted?	Risk acceptance by clinical leadership	___/5
6	Are high-risk systems subject to enhanced oversight?	Additional monitoring, review frequency	___/5
7	Is risk reassessment performed after changes?	Model updates, new use cases trigger reassessment	___/5
8	Are near-miss events analyzed for risk insights?	Incident learning, pattern identification	___/5
9	Is FDA risk classification aligned with internal assessment?	Consistency with regulatory framework	___/5
10	Are risk assessment results communicated to stakeholders?	Clinical staff, leadership, governance bodies	___/5

KEY TAKEAWAYS: WHAT GOOD LOOKS LIKE

Every AI system — regardless of risk tier — has a documented risk assessment completed before deployment. High-risk systems are reassessed after any model update, significant clinical event, or material change in the patient population. Near-miss events are systematically reviewed as a learning tool. Risk findings are shared with governance committees, not buried in a file cabinet.

CHAPTER 3:
ALGORITHM VALIDATION

TRUST BUT VERIFY — ESPECIALLY WHEN MACHINES ARE MAKING THE CALL

THE BIG IDEA

A new drug must prove it works through clinical trials before it reaches patients. An AI algorithm deserves the same scrutiny. Validation is the process of rigorously testing whether an AI system actually does what it claims to do — on your patients, in your environment, with your data.

The uncomfortable truth: an algorithm that performs brilliantly in a published research paper may perform poorly when deployed at your specific institution. Differences in patient demographics, documentation practices, EHR configurations, and clinical workflows all affect real-world performance. Validation is how you find out before patients are harmed.

SKILLWEED ACTION ITEM

Two radiology AI systems currently lack external validation. This is classified as a HIGH-risk finding in the Action Plan, with a target completion date for conducting external validation using data from a partner hospital.

CORE CONCEPTS

TYPES OF VALIDATION

- Internal Validation — Testing on data from the same institution that trained the model. Necessary but not sufficient.
- Temporal Validation — Testing on data from a more recent time period than the training data. Detects whether performance degrades over time.
- External Validation — Testing on data from a completely different institution. The gold standard for generalizability.
- Prospective Validation — Testing in a live clinical environment before full deployment. The closest to real-world conditions.

KEY PERFORMANCE METRICS

Metric	What It Measures	Why It Matters
Sensitivity	True positive rate: how often the AI catches a real finding	Critical for screening tools — missing a diagnosis is dangerous
Specificity	True negative rate: how often the AI correctly rules out a finding	Critical for treatment decision tools — over-treating is costly and harmful
AUC-ROC	Overall discrimination ability across all decision thresholds	AUC > 0.80 is generally considered clinically useful
PPV / NPV	Positive/Negative Predictive Value — accounts for disease prevalence	A test with great sensitivity may have low PPV in a low-prevalence population
Calibration	Does the predicted probability match the actual event rate?	An AI that says '80% risk' should be right about 80% of the time

SUBGROUP ANALYSIS: THE HIDDEN REQUIREMENT

Overall accuracy numbers can hide serious performance gaps in specific patient populations. A model may show 90% sensitivity overall, but only 70% sensitivity in elderly patients, patients with multiple comorbidities, or patients from certain ethnic backgrounds. Subgroup analysis is the practice of systematically examining performance across:

- Age groups (pediatric, adult, geriatric)
- Gender and sex
- Race and ethnicity
- Disease severity or acuity
- Insurance status or socioeconomic proxy variables

Failing to perform subgroup analysis is one of the most common — and consequential — gaps in AI validation programs.

THE CALIBRATION CONCEPT MADE SIMPLE

Imagine a weather app that says 'there is a 70% chance of rain' every single day for a month. If it actually rained on 70 of those days, the app is well-calibrated. If it only rained on 20 days, the app is overconfident. Well-calibrated AI gives clinicians reliable confidence scores they can actually use in decision-making.

THE AUDIT LENS: ASSESSMENT QUESTIONS

#	Assessment Question	Evidence to Request	Score
1	Is validation performed before clinical deployment?	Internal, temporal, and external validation	___/5
2	Are performance metrics clearly defined?	Sensitivity, specificity, PPV, NPV, AUC-ROC	___/5
3	Is validation performed on independent test set?	Hold-out data not used in training	___/5

4	Are subgroup analyses conducted?	Performance by age, gender, race, disease severity	___/5
5	Is temporal validation performed?	Testing on recent data to detect drift	___/5
6	Are validation results documented and reviewed?	Statistical reports, clinical review	___/5
7	Is clinical endpoint validation performed?	Patient outcomes, diagnostic accuracy	___/5
8	Are false positive/negative rates acceptable?	Clinical thresholds defined and met	___/5
9	Is model calibration assessed?	Brier score, calibration plots	___/5
10	Are validation results compared to clinical benchmarks?	Physician performance, existing standards	___/5

KEY TAKEAWAYS: WHAT GOOD LOOKS LIKE

No AI system touches a patient without documented validation on an independent test set — not the same data used for training. High-risk systems achieve external validation before or shortly after deployment. Subgroup analyses are conducted and disclosed. Calibration is measured, not assumed. Validation results are compared against clinician performance and documented benchmarks.

CHAPTER 4:
DATA GOVERNANCE

GARBAGE IN, GARBAGE OUT — AND IN HEALTHCARE, GARBAGE CAN KILL

THE BIG IDEA

An AI model is only as good as the data it was trained on. Poor training data produces poor models — biased, fragile, or simply wrong in ways that are difficult to detect until something goes wrong clinically. Data governance is the discipline of ensuring that the data feeding your AI systems is accurate, representative, secure, and ethically obtained.

Healthcare data is uniquely sensitive. It contains Protected Health Information (PHI), carries legal obligations under HIPAA, and reflects the lives and health of real people who trust their care team. That trust extends to how their data is used to train the machines that now influence their care.

THE DATA QUALITY CHECKLIST

- Completeness — Are all required fields populated? Missing data can bias models in subtle ways.
- Accuracy — Are diagnoses, lab values, and timestamps correct? Transcription errors and EHR copy-paste culture create noise.
- Representativeness — Does the training set reflect the full diversity of the population the model will serve?
- Timeliness — Is the data recent enough to reflect current clinical practice? Clinical guidelines change; older data may encode outdated care patterns.
- Consistency — Are the same concepts labeled the same way across data sources? ICD coding practices vary by institution and coder.

HIPAA & DE-IDENTIFICATION

Before any patient data is used to train an AI model, it must be properly de-identified under either the Safe Harbor Method (removing 18 specific identifiers) or the Expert Determination Method (statistical analysis confirming re-identification risk is very small). Failure to de-identify correctly is a HIPAA breach — regardless of intent.

ANNOTATION QUALITY & INTER-RATER RELIABILITY

Many healthcare AI models rely on labeled training data — images marked as 'cancer' or 'normal,' notes labeled with a diagnosis. The quality of these labels directly determines the quality of the model. The gold-standard measure for annotation consistency is Cohen's Kappa:

- $\kappa > 0.80$ — Almost perfect agreement. Excellent.
- $\kappa = 0.61–0.80$ — Substantial agreement. Good.
- $\kappa = 0.41–0.60$ — Moderate agreement. Investigate disagreements.
- $\kappa < 0.40$ — Fair to poor agreement. The annotations are unreliable; the model trained on them will be too.

Best practice is to require $\kappa > 0.70$ for clinical AI annotation projects, with explicit annotation guidelines and periodic calibration sessions among annotators.

DATASHEETS FOR DATASETS

The AI research community has developed a practice of 'datasheets for datasets' — structured documentation that describes what data was collected, how, from whom, under what consent, with what known limitations. Model cards are the equivalent for trained models. Both should be standard documentation for every AI system at Skillweed.

THE AUDIT LENS: ASSESSMENT QUESTIONS

#	Assessment Question	Evidence to Request	Score
1	Is training data quality assessed?	Missing data, outliers, label accuracy	___/5
2	Are datasets adequately sized per statistical requirements?	Power analysis, minimum sample sizes	___/5
3	Is data representative of target population?	Demographics match clinical population	___/5
4	Are annotation quality and inter-rater reliability measured?	Cohen kappa >0.7, annotation guidelines	___/5
5	Is PHI properly de-identified in training data?	HIPAA Safe Harbor or Expert Determination	___/5
6	Are data security controls implemented?	Encryption, access controls, audit logging	___/5
7	Is data lineage and provenance documented?	Data sources, transformations, versioning	___/5
8	Are dataset imbalances identified and addressed?	Class distribution, oversampling/undersampling	___/5
9	Is training data periodically refreshed?	Currency of data, concept drift monitoring	___/5
10	Are datasheets/model cards maintained?	Dataset documentation per best practices	___/5

KEY TAKEAWAYS: WHAT GOOD LOOKS LIKE

Every training dataset has a datasheet documenting its source, size, demographics, known limitations, and de-identification method. Annotation projects require demonstrated inter-rater reliability above $\kappa = 0.70$. Data imbalances are identified and explicitly addressed. Data is periodically refreshed to prevent concept drift. PHI handling is auditable with access logs.

CHAPTER 5:
BIAS & FAIRNESS TESTING

WHO GETS THE BENEFIT OF AI — AND WHO GETS LEFT BEHIND?

THE BIG IDEA

Healthcare has a long history of disparities — outcomes that differ by race, gender, income, and geography in ways that cannot be explained by biology alone. AI systems trained on historical healthcare data can inherit, amplify, and automate these disparities at scale. Bias and fairness testing is the systematic effort to detect and mitigate these inequities before they harm patients.

This is not just an ethical obligation — it is increasingly a regulatory one. The FDA's action plan for AI and machine learning, the EU AI Act, and evolving CMS guidance all point toward mandatory equity requirements for healthcare AI. Getting ahead of this now is smart strategy.

SKILLWEED ACTION ITEM

Current equity testing reveals a 12% performance gap for certain demographic subgroups in one deployed model. This is a HIGH-risk finding. The action plan calls for retraining with a balanced dataset and implementing mitigation strategies — with the AI Ethics Board overseeing the remediation process.

TYPES OF BIAS IN HEALTHCARE AI

- Historical Bias — The training data reflects historical discrimination. Example: a model trained on data from an era when certain groups had less access to diagnostic testing learns that those groups have fewer diagnoses — and under-flags them going forward.
- Representation Bias — Certain groups are underrepresented in training data. A dermatology AI trained predominantly on lighter-skinned patients may perform poorly on darker-skinned patients.
- Measurement Bias — Proxy variables used in models (e.g., healthcare cost as a proxy for health need) can encode socioeconomic bias.
- Feedback Loop Bias — Model outputs influence future data collection, reinforcing existing disparities over time.

FAIRNESS METRICS

There is no single definition of 'fairness' — different metrics capture different equity goals:

Fairness Metric	What It Means	When to Use It
Demographic Parity	Equal positive prediction rates across groups	When equal access to a beneficial intervention is the goal
Equal Opportunity	Equal true positive rates across groups	When it matters most that all groups benefit equally from correct predictions
Equalized Odds	Equal true positive AND false positive rates	When both types of errors have significant consequences

Fairness Metric	What It Means	When to Use It
Disparate Impact (80% Rule)	Minority group selection rate ≥ 80% of majority rate	A commonly used legal standard for employment; adapted for clinical AI

BIAS MITIGATION STRATEGIES

- Pre-processing — Fix the training data: resample underrepresented groups, reweight samples, or collect additional targeted data.
- In-processing — Build fairness constraints directly into the model training algorithm.
- Post-processing — Adjust model thresholds or outputs differently for different subgroups to equalize key metrics.

No mitigation strategy is a silver bullet. Each involves trade-offs. The key is transparency: document what was done, why, and what trade-offs were accepted.

THE FAIRNESS PARADOX

It is mathematically impossible to satisfy all fairness metrics simultaneously when base rates differ across groups. This is not a reason to give up — it is a reason to have explicit conversations with ethics boards and clinical leadership about which fairness definition best aligns with your organizational values and the specific clinical context.

THE AUDIT LENS: ASSESSMENT QUESTIONS

#	Assessment Question	Evidence to Request	Score
1	Is bias testing performed across demographic subgroups?	Age, gender, race, ethnicity, insurance status	___/5
2	Are fairness metrics calculated?	Demographic parity, equal opportunity, equalized odds	___/5
3	Is disparate impact assessed?	80% rule for protected characteristics	___/5
4	Are performance differences statistically tested?	Significance testing for subgroup differences	___/5
5	Are identified disparities investigated?	Root cause analysis, data or model issues	___/5
6	Are bias mitigation strategies implemented?	Pre-processing, in-processing, post-processing	___/5
7	Is equity testing repeated after model updates?	Retraining triggers new fairness assessment	___/5
8	Are results reviewed by AI Ethics Board?	Ethics committee oversight of equity	___/5
9	Is equity performance monitored in production?	Ongoing subgroup tracking	___/5
10	Are disparities communicated to clinical users?	Transparency about limitations	___/5

KEY TAKEAWAYS: WHAT GOOD LOOKS LIKE

Bias testing is mandatory for every clinical AI system before deployment and after every model update. Results are reviewed by the AI Ethics Board. Identified disparities trigger formal investigation and documented remediation. Clinical users are informed of known limitations. Equity monitoring is ongoing — not a one-time checkbox.

CHAPTER 6:
EXPLAINABILITY & TRANSPARENCY

IF YOU CAN'T EXPLAIN IT, YOU CAN'T DEFEND IT — OR TRUST IT

THE BIG IDEA

A clinician makes a recommendation. The patient asks 'Why?' The clinician explains — the lab values, the imaging findings, the clinical history, the guidelines. That explanation is the foundation of informed consent, professional accountability, and patient trust.

Now an AI system makes a recommendation. The patient asks 'Why?' If the answer is 'the algorithm said so, and we can't really explain it,' that is a clinical, ethical, and legal problem. Explainability is the capacity to provide human-understandable reasons for AI outputs. It is not optional in high-stakes healthcare contexts.

THE EXPLAINABILITY SPECTRUM

- Inherently Interpretable Models — Logistic regression, decision trees, linear models. You can read the formula and understand exactly what drives each prediction. Preferred when clinical stakes are high and data complexity permits.
- Post-hoc Explanation Methods — Applied to 'black box' models (deep neural networks) to approximate their reasoning:
  - SHAP (SHapley Additive exPlanations) — Quantifies how much each input feature contributed to a specific prediction.
  - LIME (Local Interpretable Model-agnostic Explanations) — Builds a local, interpretable approximation around a specific prediction.
  - Attention Maps / Saliency Maps — In imaging AI, highlights which pixels or regions most influenced the model's output.
  - Counterfactual Explanations — 'The AI would have changed its recommendation if your creatinine were 1.2 instead of 2.4.'

SHAP IN PLAIN ENGLISH

Imagine your sepsis risk score is 0.78. SHAP would tell you: 'Heart rate contributed +0.15, white blood cell count contributed +0.22, temperature contributed +0.08, and normal urine output contributed -0.07.' That story is something a clinician can validate, question, or override with clinical judgment.

MODEL CARDS: DOCUMENTATION FOR TRANSPARENCY

A model card is a one-to-two page document accompanying every AI system that discloses:

- What the model does and who it is intended for
- What data it was trained on (size, source, demographics)
- How it was validated and what metrics it achieved

- Known limitations and failure modes
- Ethical considerations and potential for bias
- Performance across demographic subgroups

Model cards are becoming a regulatory expectation. Think of them as the package insert for an AI drug.

THE AUTOMATION BIAS TRAP

Studies show that clinicians often over-rely on AI recommendations — even when those recommendations are wrong — simply because the AI presented them confidently. Training clinicians to critically evaluate AI outputs (not just accept them) is as important as making the explanations available in the first place.

THE AUDIT LENS: ASSESSMENT QUESTIONS

#	Assessment Question	Evidence to Request	Score
1	Are explainability methods implemented?	SHAP, LIME, attention mechanisms, saliency maps	___/5
2	Are model cards/documentation provided?	Model architecture, training data, limitations	___/5
3	Can clinicians understand AI reasoning?	Human-readable explanations, feature importance	___/5
4	Are prediction uncertainties communicated?	Confidence scores, prediction intervals	___/5
5	Are model limitations clearly documented?	Known failure modes, edge cases	___/5
6	Is training data characteristics disclosed?	Source, size, demographics, quality	___/5

#	Assessment Question	Evidence to Request	Score
7	Are clinicians trained on AI interpretation?	Education on outputs, limitations, appropriate use	___/5
8	Is explainability validated with clinical users?	Comprehension testing, usability studies	___/5
9	Are decisions attributable to specific inputs?	Traceability of reasoning	___/5
10	Is model performance transparently reported?	Metrics, validation results, ongoing monitoring	___/5

KEY TAKEAWAYS: WHAT GOOD LOOKS LIKE

Every high-risk AI system has a model card and implements at least one explainability method accessible to end users. Clinicians receive training on interpreting AI outputs — including when to question or override them. Confidence scores or uncertainty ranges accompany all predictions. Patient-facing disclosure explains that AI is used in their care.

CHAPTER 7:
SAFETY MONITORING

DEPLOYMENT IS NOT THE FINISH LINE — IT IS THE STARTING LINE

THE BIG IDEA

Clinical trials end at approval. Patient safety monitoring begins at approval. The same logic applies to AI systems — validation before deployment is necessary but not sufficient. Real-world performance differs from test-set performance. Patient populations shift. Clinical workflows evolve. Models drift. Problems that were invisible in testing emerge in production.

Safety monitoring is the continuous process of watching what your deployed AI systems actually do in the real world — and acting quickly when something is not right.

SKILLWEED ACTION ITEM

Alert fatigue metrics are currently exceeding acceptable thresholds for one safety monitoring system. This is a MEDIUM-risk finding. The CMIO is responsible for optimizing alert thresholds and implementing alert prioritization by the target date.

WHAT TO MONITOR

MODEL DRIFT DETECTION

- Data Drift — The statistical distribution of input variables shifts over time (e.g., patient demographics change, new clinical documentation practices take hold).
- Concept Drift — The relationship between inputs and outcomes changes (e.g., a new treatment protocol means that the same lab values now predict different outcomes than they did during training).
- Label Drift — The prevalence of the outcome being predicted changes in the real world.

Drift detection requires monitoring input and output distributions continuously — not just checking performance metrics quarterly.

KEY SAFETY METRICS TO TRACK

Metric	What It Tracks	Threshold to Investigate
Sensitivity / Specificity	Core performance — is the model still accurate?	> 5% degradation from validation baseline
Alert Override Rate	How often clinicians dismiss AI recommendations	> 30% override rate may signal poor model or poor integration
Alert Fatigue Ratio	False alert rate relative to true alerts	Should align with clinical tolerance defined in deployment specs

Metric	What It Tracks	Threshold to Investigate
Adverse Event Rate	Patient harm events associated with AI use	Any event triggers formal review; trends trigger redesign
Time to Action	Lag between AI alert and clinician response	Increasing lag may indicate alert fatigue or workflow friction

FDA MEDWATCH REPORTING

If an AI system is an FDA-cleared or approved medical device, adverse events associated with it may require MedWatch reporting. Mandatory reporting applies when the device may have caused or contributed to serious injury or death. Voluntary reporting is encouraged for any safety concern. Failure to report required events is a regulatory violation.

CORRECTIVE ACTION PLAYBOOK

When safety monitoring identifies a problem, escalation should follow a pre-defined playbook: (1) Document the finding. (2) Assess clinical impact immediately. (3) If high-risk, consider temporary suspension. (4) Notify governance committee. (5) Conduct root cause analysis. (6) Implement fix and re-validate. (7) Document and close the loop.

THE AUDIT LENS: ASSESSMENT QUESTIONS

#	Assessment Question	Evidence to Request	Score
1	Is ongoing performance monitoring in place?	Continuous KPI tracking, dashboards	___/5
2	Are adverse events tracked and investigated?	Incident reporting, root cause analysis	___/5
3	Is FDA MedWatch reporting performed when required?	Timely reporting of device-related events	___/5
4	Are near-miss events captured and analyzed?	Learning from close calls	___/5
5	Is model drift detected and addressed?	Data drift, concept drift, label drift	___/5
6	Are alert fatigue metrics monitored?	False positive rates, override rates	___/5
7	Are clinical outcomes tracked?	Patient harm, diagnostic errors, treatment appropriateness	___/5
8	Is safety data reviewed by governance committees?	AI Safety Board, Clinical AI Review Board	___/5
9	Are safety issues escalated appropriately?	Clear escalation paths, decision authority	___/5
10	Is corrective action taken promptly?	Model fixes, alerts, retraining, deactivation	___/5

KEY TAKEAWAYS: WHAT GOOD LOOKS LIKE

Every deployed AI system has an active monitoring dashboard with defined KPIs and thresholds. Adverse events and near-misses are reported in a standardized system and reviewed at AI Safety Board meetings. Drift detection runs continuously. Alert fatigue metrics are tracked. The escalation path from monitoring alert to corrective action is documented and tested.

CHAPTER 8:
HUMAN OVERSIGHT

AI IS A TOOL. THE CLINICIAN IS STILL THE PROFESSIONAL.

THE BIG IDEA

In 2016, a chess-playing computer beat the world champion. Nobody suggested eliminating chess professionals — or letting the computer make decisions about who gets to play at all. The relationship between AI capability and human judgment is collaborative, not zero-sum.

In healthcare AI, human oversight means that no AI system makes autonomous consequential decisions about patient care without meaningful human review and the genuine ability to override. It also means designing systems that support — rather than overwhelm or replace — clinical judgment.

THE HUMAN-IN-THE-LOOP PRINCIPLE

At Skillweed, the policy is clear: every AI system that influences a patient care decision requires human review before that decision is implemented. There are no exceptions for high-risk systems. The question is not 'Should humans be involved?' — it is 'How do we design the workflow so that human involvement is meaningful, not just nominal?'

DESIGNING FOR MEANINGFUL OVERSIGHT

THE OVERRIDE CAPABILITY REQUIREMENT

Every AI system must allow clinicians to override recommendations with appropriate ease. An override mechanism is not meaningful if:

- It requires five extra clicks in a rushed workflow
- There is no place to document the reason for the override
- Override data is never reviewed or acted upon

Override rates are a signal. Very high rates (> 30-40%) may indicate the model is not useful or not trusted. Very low rates (< 5%) on high-stakes recommendations may indicate automation bias — clinicians deferring to AI even when they have concerns.

ALERT DESIGN PRINCIPLES

Poor alert design is a patient safety issue. The best-practice principles:

- Actionable — Every alert should prompt a specific, clear clinical action.
- Prioritized — Not all alerts are equal. Tier them by urgency.
- Contextual — Show why the alert fired, not just that it fired.
- Non-interruptive where possible — Use passive notifications for lower-acuity signals.
- Threshold-adjustable — Allow local customization to match clinical context.

THE NIELSEN NORMAN WISDOM

Usability experts found that clinicians spend an average of 3.2 hours per day managing EHR-related interruptions, including clinical decision support alerts. Poorly designed AI alerts add to this burden. Good AI alert design reduces it. The question 'Does this alert help clinicians care better, or just comply better?' should guide every design decision.

CLINICIAN TRAINING: THE MISSING PIECE

A technically excellent AI system will underperform if clinicians do not understand how to use it appropriately. Training should cover:

1. What the system does and does not do (scope and limitations)
2. How to interpret the outputs, including confidence scores
3. When to follow the recommendation vs. when to question it
4. How to document an override and why it matters
5. Where to report concerns or unexpected behavior

THE AUDIT LENS: ASSESSMENT QUESTIONS

#	Assessment Question	Evidence to Request	Score
1	Do all AI systems require human review?	No autonomous decisions without clinician oversight	___/5
2	Can clinicians override AI recommendations?	Override capability in all systems	___/5
3	Are overrides documented with reasons?	Structured data capture, free text	___/5
4	Is override rate monitored?	Tracking and trending by system, clinician	___/5

#	Assessment Question	Evidence to Request	Score
5	Are alerts designed appropriately?	Clear, actionable, prioritized, non-interruptive	___/5
6	Is workflow integration optimized?	Fits clinical workflow, minimizes burden	___/5
7	Are clinicians trained on AI systems?	Initial and ongoing education	___/5
8	Is clinician feedback collected?	Usability, accuracy, workflow impact	___/5
9	Are alert thresholds adjustable?	Sensitivity/specificity trade-offs, local customization	___/5
10	Is human-AI collaboration evaluated?	Combined performance vs. human or AI alone	___/5

KEY TAKEAWAYS: WHAT GOOD LOOKS LIKE

Every AI system has a documented, accessible override mechanism. Override rates are monitored, trended, and reviewed. Alert design follows usability standards and is validated with end users. Clinicians complete structured training before using any clinical AI system. Human-AI team performance (not just AI-alone performance) is tracked and reported.

CHAPTER 9:
MODEL MAINTENANCE & RETRAINING

THE MODEL YOU DEPLOYED YESTERDAY MAY NOT BE THE MODEL YOU NEED TODAY

THE BIG IDEA

AI models are not static artifacts. They are dynamic systems operating in a changing world. Clinical practices evolve. Patient populations shift. New diseases emerge. Health systems adopt new EHR versions and documentation standards. Drug formularies change. A model trained two years ago on yesterday's reality may be quietly wrong about today's.

Model maintenance is the planned, systematic process of keeping AI systems accurate, current, and clinically relevant throughout their lifecycle. Retraining is one tool — but it is not the only tool, and it is not free. It requires the same rigor as initial validation.

THE MODEL LIFECYCLE ANALOGY

Think of an AI model like a clinical protocol. Protocols are evidence-based when written, but evidence evolves. Protocols are reviewed, updated, and retired when better options exist. AI models deserve the same lifecycle discipline — regular review, evidence-triggered updates, and formal retirement with transition planning.

TRIGGERS FOR RETRAINING

Retraining should not be scheduled arbitrarily or triggered only by disaster. A mature maintenance program defines explicit triggers:

- Performance-based — A KPI drops below a defined threshold (e.g., sensitivity falls more than 5% below validation benchmark)
- Data-based — Significant drift detected in input variable distributions
- Clinical-based — A major change in clinical practice, guidelines, or patient population
- Event-based — A serious adverse event is attributed to model failure
- Time-based — Annual mandatory review, regardless of performance metrics (for high-risk systems)

THE RETRAINING-VALIDATION CYCLE

Retraining a model does not automatically make it better. Every retrained model must go through the same validation rigor as the original:

1. Data quality control on the new training data
2. Independent test set evaluation
3. Subgroup analysis and bias testing
4. Calibration assessment
5. Clinical review and sign-off
6. A/B testing in production before full rollout
7. FDA notification if the change is significant (for regulated devices)

THE FDA SAMD CHANGE PROTOCOL

For AI-based Software as a Medical Device (SaMD), the FDA's recent guidance distinguishes between 'locked' algorithms (fixed after training) and 'adaptive' algorithms (that continue learning in deployment). Significant changes to a cleared device — including major retraining — may require a new 510(k) submission. This determination should involve your regulatory counsel.

VERSION CONTROL & ROLLBACK

Every model version should be tagged, archived, and retrievable. If a new version performs worse than its predecessor, the organization needs the ability to roll back quickly. This requires:

- Unique version identifiers for every model iteration
- Immutable storage of model weights and configurations
- Documented rollback procedures and decision authority
- Clinician notification when a model version changes

THE AUDIT LENS: ASSESSMENT QUESTIONS

#	Assessment Question	Evidence to Request	Score
1	Is model performance continuously monitored?	Automated dashboards, KPI tracking	___/5
2	Are performance degradation triggers defined?	Thresholds for intervention (e.g., >5% drop)	___/5
3	Is model retraining performed when needed?	Scheduled or triggered retraining	___/5
4	Is retraining data quality controlled?	New data meets quality standards	___/5

#	Assessment Question	Evidence to Request	Score
5	Are retrained models validated before deployment?	A/B testing, validation on new data	___/5
6	Is version control maintained?	Model versioning, rollback capability	___/5
7	Is change management process followed?	Testing, approvals, phased rollout	___/5
8	Is FDA notification performed when required?	Significant updates to 510(k) devices	___/5
9	Are model updates communicated to users?	Clinician notification, training updates	___/5
10	Is model retirement process defined?	End-of-life, data retention, transition	___/5

KEY TAKEAWAYS: WHAT GOOD LOOKS LIKE

Every AI system has a formal maintenance plan with defined performance thresholds that trigger review or retraining. Retrained models go through full validation before deployment. All model versions are version-controlled with rollback capability. Clinicians are notified of significant model changes. Model retirement follows a formal process with data retention and transition planning. FDA notification is assessed for every significant change to regulated systems.

CHAPTER 10:
AI ETHICS & GOVERNANCE

BECAUSE DOING AI RIGHT IS NOT JUST TECHNICALLY HARD — IT IS MORALLY HARD

THE BIG IDEA

Every other chapter in this book has been, at its core, an ethics chapter. Validation is about whether patients can trust what AI tells us about them. Bias testing is about whether that trust is equally available to all patients. Safety monitoring is about whether we keep our promise to do no harm. But this chapter goes one level deeper — to the structural governance that makes ethical behavior consistent, accountable, and improvable over time.

Ethics without governance is aspiration. Governance without ethics is compliance theater. The goal is to build systems, processes, and cultures where doing the right thing with AI is also the path of least resistance.

THE FOUR PRINCIPLES OF BIOMEDICAL ETHICS — APPLIED TO AI

The foundational framework of healthcare ethics maps directly onto AI governance:

- Beneficence: AI systems must actively help patients, not just avoid harming them.
- Non-Maleficence: Do no harm — prevent errors, bias, and unsafe automation.
- Autonomy: Patients have the right to know when AI influences their care and to refuse it.
- Justice: AI benefits must be equitably distributed; no group should bear disproportionate AI risk.

THE AI ETHICS BOARD

An AI Ethics Board (sometimes called an AI Governance Committee or Clinical AI Review Board) is the organizational structure through which ethical oversight is exercised. At Skillweed, this board should include:

- Clinical leadership (CMO, CNO, department chairs)
- Data science and AI engineering representation
- Legal and compliance counsel
- Patient/community advocates
- Information security leadership
- Ethics faculty or external ethicists

High-risk AI systems should require formal Ethics Board review before deployment. The board should also be the escalation destination for ethical concerns raised by any staff member.

PATIENT TRANSPARENCY & INFORMED CONSENT

Patients increasingly have a right to know when AI influences their care. This takes several forms:

- General disclosure — A policy-level statement that the organization uses AI in clinical care
- System-specific disclosure — For high-risk decisions, informing patients that an AI recommendation was involved
- Opt-out provisions — Where clinically and technically feasible, allowing patients to request AI-free decision pathways

This is evolving territory. State laws vary. Patient expectations are rising. Building the infrastructure for transparency now will prevent reactive scrambling later.

ACCOUNTABILITY: THE MISSING PRINCIPLE

When an AI system contributes to patient harm, who is accountable? The vendor? The clinician who followed the recommendation? The hospital that deployed the system? The answer must be defined in advance — in contracts, in credentialing policies, in clinical bylaws. Accountability that is undefined in advance becomes a legal dispute after the fact.

BUILDING AN ETHICAL AI CULTURE

Governance structures matter — but culture matters more. An organization with robust policies and a board that never meets will underperform an organization with simpler structures and genuine leadership commitment. Hallmarks of an ethical AI culture:

- Leaders who openly discuss AI limitations, not just capabilities
- Psychological safety for clinicians to raise concerns about AI systems
- Reward structures that value quality and equity, not just efficiency

- Regular ethics education woven into AI training, not tacked on as a one-time module
- Transparent reporting of AI performance — including failures — to governance bodies

THE AUDIT LENS: ASSESSMENT QUESTIONS

#	Assessment Question	Evidence to Request	Score
1	Is AI Ethics Board established and active?	Multidisciplinary committee, regular meetings	___/5
2	Are all high-risk AI systems reviewed by Ethics Board?	Mandatory ethics review	___/5
3	Are ethical principles defined and applied?	Beneficence, non-maleficence, autonomy, justice	___/5
4	Is patient autonomy respected?	Informed consent, opt-out options	___/5
5	Are equity and justice considerations addressed?	Fair access, equitable performance	___/5
6	Is transparency provided to patients?	Disclosure of AI use in care	___/5
7	Are privacy protections in place?	PHI security, data minimization	___/5
8	Is accountability clearly assigned?	Responsible parties for AI decisions	___/5
9	Are ethical concerns raised and addressed?	Escalation process, issue resolution	___/5
10	Is ethics guidance updated regularly?	Alignment with evolving standards, regulations	___/5

KEY TAKEAWAYS: WHAT GOOD LOOKS LIKE

An active, multidisciplinary AI Ethics Board meets regularly and reviews all high-risk systems before deployment. Ethical principles are codified in policy and actively applied in governance decisions. Patients receive meaningful disclosure about AI use in their care. Accountability for AI-related harm is clearly defined in contracts and clinical policies. The ethics function evolves with regulations, public expectations, and organizational learning.

CONCLUSION: THE GOVERNANCE MINDSET

You have made it through ten domains of AI governance. If there is one idea to carry forward, it is this: AI governance in healthcare is not a compliance exercise. It is an expression of your organization's commitment to the patients you serve.

Every inventory entry is a patient whose care depends on that system being trustworthy. Every bias test is a patient who deserves an equitable shot at accurate diagnosis. Every override mechanism is a clinician who deserves to stay in control. Every ethics review is a human being whose dignity matters more than any algorithm's efficiency.

The 10 domains in this guide are not bureaucratic boxes to check. They are the structural expression of a healthcare culture that takes AI seriously — not as a magic solution, but as a powerful tool that requires the same professional stewardship as every other tool in medicine.

THE FOUR COMMITMENTS OF HEALTHCARE AI GOVERNANCE

	Commitment	What It Means in Practice
1	Know Your Systems	Complete, current inventory; clear risk classification; defined owners
2	Validate Rigorously	Independent testing; subgroup analysis; calibration; external validation
3	Monitor Continuously	Safety dashboards; drift detection; adverse event review; performance trending
4	Govern with Integrity	Ethics board; patient transparency; equity testing; accountable leadership

AI and robotics will continue to transform healthcare. The organizations that govern them well will build the trust needed to use them at scale. The organizations that do not will face the consequences — in patient outcomes, in regulatory action, and in the erosion of the trust that healthcare runs on. Skillweed Healthcare System has 53 AI systems, 7 high-risk, 18 FDA-regulated, and a team of people committed to getting this right. This book is for all of them.


SkillWeed