Implementing AI in Academic Medicine

A guide to governance and principled practice

Sean Davis, MD, PhD

2026-04-29

IMPLEMENTING AI IN ACADEMIC MEDICINE

A Practical Guide to Governance, Barriers, and Principled Practice

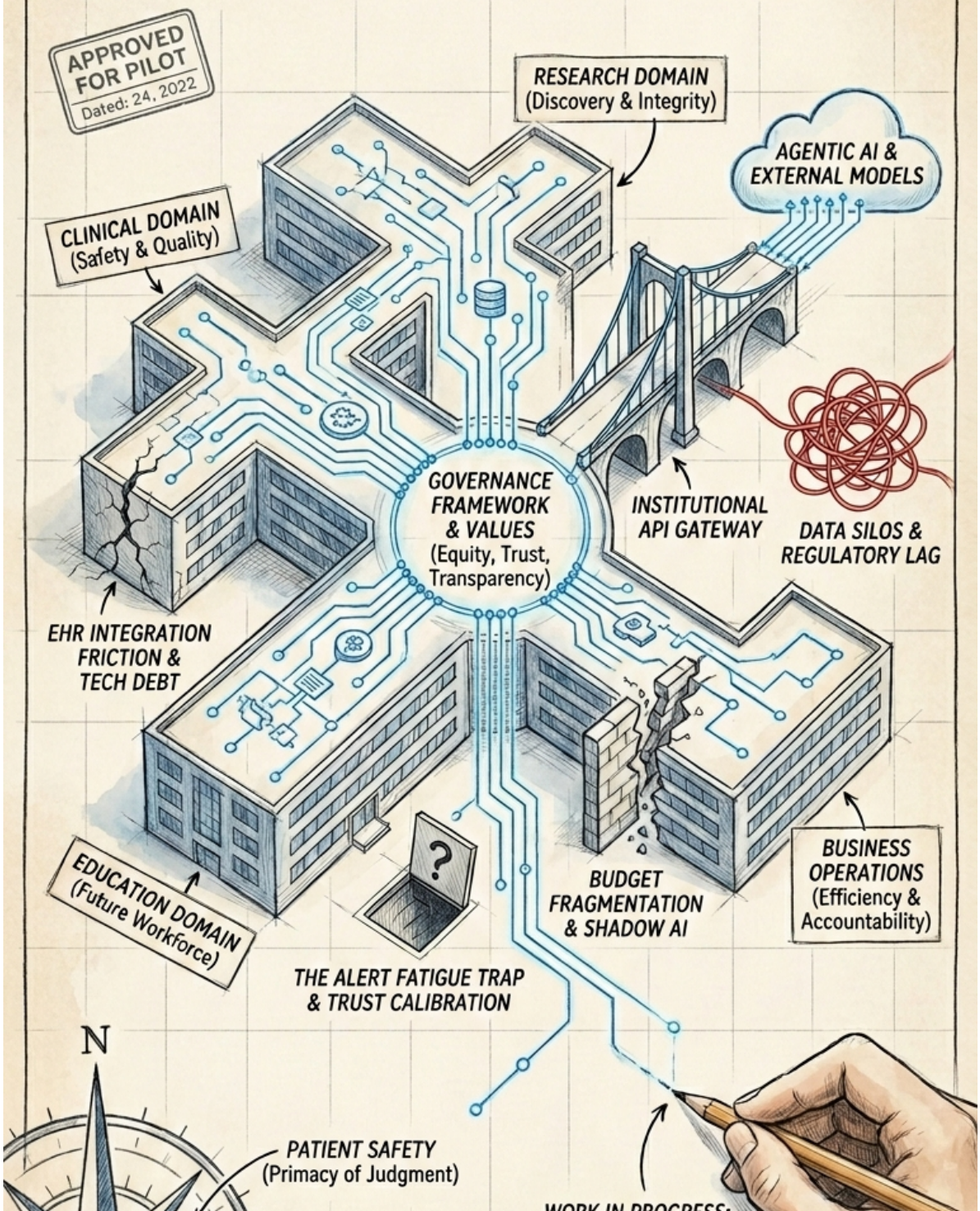


Table of contents

Preface 11

1 Overview 13
- 1.1 What changed between 2020 and 2024 . . . 13
- 1.2 The evidence: what is actually working . . . 14
- 1.3 The evidence: where it has gone wrong . . . 15
- 1.4 Where governance stands right now . . . 16
- 1.5 What this book is and isn't . . . 16

2 Values and Principles 18
- 2.1 Patient Safety and the Primacy of Clinical Judgment . . . 18
- 2.2 Equity as Performance Requirement . . . 19
- 2.3 Transparency: From Principle to Attribute List . . . 19
- 2.4 Human Oversight and the Agentic Threshold . . . 20
- 2.5 Accountability: The Structural Turn . . . 20
- 2.6 How Peer Institutions Have Operationalized These Principles . . . 21

I Implementation 23

3 The Framework 24
- 3.1 The Four Domains . . . 24
- 3.2 The Five Workstreams . . . 25
- 3.3 Why This Structure . . . 26
- 3.4 How to Use This Framework . . . 28

4 Domain Implementation Guide 29
- 4.1 The Domain Divergence . . . 29
- 4.2 Normalization, Not Installation . . . 30
- 4.3 The Clinical Domain: Safety as the Primary Constraint . . . 31
- 4.4 The Research Domain: Integrity Before Efficiency . . . 31
- 4.5 The Education Domain: Literacy as a Prerequisite . . . 33
- 4.6 The Business Operations Domain: Efficiency With Accountability . . . 34
- 4.7 The Champion Infrastructure . . . 34
- 4.8 Where to Start . . . 35
  - 4.8.1 Starter Project 1: Domain Implementation Roadmap . . . 35

4.8.2 Starter Project 2: Clinician Champion Cohort . . . 35

5 Barriers and How to Address Them **36**
5.1 The Governance Vacuum . . . 36
5.2 EHR Integration Friction . . . 37
5.3 The Alert Fatigue Trap . . . 37
5.4 The Trust Calibration Problem . . . 39
5.5 Budget Fragmentation . . . 40
5.6 The Barriers You Cannot Fix . . . 41
5.7 Where to Start . . . 42
5.7.1 Starter Project 1: Barrier Audit and Governance Gap Analysis . . . 42
5.7.2 Starter Project 2: Shadow AI Discovery and Triage . . . 42

II Domain Resources **44**

6 Clinical Domain **46**
6.1 The Guiding Principles Framework . . . 46
6.1.1 Principle 1: AI Tools Should Aim to Alleviate Existing Health Disparities 48
6.1.2 Principle 2: AI Tools Should Produce Clinically Meaningful Outcomes . 48
6.1.3 Principle 3: AI Tools Should Reduce Overdiagnosis and Overtreatment 49
6.1.4 Principle 4: AI Tools Should Have High Healthcare Value . . . 49
6.2 Principles 5–8: From Ethics to Operations . . . 49
6.2.1 Principle 5: Incorporating Biographical and Structural Drivers of Health 49
6.2.2 Principle 6: Local Calibration for the Local Population . . . 50
6.2.3 Principle 7: Promoting a Learning Healthcare System . . . 50
6.2.4 Principle 8: Facilitating Shared Decision-Making Through Explainability 51
6.3 Clinical AI in Practice . . . 52
6.3.1 Ambient Documentation and the Documentation Burden . . . 52
6.3.2 The Federal Regulatory Landscape . . . 53
6.3.3 Governing Adaptive AI: FDA SaMD and the PCCP Framework . . . 55
6.3.4 Explainability, Trust, and the Liability Frontier . . . 55
6.3.5 Patient-Facing AI and the Consent Gap . . . 57
6.4 Where to Start . . . 58
6.4.1 Starter Project 1: Clinical AI Inventory and Regulatory Risk Assessment 58
6.4.2 Starter Project 2: Ambient Documentation Pilot with Pre/Post Measurement . . . 59

7 Research Domain **60**
7.1 The Information Crisis in Biomedical Research . . . 60
7.2 Literature Discovery and Synthesis . . . 61
7.3 Hypothesis Generation and Study Design . . . 62
7.4 Code Generation and Reproducible Analysis . . . 63

7.5 Manuscript Drafting, Authorship, and the Policy Landscape . . . 63
7.6 Research Integrity Risks . . . 65
7.7 Human Subjects, Privacy, and Secure Infrastructure . . . 65
7.8 Where to Start: Two Starter Projects . . . 67
7.8.1 Project 1: Institutional Literature Review Toolkit . . . 67
7.8.2 Project 2: Secure AI Gateway for Research Computing . . . 67

8 Education Domain **69**
8.1 The Collapse of the Proxy . . . 69
8.2 The Detection Trap . . . 70
8.3 A Workable Framing: Tiered Policies and Process Grading . . . 71
8.4 The USMLE and the Limits of Licensing Exam Reform . . . 71
8.5 AI Literacy as Curriculum . . . 73
8.6 Institutional Policies: What Has Actually Been Published . . . 74
8.7 Academic Integrity as a Patient Safety Issue . . . 76
8.8 Where to Start: Two Starter Projects . . . 76
8.8.1 Project 1: AI Literacy Module for All Health Professions Students . . . 76
8.8.2 Project 2: Assessment Redesign Workshop for Course Directors . . . 77

9 Business Operations **78**
9.1 A Domain Unlike the Others . . . 78
9.2 What Business Operations AI Actually Does . . . 79
9.3 The Regulatory Layer . . . 82
9.4 The Shadow-IT Problem . . . 83
9.5 Enterprise AI Platforms . . . 83
9.6 Governance as Procurement . . . 85
9.7 What Success Looks Like . . . 86
9.8 Where to Start: Two Starter Projects . . . 88
9.8.1 Project 1: Enterprise AI Gateway Deployment . . . 88
9.8.2 Project 2: Internal Policy Q&A Chatbot (RAG over Internal Documents) 88

10 Regulatory and Policy Landscape **90**
10.1 The Federal Regulatory Baseline . . . 90
10.2 The Executive Pivot: Federal Deregulation and State Divergence . . . 91
10.3 The State Legislative Wave . . . 93
10.4 Professional Sovereignty and Accreditation Standards . . . 94
10.5 International Requirements for Global Research Partners . . . 95
10.6 Where to Start . . . 96
10.6.1 Starter Project 1: AI Regulatory Compliance Mapping . . . 96
10.6.2 Starter Project 2: Annual AI Governance Report to the Board . . . 96

III Agentic AI in Practice **97**

11 Agentic Safety and Guardrails **98**
11.1 From Advisory to Agentic: The Autonomy Spectrum . . . 98
11.2 Agentic Failure Modes . . . 99
11.3 Human-in-the-Loop Architecture . . . 100
11.4 Kill Switches and Circuit Breakers . . . 101
11.5 Least Privilege and Scope Limitation . . . 101
11.6 The Regulatory Picture for Agentic AI . . . 103
11.7 The Action Risk Authorization Matrix . . . 104
11.8 Where to Start . . . 105
11.8.1 Starter Project 1: Agentic AI Audit and Tiering . . . 105
11.8.2 Starter Project 2: Kill-Switch Design and Testing . . . 105

12 Patient and Community Trust **107**
12.1 The Empirical Trust Landscape . . . 107
12.2 The Historical Roots of Differential Trust . . . 108
12.3 Meaningful Disclosure vs. Boilerplate . . . 109
12.4 Informed Consent and Its Gaps . . . 111
12.5 Patient AI Advisory Councils . . . 111
12.6 The Regulatory Landscape of Disclosure . . . 112
12.7 Trust Recovery After Adverse AI Events . . . 113
12.8 Where to Start . . . 114
12.8.1 Starter Project 1: Patient AI Disclosure Audit and Language Standardization . . . 114
12.8.2 Starter Project 2: Establish a Patient AI Advisory Council . . . 114

13 Professional Wellness and Reducing Cognitive Burden **116**
13.1 The Structural Crisis of the Administrative Burden . . . 116
13.2 Pajama Time: The Mechanism of Burnout . . . 117
13.3 Ambient AI Documentation: The Evidence Base . . . 118
13.4 The Ambient Consent Architecture . . . 119
13.5 AI-Assisted Inbox Management . . . 120
13.6 Automation Complacency and the Vigilance Gap . . . 121
13.7 Nursing and Advanced Practice Burden . . . 122
13.8 Where to Start . . . 122
13.8.1 Starter Project 1: Ambient Documentation Pilot with Pre/Post Wellness Measurement . . . 123
13.8.2 Starter Project 2: Prior Authorization Workflow Automation Assessment 123

IV Workstream Resources **125**

14 IT Infrastructure and Security **126**
14.1 The Buy/Build/Connect Trilemma . . . 126
14.2 The Institutional API Gateway . . . 127
14.3 Clinical RAG Architecture . . . 127
14.4 The Security Stack . . . 130
14.5 Ambient AI and the EHR Integration Layer . . . 130
14.6 Agentic Infrastructure: Beyond the Advisory Model . . . 131
14.7 Sovereign Cloud and On-Premises Deployment . . . 132
14.8 Defeating Shadow AI . . . 133
14.9 Where to Start . . . 133
14.9.1 Starter Project 1: Institutional API Gateway Deployment . . . 133
14.9.2 Starter Project 2: Internal AI Assistant with Clinical RAG . . . 134

15 Training and Workforce Development **135**
15.1 The Four-Tier Competency Model . . . 135
15.2 National Competency Frameworks . . . 136
15.3 The Clinical Human-in-the-Loop Mandate . . . 137
15.4 Shadow AI as a Training Priority . . . 138
15.5 Measuring the Gap: The Evidence the Chapter's Claims Rest On . . . 138
15.6 The Accreditor Mandate: 2025 and Beyond . . . 140
15.7 Building a Living Curriculum . . . 141
15.8 The Faculty Development Gap . . . 141
15.9 Where to Start . . . 142
15.9.1 Starter Project 1: Role-Based AI Literacy Module Deployment . . . 142
15.9.2 Starter Project 2: AI Champions Program . . . 142

16 Ethics, Equity, and Institutional Accountability **144**
16.1 Algorithmic Bias as a Structural Problem . . . 144
16.2 Health Equity as a Performance Metric . . . 145
16.3 Informed Consent in the Continuous AI Era . . . 146
16.4 Intellectual Property and the AI Authorship Gap . . . 146
16.5 The Regulatory Turn: HHS Section 1557 and the Duty to Mitigate . . . 147
16.6 Beyond Obermeyer: Recent Cases of Algorithmic Bias . . . 147
16.7 State Privacy Laws and the Post-HIPAA Landscape . . . 149
16.8 The Workforce and Labor Dimension . . . 150
16.9 Community Trust and the Social License to Deploy . . . 150
16.10 Liability, the Standard of Care, and the Duty to Use . . . 151
16.11 Where to Start . . . 152
16.11.1 Starter Project 1: Equity Audit of Deployed Clinical AI . . . 152
16.11.2 Starter Project 2: Clinical AI Ethics and Accountability Policy . . . 152

17 Data Access and Governance **154**
17.1 The AMC Data Mosaic . . . 154
17.2 Data Classification for AI . . . 155
17.3 The Limits of De-identification . . . 155
17.4 The AI-Ready Honest Broker . . . 156
17.5 FHIR and OMOP as AI Substrate . . . 156
17.6 Vendor Contracts: The Non-Negotiables . . . 158
17.7 TEFCA and the Nationwide Exchange Layer . . . 159
17.8 The NIH Data Management and Sharing Policy Tension . . . 159
17.9 Synthetic Data as a Governance Instrument . . . 160
17.10 Federated Learning and the Governance of Distributed Data . . . 161
17.11 Where to Start . . . 162
17.11.1 Starter Project 1: AI Data Governance Policy and BAA Audit . . . 162
17.11.2 Starter Project 2: Institutional AI Data Enclave . . . 162

18 Project Management and AI Portfolio Governance **164**
18.1 The AISC as Portfolio Manager . . . 164
18.2 The Intake Engine: Triage Before Resource Commitment . . . 165
18.3 Stage-Gate Discipline: From Ideation to Scale . . . 167
18.4 The Integration Tax and Pilot Design . . . 168
18.5 The Total Product Lifecycle . . . 169
18.6 New Human Infrastructure: Architects and Champions . . . 170
18.7 Valuing AI: The Return on Health Framework . . . 171
18.8 Where to Start . . . 172
18.8.1 Starter Project 1: Intake Process and Stage-Gate Framework . . . 172
18.8.2 Starter Project 2: Deployed AI Portfolio Dashboard . . . 173

19 Evaluation and Monitoring **175**
19.1 Why Benchmark Performance Does Not Transfer . . . 175
19.2 Pre-Deployment Evaluation . . . 176
19.3 Post-Deployment Monitoring: Catching Drift . . . 177
19.4 KPI Architecture: Three Tiers . . . 178
19.5 Stakeholder Feedback as a Monitoring Method . . . 179
19.6 Domain-Specific Dimensions . . . 180
19.6.1 Clinical . . . 180
19.6.2 Research . . . 182
19.6.3 Education . . . 182
19.6.4 Business Operations . . . 183
19.7 Decommissioning . . . 184
19.8 Minimum Responsible Bar . . . 185

Appendices **186**

A AI Principles Across Governance Frameworks **186**
A.1 The Frameworks . . . 186
A.2 Where the Frameworks Agree . . . 187

B OSTP Blueprint for an AI Bill of Rights **190**
B.1 The Five Principles . . . 190
B.2 Current Status . . . 191
B.3 The Blueprint's Limitations . . . 191
B.4 Primary Sources . . . 192

C ONC HTI-1: Algorithm Transparency and Interoperability **193**
C.1 The Decision Support Intervention Framework . . . 193
C.2 What This Means for AMC Governance . . . 193
C.3 USCDI v3 and AI Data Access . . . 194
C.4 Information Blocking and AI Data Access . . . 194

D How This Book Was Written: A Multi-Model AI Authorship Workflow **196**
D.1 The Workflow in Brief . . . 196
D.2 Why These Tools in These Roles . . . 197
D.3 The Citation Problem . . . 198
D.4 Costs and Volume . . . 200
D.5 The Review Pass . . . 201
D.6 What This Means for AI-Assisted Knowledge Work . . . 202

References **204**

Preface

I started writing this framework in 2023, when the dominant institutional question at most academic medical centers was still "should we let people use ChatGPT?" That question has been overtaken by events. The question now is how to govern AI programs that are already real, already large, and already affecting patients, trainees, researchers, and staff in ways that most institutions do not yet fully see.

This book grew out of a working document — an attempt to organize the range of AI-related decisions an AMC has to make into something coherent enough to act on. It is not a strategic plan, and it is not meant to be prescriptive. Every institution has different constraints, different governance history, and different places where the technology has already landed. What I hope to offer is a framework flexible enough to be useful across those differences.

The central structural insight, and the one I keep returning to, is that an AMC is not a single AI deployment environment. It is four. Clinical care, research, education, and business operations each have different risk profiles, different regulatory obligations, different definitions of success, and different leadership structures. A governance program that treats them as a single domain will either be so restrictive that the research and education communities route around it, or so permissive that clinical AI gets deployed without the oversight patient safety requires. The domain structure in this framework is an attempt to hold that complexity without collapsing it.

Across all four domains, the same five operational questions recur. Who can access what data and under what conditions? What technical infrastructure and security controls are required? What are the ethical and regulatory obligations? How does the institution build and sustain the workforce capacity to use AI responsibly? And who manages the program across its full lifecycle, from the first intake request to eventual decommissioning? These are the workstreams — the horizontal dimension of the framework — and they are where most of the practical governance work actually lives.

The chapters that have grown up around this framework are substantially more detailed than the working document I started with, and they reflect a lot of thinking from peer institutions, published governance frameworks, and the accumulating evidence on what makes clinical AI deployments succeed and fail. I have tried to be honest about what is well-established and what is still contested, and to flag where the regulatory landscape is still moving.

This is a living document. The regulatory environment described in the governance chapter will continue to change. The AI tools described in the domain chapters will be superseded.

The governance programs at peer institutions will evolve. The version you are reading is as current as it was when it was last revised — the date at the top of each chapter is your guide to how much that should concern you.

It is also, itself, a product of AI collaboration. The research, drafting, and ongoing maintenance of this book are conducted with the assistance of large language models — primarily Claude and Gemini — working under human direction and review. The workflow appendix (Appendix D) describes that process in detail, including where the AI contributes most and where human judgment remains the necessary check.

This is a living document. Scope and content can and will continue to change and evolve. All the material is open for comment. Contributions are welcome via the github repo[1]. Reuse and sharing are encouraged.

[1] https://githbub.com/seandavi/amc-ai-governance

1 Overview

At most academic medical centers right now, AI is being deployed faster than the governance structures meant to oversee it. The 2025 CHIME/Censinet survey found that 84 percent of U.S. health systems have some form of AI steering committee — but only 10 percent maintain an automated inventory of which AI tools are actually running in clinical environments (College of Healthcare Information Management Executives and Censinet 2025). The committee exists. The operational visibility does not.

That gap is not a small administrative oversight. Ambient documentation tools are now used by more than 600,000 clinicians across U.S. health systems. Epic[1] has embedded AI across the clinical workflow at thousands of hospitals without requiring most of those hospitals to make a deliberate procurement decision. Research teams are running LLM-assisted literature synthesis, grant applications, and protocol drafts whether or not anyone has worked through the research integrity implications. Administrative staff are using consumer AI for tasks that touch sensitive data, often without knowing what the privacy exposure actually is. The question is no longer whether AI has been deployed at your AMC. The question is whether anyone knows what has been deployed, and whether the governance structures are in place to ensure it is being used well.

This book exists because the answer to that second question, at most institutions, is not yet yes.

> **Note**
>
> This book is fully compliant with the llms.txt standard[a]. Every page is available in plain Markdown format for use with AI assistants and large language models. See the Quarto llms.txt documentation[b] for details.
>
> [a] https://llmstxt.org/
> [b] https://quarto.org/docs/websites/website-llms.html

1.1 What changed between 2020 and 2024

Previous waves of AI in healthcare did not land like this. Expert systems in the 1980s and 1990s required custom development and remained isolated experiments. The early machine

[1] https://www.epic.com

learning wave of the 2010s produced promising models in research settings — genomic risk prediction, sepsis alert systems, radiology AI — but deployment at scale was slow, expensive, and required deep integration with clinical informatics teams. The tooling was hard. The data pipelines were fragile. Most models never left the institution that built them.

What changed was the API. Starting in 2023, Microsoft Azure OpenAI[2], AWS Bedrock[3], and Google Vertex AI[4] began offering foundation model access through enterprise APIs with signed Business Associate Agreements, U.S.-only data residency options, and zero-data-retention configurations for prompt content. The computational barrier to deploying a capable language model collapsed. An AMC with no machine learning infrastructure could connect a clinical workflow tool to a frontier model in weeks rather than years.

The second change was the nature of the tools. Previous clinical AI was episodic and discrete: this radiology image, this EHR record, this risk score calculated at a specific decision point. The new AI is ambient and continuous. An ambient documentation system is active in every patient encounter, listening to a conversation, and generating a clinical note that the physician then attests to as their professional documentation. A predictive readmission model runs on every patient in the hospital, updating continuously as new data arrives. A care gap identification algorithm touches every patient in the panel, every night. These tools do not generate outputs at discrete moments when a clinician is paying attention. They operate continuously, at scale, in the background.

That shift from episodic to ambient changes everything about governance. A governance model designed for discrete, intentional AI queries does not cover a system that is continuously analyzing every patient encounter without any individual triggering event.

1.2 The evidence: what is actually working

The ambient documentation case is the clearest current evidence of clinical AI value. Studies of early adopters at academic medical centers have found consistent reductions in documentation time — typically 10 to 15 minutes per patient encounter — alongside improvements in note quality and physician satisfaction (Tierney et al. 2024). The AMA's 2023 survey found that most physicians who reported using AI tools had positive perceptions of their impact on efficiency, though they were more skeptical about the tools' accuracy and their own ability to verify AI-generated content (American Medical Association 2023). The burnout implications are real: documentation burden is a primary driver of physician attrition, and attrition at AMCs is measured in millions of dollars per departing physician. Ambient documentation is not just a convenience tool. It is a workforce retention intervention.

[2] https://azure.microsoft.com/en-us/products/ai-services/openai-service
[3] https://aws.amazon.com/bedrock/
[4] https://cloud.google.com/vertex-ai

Diagnostic AI in radiology and pathology has accumulated the strongest clinical evidence base outside of documentation. FDA-cleared AI tools for diabetic retinopathy screening, mammography triage, pulmonary nodule detection, and stroke identification have demonstrated performance at or near specialist-level accuracy in prospective validation studies. The evidence for diagnostic AI reducing time-to-diagnosis and, in some cases, improving outcomes in underserved populations where specialist access is limited is compelling enough that the liability question has begun to cut both ways: institutions may face exposure not only for harms caused by deploying AI but, increasingly, for failing to deploy AI tools that have become part of the standard of care for specific diagnostic tasks.

1.3 The evidence: where it has gone wrong

The counterpart to this evidence base is a set of high-profile failures that share a common diagnosis. IBM Watson for Oncology was deployed at major cancer centers with confident marketing claims about its ability to recommend treatment plans. Physicians at several institutions found its recommendations unsafe, based on synthetic training cases rather than real patient records, and in direct conflict with clinical judgment. The product was eventually discontinued. The failure was not primarily algorithmic. It was a governance failure: inadequate validation against the populations and workflows where the tool was actually deployed, and institutional decisions made on the basis of vendor claims rather than independent evidence.

The Epic Sepsis Model story is more instructive because it involves a tool that was widely deployed across real clinical environments and subjected to rigorous external validation. When Wong and colleagues at the University of Michigan validated the model against their own patient population, they found that its area under the curve — 87 percent in the vendor's reported validation — dropped to 63 percent in their environment (Wong et al. 2021). More significantly, when they analyzed what the model was actually predicting, they found it was largely capturing patients who were already suspected of having sepsis and who had already had diagnostic cultures ordered. As a predictive tool that could trigger earlier intervention, it was performing close to chance. The model was doing something, just not what the deployment decision assumed it was doing.

This pattern — a tool that performs well on a vendor-provided validation set and underperforms in the real clinical environment — is not an Epic-specific failure. It is a predictable consequence of deploying models without independent local validation. The validation set reflects the population and workflow context where the model was developed. Your population and workflow are different. Sometimes the difference is small. Sometimes it is the difference between an 87 percent AUC and a 63 percent AUC.

The algorithmic bias literature documents a third failure mode that is less about performance in aggregate and more about who the performance failures fall on. Obermeyer and colleagues' demonstration that a commercial risk stratification algorithm systematically underestimated the health needs of Black patients — because it used healthcare cost as a proxy for health need,

encoding unequal access into the model's outputs — remains the clearest published example of how a technically functional AI tool can produce inequitable outcomes (Obermeyer et al. 2019). The algorithm was working as designed. The design encoded an injustice.

1.4 Where governance stands right now

The governance response to this evidence — both the successes and the failures — has been substantial in scope and uneven in implementation. A handful of academic medical centers have built genuinely operational AI governance programs. Duke Health published a framework for Algorithm-Based Clinical Decision Support oversight that treats deployed algorithms as clinical assets with full lifecycle management requirements: a clinical owner, a technical owner, a silent evaluation phase before any tool influences clinical decisions, and a registry that maintains visibility into every algorithm in the environment (Bedoya et al. 2022). UCSF developed a Trustworthy AI playbook grounded in six operating principles — Fair, Robust, Transparent, Responsible, Privacy, and Safe — with mandatory checkpoints at data validation, pilot deployment, and enterprise scale. Vanderbilt built a REDCap-based intake process that applies structured triage to every AI tool proposal before it consumes governance committee bandwidth.

These are working models. They are not yet the norm. The same CHIME/Censinet survey that found 84 percent of health systems with AI governance committees found that only 59 percent have a formal intake process for evaluating new AI tools, and only 10 percent have automated inventory of what is actually deployed (College of Healthcare Information Management Executives and Censinet 2025). The governance aspiration is widespread. The operational machinery is not.

At the same time, the regulatory environment has moved from guidance to enforcement. The ONC Health Data, Technology, and Interoperability rule took effect in 2025, requiring EHR vendors to surface 31 structured source attributes — training data provenance, demographic performance breakdowns, known limitations — for every certified AI-enabled clinical decision support tool in the workflow. The HHS Section 1557 nondiscrimination rule now holds covered entities liable for deploying patient care decision-support tools that produce discriminatory outcomes. Colorado's AI Act requires annual impact assessments for high-risk AI. These rules are described in detail in Chapter 10. For present purposes, the point is that the option of deploying AI and revisiting governance later is closing.

1.5 What this book is and isn't

This is a working framework, not a finished playbook. It was developed for a specific context — an academic medical center trying to organize the deployment of AI tools across four semi-independent organizational domains (clinical care, research, education, and business operations)

while maintaining coherent governance — and that context shapes every recommendation in it.

The framework is organized around the recognition that an AMC is not a single AI deployment environment. It is four. Clinical AI governance is shaped by patient safety obligations, FDA regulation, and EHR integration realities that have nothing to do with research AI governance. Research AI governance is shaped by IRB requirements, data sharing agreements, and publication integrity standards that are irrelevant to the education domain. Each domain has its own risk profile, its own leadership structure, its own budget authority, and its own pace of adoption. A governance program designed for clinical AI and applied wholesale to educational uses will miss things that matter. The converse is equally true.

Across all four domains, the same five operational questions recur: who can access what data under what conditions, how is the technical infrastructure governed and secured, what are the ethical and legal obligations, how does the workforce develop the competency to use AI responsibly, and who manages the AI program across its full lifecycle from intake to decommission. These five questions are the workstreams that cross-cut the domain structure, and they are the organizational scaffolding of this book.

The chapters that follow are written to be useful independently as well as together. A CMIO who needs to stand up a clinical AI governance program can read the clinical chapter, the infrastructure chapter, and the project management chapter without reading everything in between. A CHRO who needs to build a workforce AI literacy program can read the workforce chapter without needing the data governance chapter. The cross-references are there for context, not for prerequisite reading.

I wrote the first version of this framework in 2023, when the dominant institutional question was still "should we let people use ChatGPT?" That question has been overtaken by events. The question now is how to govern AI programs that are already real, already large, and already affecting patients, researchers, students, and staff in ways that most institutions do not yet have full visibility into. This version of the book tries to be useful to that question. Whether it succeeds is something you will be better positioned to judge than I am.

2 Values and Principles

Every AMC that has published an AI governance framework lists roughly the same values: patient safety, equity, transparency, accountability, privacy, and human oversight. The lists are not wrong. They are also not differentiating. The meaningful question is not which values an institution holds — that convergence is real and appropriate — but whether those values are encoded in governance structures that make them operational, or whether they remain aspirational statements that authorize AI deployments without constraining them.

This distinction matters because value statements are not self-enforcing. An institution can publish a principle of "transparency" and simultaneously deploy clinical AI tools with no documentation of training data provenance, no disclosure to clinicians of the model's known limitations, and no audit trail connecting AI outputs to clinical decisions. The principle is present. The operational mechanism that would give it meaning is not. Three-quarters of U.S. health systems now report deploying at least one AI application (Fierce Healthcare 2026). At that scale, the gap between a values statement and an operational mechanism is not a planning problem. It is already consequential. The chapters that follow this one are primarily about the mechanisms. This chapter is about why the principles matter and how peer institutions have translated them into governance practice.

2.1 Patient Safety and the Primacy of Clinical Judgment

Patient safety is the foundational principle in clinical AI governance, and it is the one most at risk of being treated as self-evident when it requires active structural support. The automation complacency literature documents consistently that clinicians who work with accurate AI tools gradually reduce their independent scrutiny of AI outputs — not through carelessness but through rational adaptation to an environment where the tool is usually right (Parasuraman and Manzey 2010). The risk is not that clinicians distrust AI. It is that they come to trust it in ways that bypass the critical evaluation the technology requires.

Patient safety as an operational governance principle means two specific things. First, every deployed clinical AI tool requires an explicit human oversight architecture: who reviews AI outputs, under what conditions, with what documentation requirements, and what happens when the tool is wrong. This is not a policy preference. It is the design constraint that determines whether a deployment is safe. Second, training programs for clinicians using AI tools should include explicit instruction on the specific failure modes of each tool they use, not

just its aggregate accuracy. A physician who knows that ambient documentation tools are more likely to omit negative findings than to hallucinate positive ones is a better human-in-the-loop than one who knows only that the tool achieves high average accuracy.

2.2 Equity as Performance Requirement

Equity is the principle that most AMC AI governance programs acknowledge and fewest operationalize. The acknowledgment is easy: any institutional AI values statement will include language about not exacerbating health disparities. The operationalization requires something specific: demographic stratification of performance metrics as a standard component of AI validation, not an optional audit.

The Obermeyer 2019 demonstration that a widely deployed risk stratification algorithm systematically underestimated the health needs of Black patients was not a finding about a biased algorithm (Obermeyer et al. 2019). It was a finding about a technically correct algorithm that encoded an existing inequity by using healthcare utilization as a proxy for health need. The algorithm did exactly what it was designed to do. The design was the problem. Catching this kind of structural inequity before deployment requires stratified performance analysis across race, ethnicity, age, insurance status, and language as a baseline requirement — not a step that happens if someone raises a concern.

Badal and colleagues proposed a framework that frames alleviation of health disparities as the first principle of responsible clinical AI, not a supplementary consideration (Badal et al. 2023). In practice, this means that a model achieving 85 percent accuracy overall but 70 percent accuracy for the subpopulation with the highest disease burden is not a high-performing model. It performs best for the patients who need it least. The equity audit process described in Section 16.11 is the operational mechanism for this principle. Without that audit, equity is a value statement, not a program element.

2.3 Transparency: From Principle to Attribute List

Transparency is now a regulatory requirement, not just an ethical aspiration. The ONC Health Data, Technology, and Interoperability rule requires EHR vendors to surface 31 structured source attributes for every certified AI-enabled clinical decision support intervention (Office of the National Coordinator for Health Information Technology 2024). Those attributes include training data sources and date ranges, performance characteristics on the populations validated, known limitations and failure modes, instructions for appropriate use, and update history. This is what operationalized transparency looks like: a specific list of information that must be available at the point of clinical use.

For AMCs, this regulatory baseline is a floor, not a ceiling. Mitchell and colleagues' model card framework — now widely adopted by major AI vendors and required by the Coalition for Health AI[1] — defines a similar set of disclosure requirements that apply to any model, regardless of whether it falls under ONC certification (Mitchell et al. 2019). An AMC that requires model cards from every vendor providing an AI tool, and publishes equivalent documentation for internally developed tools, has operationalized transparency in a way that the values statement alone never could.

The transparency principle extends to patients. The WHO[2] ethics guidance on AI for health recommends meaningful disclosure to patients about which AI tools are used in their care and how those tools affect clinical decisions (World Health Organization 2024). California AB 3030 now requires disclosure on AI-generated patient communications. The operational mechanism here is the consent architecture described in Chapter 16: an institutional policy on when and how patients are informed about clinical AI use, not a case-by-case determination left to individual clinicians.

2.4 Human Oversight and the Agentic Threshold

Human oversight has become more complicated as AI has become more agentic. When the tool is advisory — an AI recommends, a clinician decides — the oversight model is relatively straightforward. As AI tools gain the ability to take autonomous actions — ordering tests, routing referrals, updating records, communicating with patients — the oversight architecture requires more specific design. Who reviews autonomous AI actions? How quickly must review occur to be meaningful? What actions require pre-authorization rather than post-hoc review? These questions are addressed specifically in Chapter 11; the relevant principle here is that human oversight is not a passive condition but an active design requirement that must be specified for each tool's risk profile and autonomy level.

2.5 Accountability: The Structural Turn

Accountability in AI governance is structural, not individual. An ambient documentation system that produces inaccurate notes fails not because the clinician was careless in attestation — though the attestation does carry professional accountability — but because the deployment decision did not include adequate validation of the tool in the clinical context, the training program did not adequately prepare clinicians to recognize the tool's specific failure modes, and the monitoring system did not detect the pattern of errors before it became a patient safety event. Individual accountability is real. It is also insufficient as an accountability framework for tools that operate at scale across thousands of encounters.

[1] https://www.coalitionforhealthai.org
[2] https://www.who.int

The structural accountability mechanisms in this book are: the stage-gate governance process in Chapter 18 that requires documented validation before tools influence clinical decisions; the equity audit requirement in Chapter 16 that creates an institutional record of what was known about demographic performance before deployment; the total product lifecycle monitoring in Chapter 18 that creates ongoing visibility into whether deployed tools are performing as validated; and the board-level reporting requirement in Section 10.6 that makes AI governance visible at the level of institutional accountability where it ultimately resides.

2.6 How Peer Institutions Have Operationalized These Principles

The convergence of leading AMC AI governance programs on similar structures is itself informative. Duke Health's Algorithm-Based Clinical Decision Support framework treats every deployed algorithm as a clinical asset with a named clinical owner, a named technical owner, a mandatory silent evaluation phase before any tool influences clinical decisions, and a registry that maintains continuous visibility into the deployed portfolio (Bedoya et al. 2022). The registry — a simple idea that turns out to be operationally significant — is what makes the 10 percent automated inventory figure so striking. Most institutions that have AI governance committees do not have the basic visibility into their own environment that accountability requires.

UCSF's Trustworthy AI playbook grounds its six principles in mandatory checkpoints at three points in the deployment lifecycle: before a tool accesses data, before a tool enters a pilot, and before a tool reaches enterprise scale. The checkpoint structure converts principles into decision gates. A tool does not advance through the lifecycle by meeting soft criteria; it advances by clearing documented requirements that are enforced by the governance committee.

Vanderbilt's AVAIL program requires every AI tool to pass through a REDCap-based intake process before it reaches the governance committee's attention. The intake triage is itself a governance mechanism: it ensures that every tool is assessed against a consistent set of criteria before institutional resources are committed to evaluation, and it builds the documented record that the annual impact assessment requirements of Colorado SB 24-205 and the equity audit requirements of HHS Section 1557 ultimately require.

Table 2.1: Governance principles mapped to regulatory requirements and operational mechanisms. The mechanism column is the bridge between aspiration and program.

Principle	Regulatory Mechanism	Operational Mechanism	Chapter
Safety	CMS MA rule: AI cannot substitute for human clinical review	Human oversight architecture, specific failure mode training	Chapter 6, Chapter 13

Principle	Regulatory Mechanism	Operational Mechanism	Chapter
Equity	HHS Section 1557: no discriminatory patient care tools	Stratified performance validation, demographic drift monitoring	Chapter 16
Transparency	ONC HTI-1: 31 source attributes for certified DSIs	Model card requirement at intake, institutional disclosure policy	Chapter 17, Chapter 18
Human oversight	FDA PCCP: human oversight for adaptive AI devices	HITL checkpoints, override documentation	Chapter 11
Accountability	Colorado SB 24-205: annual impact assessments	Stage-gate governance, board reporting	Chapter 18, Chapter 10
Privacy	HIPAA BAA + zero-data-retention provisions	Data classification framework, honest broker function	Chapter 17

These examples share a common architecture: principles that are specific enough to be falsifiable, checkpoints that create documented records of compliance, and ownership structures that ensure someone is accountable for the principle being honored. That architecture is what this book is about. The values are the starting point. The mechanisms are the work.

Part I

Implementation

3 The Framework

In order to successfully integrate AI across an academic medical center, it helps to start by acknowledging what an AMC actually is. It is not a single organization with a unified mission, a single budget, and consistent risk tolerance. It is a federation of at least four semi-independent domains — clinical care, research, education, and business operations — that share governance overhead and physical infrastructure but operate under largely independent leadership, funding models, regulatory obligations, and definitions of what success looks like. A tool or policy that makes perfect sense for clinical operations may be irrelevant or actively counterproductive in the research context. An AI use case that is urgent in the business operations domain may be low priority in the education domain, and vice versa.

This structural reality is the organizing principle of the framework. Rather than specifying a single institutional AI program, the framework recognizes that AI deployment in an AMC has both domain-specific and cross-domain dimensions, and that good governance requires handling each dimension appropriately. The domains are organized vertically, reflecting their semi-independence. The workstreams are organized horizontally, reflecting the shared capabilities and governance functions that cut across all domains. The AI Steering Committee sits above both, providing the portfolio management and accountability function that neither the domains nor the workstreams can provide alone.

3.1 The Four Domains

The domain structure reflects a basic fact about how AMCs work: clinical, research, educational, and business operations communities have different relationships to AI risk, different institutional cultures, and different processes for making deployment decisions.

Clinical encompasses patient care — the AI tools used in diagnosis, treatment planning, documentation, patient communication, and care coordination. Clinical AI carries the highest patient safety risk and the most developed regulatory framework. Tools in this domain may be subject to FDA oversight as Software as a Medical Device, must comply with the CMS human-review requirement for coverage decisions, and are now subject to the HHS Section 1557 nondiscrimination mandate. Clinical AI deployment decisions involve the CMO, the CMIO, patient safety leadership, and clinical governance structures that have no parallel in the other domains.

Research encompasses basic, translational, and clinical research programs — the AI tools used in literature synthesis, grant writing, protocol development, data analysis, and research operations. Research AI has a different risk profile: the primary risks are data provenance, IRB compliance, publication integrity, and the integrity of the scientific record, rather than immediate patient harm. Research AI deployment is governed by IRB review, data sharing agreements, federal research compliance frameworks, and professional norms around authorship and reproducibility.

Education encompasses teaching, learning, and assessment across medical school, graduate programs, residency, and continuing education. AI in education introduces academic integrity questions that have no parallel in the clinical or research domains — how to assess student competence in an environment where AI can generate sophisticated responses to most standard assessments, how to design learning experiences that build genuine clinical reasoning rather than delegating it to AI tools, and how to ensure that faculty can teach and evaluate AI-related competencies they are still developing themselves.

Business operations encompasses administrative and operational functions: revenue cycle, supply chain, facilities management, human resources, scheduling, and financial operations. Business operations AI typically has the lowest patient safety risk and the most tractable ROI calculation, but it touches sensitive financial and employment data and introduces process automation that changes staff roles in ways that require careful change management.

3.2 The Five Workstreams

Within each domain, the same five operational questions recur regardless of the specific AI use case. These recurring questions are the workstreams — the cross-cutting capabilities that the institution needs to develop and maintain whether it is deploying clinical AI, research AI, or administrative AI.

Data Access and Governance asks: what data can be used for AI development and deployment, under what conditions, governed by which agreements, and with what privacy protections? The data governance workstream is described in detail in Chapter 17. Its answers determine which AI use cases are feasible and which are not.

IT, Security, and Infrastructure asks: what technical architecture, API management, security controls, and monitoring infrastructure are required to deploy AI responsibly at scale? The infrastructure workstream, described in Chapter 14, is the difference between AI tools that are governed and AI tools that are not — the institutional API gateway is the chokepoint through which governance is enforced.

Ethical, Legal, and Social asks: what are the ethical obligations, regulatory requirements, and social implications of deploying this AI tool in this context? The ethics workstream, described in Chapter 16, is not the same question in all four domains. Research AI raises different equity questions than clinical AI; educational AI raises different accountability questions

than business operations AI. The workstream provides shared frameworks and governance structures while leaving domain-specific application to the domain teams.

Training and Workforce Development asks: what do the people using, deploying, and overseeing AI tools need to know, and how does the institution build and sustain that capacity? The workforce workstream, described in Chapter 15, includes the four-tier competency model (consumers, translators, developers, governors) and the faculty development programs that prevent AI literacy from degrading as the technology evolves.

Project Management and Support asks: how does the institution manage the AI portfolio across its full lifecycle — from intake through deployment through decommissioning — and how does it maintain the organizational infrastructure (champions, architects, governance committees) that makes deployment sustainable? The project management workstream, described in Chapter 18, is the connective tissue that keeps everything else from being a set of independent good intentions.

3.3 Why This Structure

The value of the matrix structure is not that it is theoretically elegant. It is that it matches the actual decision-making geography of an AMC.

Domain-level decisions — whether to deploy a specific AI tool in the clinical workflow, which research AI capabilities to invest in, how to address academic integrity in medical education — belong to the domain leadership. The CMO is accountable for clinical AI governance. The VP of Research is accountable for research AI governance. The dean's office is accountable for educational AI governance. These are not decisions that should be made by a central IT committee or a governance function disconnected from operational reality. Domain ownership is what makes AI governance legible to the people who actually work in those domains.

Workstream-level decisions — the data governance framework, the institutional API architecture, the workforce development curriculum, the portfolio management process — belong to centralized institutional functions because the alternatives are worse. An institution where each domain maintains its own data governance policy, its own API infrastructure, and its own workforce development program will produce four incompatible frameworks, four redundant technical stacks, and four sets of training content that conflict with each other. The workstreams are shared services not because central control is desirable in principle but because fragmentation in these specific areas is costly in practice.

The AISC is what makes the matrix function as a governance program rather than a conceptual diagram. Without the AISC — a body with actual authority over AI deployment decisions, portfolio visibility, and the power to terminate projects that fail governance review — the domains will optimize for their own priorities at the expense of institutional coherence, and the workstreams will provide advice that no one is required to follow. The AISC's role in portfolio management is described in detail in Chapter 18. The point here is structural: the domains

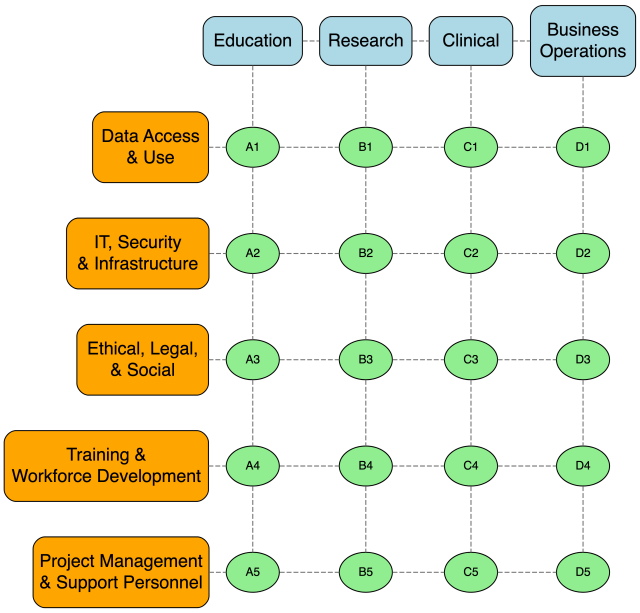


Figure 3.1: A schematic framework for organizing workstreams (orange boxes), domains (blue boxes), and work products and tasks (green ovals). Domains (vertical dimension) capture semi-independent organizations, each with largely independent use cases, budgets and business plans, priorities, and leadership. The workstreams (horizontal dimension) will often require similar or overlapping expertise, and can serve as knowledge resources to provide synergy and uniformity in implementation across domains.

and workstreams need the AISC the way a matrix organization needs its executive leadership. Remove it and you have a framework diagram without a governance program.

3.4 How to Use This Framework

This framework is a starting point, not a finished organizational design. Most AMCs will need to adapt it to their existing governance structures, their current AI deployment footprint, and their specific regulatory environment. An institution that already has a strong clinical informatics governance program may find that the clinical domain workstream structure maps naturally onto existing committees and roles. An institution that is starting from scratch may find the workstream structure useful primarily as a checklist of the functions it needs to build.

The chapters that follow describe each workstream in detail, grounded in current evidence and the practices of peer institutions that have built working programs. The domain chapters describe the AI use cases, governance considerations, and regulatory requirements specific to each domain. Neither set of chapters is meant to be read cover to cover before acting. The right starting point depends on where your institution is and what problem is most pressing.

If there is one organizing principle I would ask you to take from this framework, it is this: the question of what AI to deploy is secondary to the question of how to govern what you deploy. The deployment pipeline will fill itself. The governance program will not build itself. The work described in this book is the governance program.

4 Domain Implementation Guide

When we talk about implementing a new technology in an academic medical center, we default to the language of the IT rollout: go-live dates, user provisioning, help desk tickets. We treat the software as a finished object that needs to be deposited into the workflow, like furniture moved into an office. This mental model works reasonably well for a word processor, where the utility is clear and the risks of failure are largely individual. It fails completely for large language models. In two decades of building data infrastructure for these institutions, I have found that the technology is almost never the bottleneck. The bottleneck is the social and professional fabric of the hospital itself.

Implementing a language model is not a technological event; it is a sociotechnical negotiation. We are asking highly trained professionals to cede some portion of their cognitive labor to a system that is probabilistic, occasionally wrong, and often opaque about why it is wrong. If you come from a technical background, this looks like an optimization problem. If you come from the clinical side, it looks like a liability problem. Bridging that gap requires moving beyond generic change management checklists and toward a domain-specific understanding of how work actually happens in an AMC.

4.1 The Domain Divergence

There is no generic AI implementation in an academic medical center. The needs of a surgeon managing a complex case have almost nothing in common with the needs of a research coordinator managing a multi-center trial, or a dean of students worried about the integrity of a medical school exam. Each domain operates under different constraints, follows a different regulatory cadence, and answers to different stakeholders. When we force a one-size-fits-all implementation strategy on these groups, we end up with tools that are technically functional but socially rejected.

In the clinical domain, the primary constraint is patient safety and clinician liability. Every implementation step is filtered through those questions. The research domain cares most about integrity and data provenance — a researcher may accept a slower tool if it guarantees auditability and reproducibility. In education, the implementation must focus on literacy and assessment validity; a tool that inadvertently replaces student reasoning with AI reasoning defeats the educational purpose. Business operations is driven by efficiency and the complex

choreography of payer-provider relationships, with success measured in clicks saved and denial rates reduced.

Table 4.1: Domain implementation characteristics across the AMC. Each column represents a distinct deployment context with different risk profiles and accountability structures.

Characteristic	Clinical	Research	Education	Business Ops
Primary value	Patient safety	Scientific integrity	Academic growth	Operational efficiency
Primary risk	Diagnostic error / liability	Data provenance / IRB	Competency loss	Financial leakage
Key stakeholder	CMO / CMIO	IRB / VP Research	Dean	CFO
Regulatory lead	FDA, Joint Commission	OHRP, NIH	LCME, accreditors	CMS, payers
Success metric	Outcomes, time savings	Publication quality	Student performance	Margin, throughput

Acknowledging these differences is the prerequisite for a successful deployment. I have watched well-funded projects die because the technical team argued for efficiency to a department chair who was worried about liability. You have to speak the language of the domain you are entering. If you are in the clinic, quantify the integration tax and what the tool will actually save. If you are in the research office, explain the audit trail.

4.2 Normalization, Not Installation

To understand why some implementations stick and others disappear after the pilot phase, it helps to move beyond "adoption" and toward "embedding." Normalization Process Theory asks how a new practice becomes a normal part of daily work — not just something people are required to use, but something they use without thinking about it, the way they use their email.

The theory describes four constructs that matter for this process. Coherence — does the clinician understand what the model is doing, or does it feel like a black box? If they cannot form a mental model of when the system is likely to fail, they will not trust it appropriately. Cognitive participation — who decides to engage with the change, and who has the social capital to bring their colleagues along? This is not accomplished by memo; it requires finding the influential people in a department and getting them genuinely invested. Collective action — the actual day-to-day work of using the tool, including the hidden labor (double-checking outputs, mapping data between systems) that project plans routinely underestimate. And reflexive monitoring — the ongoing organizational process of checking whether the tool is

doing what it was supposed to do, whether its outputs still reflect current clinical practice, and whether its errors are accumulating in ways that require attention (Finlayson et al. 2021).

The value of this framework is that it focuses implementation planning on the social work rather than the technical work. The technical configuration of an API gateway takes weeks. Building genuine cognitive participation among skeptical clinicians takes months. Planning for the slower process is the difference between a successful deployment and a successful pilot that no one uses six months later.

4.3 The Clinical Domain: Safety as the Primary Constraint

In the clinical world, implementation must begin with a period of silent or shadow deployment. The tool runs in the background, consuming real patient data and generating outputs that are logged and reviewed but never shown to clinicians. This period validates the model against the clinical environment's actual patient population and workflow before any outputs influence clinical decisions.

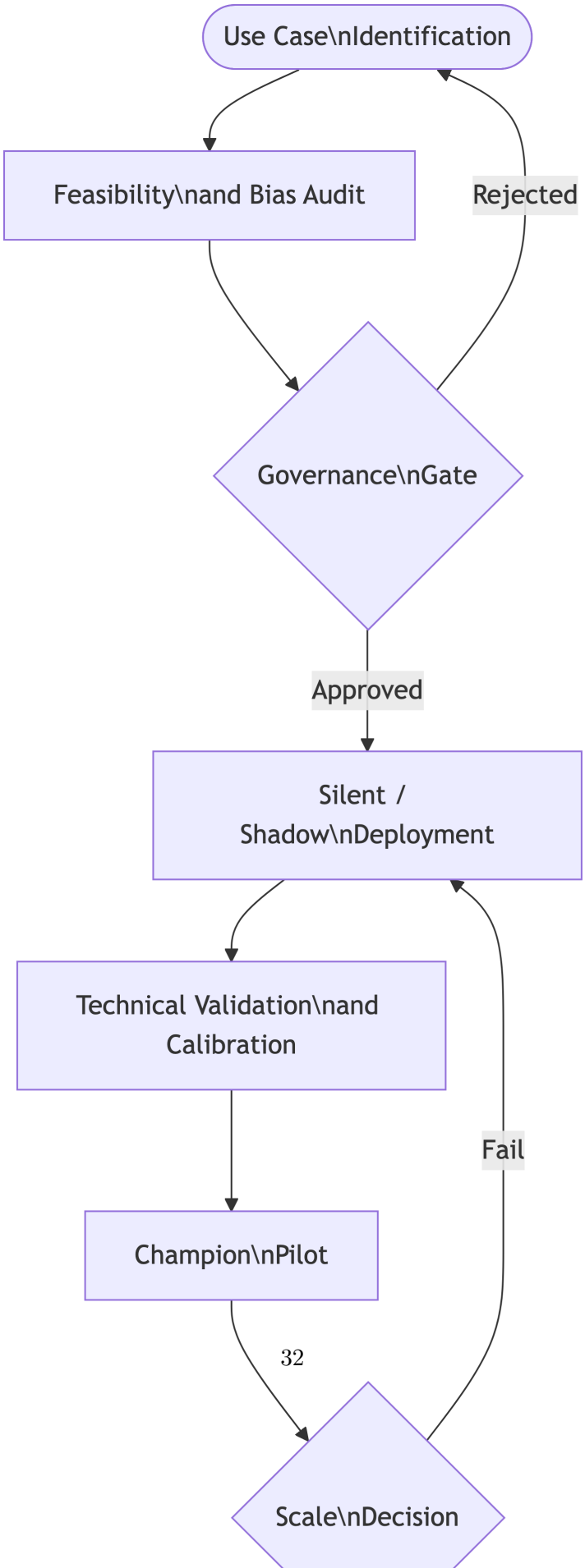
The DECIDE-AI[1] reporting guidelines for early-stage AI pilots give this process a formal structure: pre-registered primary endpoints, prospective design, and monitoring for unexpected harms (Vasey et al. 2022). The framework shifts the evaluation focus away from aggregate accuracy metrics — "area under the curve" — toward how the model actually changes clinician behavior when it is present. A model that is 95 percent accurate but whose alerts are ignored 90 percent of the time has not been implemented; it has been deployed and ignored (Wong et al. 2021).

The Duke Health Sepsis Watch deployment remains the most thoroughly documented example of this staged approach (Sendak et al. 2020). The team did not simply deploy a model — they spent months in shadow mode, involved nurses in interface design, and built a dedicated rapid-response workflow around the model's outputs. The technical performance was validated, but the implementation was designed around the social and operational reality of the units where it was deployed. That is the difference between a peer-reviewed model and a working clinical tool.

4.4 The Research Domain: Integrity Before Efficiency

Implementation in the research domain is governed by the IRB and by the demands of data provenance. A researcher may be willing to accept a slower, more cumbersome tool if it guarantees that the data remains auditable and the results are reproducible. For someone coming from the clinical side, the data governance overhead of research AI can feel like

[1] https://www.decide-ai.org/


Use Case\nIdentification
Feasibility\nand Bias Audit
Rejected
Governance\nGate
Approved
Silent /
Shadow\nDeployment
Technical Validation\nand
Calibration
Fail
Champion\nPilot
Scale\nDecision

unnecessary friction. In the research world, a single instance of unattributed AI-generated content can compromise a career and trigger institutional sanctions.

The implementation sequence here begins with data provenance mapping: where is the data coming from, where is it stored, who has access to the model outputs, and what is the audit trail connecting AI outputs to final research products. Model cards — the structured disclosure format defining training data, performance characteristics, and known limitations — are increasingly required as part of IRB submissions for AI-assisted research, and requiring them is one concrete way the institution operationalizes the transparency principle from Chapter 2.

The research domain also requires attention to the human authorship question. The ICMJE[2] standards state clearly that AI systems cannot be listed as authors, and that authors are responsible for the integrity of AI-assisted content, including any errors or fabrications introduced by AI tools. The implementation guidance here is straightforward but requires explicit communication: researchers using AI for analysis, writing, or literature synthesis are responsible for verifying AI-generated content to the same standard as any other source.

4.5 The Education Domain: Literacy as a Prerequisite

The education domain presents an implementation challenge that has no parallel in the others: if the faculty responsible for teaching and assessing students do not understand how AI tools work and where they fail, they cannot design valid assessments, evaluate AI-assisted student work, or teach the AI literacy that accreditation bodies are beginning to require.

Implementation in this domain must therefore begin with faculty development, not student-facing tools. Faculty development for AI does not require every professor to become a machine learning specialist. It requires that they understand the specific capabilities and failure modes of the tools their students have access to, and that they can redesign assessments to test for reasoning processes that AI cannot easily replicate. An assessment that can be completed by copy-pasting a prompt into a language model is not testing what it claims to test, regardless of whether the student was authorized to use AI.

The workforce chapter (Chapter 15) addresses the faculty development gap in detail. For implementation planning, the relevant point is sequencing: student-facing AI tools are appropriate to deploy after, not before, faculty have the literacy to evaluate their outputs and design around their capabilities.

[2] https://www.icmje.org

4.6 The Business Operations Domain: Efficiency With Accountability

Business operations is often the most ready for AI deployment and the most susceptible to the assumption that efficiency benefits are self-justifying. Revenue cycle management, scheduling optimization, and administrative documentation are legitimate and high-value AI use cases. They are also use cases where error consequences — incorrect billing codes, compliance violations, employment decisions driven by flawed algorithmic screening — can be significant and are sometimes invisible until they accumulate.

Implementation in the business operations domain requires the same governance structure as clinical deployment: a named owner, documented validation, and monitoring for outcomes that extend beyond throughput metrics. An automated prior authorization tool that systematically denies certain patient populations at higher rates is not an efficiency tool. It is a Section 1557 compliance problem. An AI-assisted hiring screening tool is subject to NYC Local Law 144[3]'s independent bias audit requirement if it is used for employment decisions in New York. Accountability structures in business operations AI are less visible than in clinical AI, but they are no less real.

4.7 The Champion Infrastructure

Every implementation, regardless of domain, depends on champions — people with the social capital and genuine engagement to lead their peers through the discomfort of adopting a new practice. The most effective champions are often not the most technically enthusiastic people in a department. The most effective champions are people who were initially skeptical and changed their minds based on evidence. Their skepticism makes them credible. When a skeptic says the tool saved them time, their colleagues listen in a way they would not listen to an early enthusiast.

Building a champion infrastructure means two things beyond identifying willing volunteers. First, protected time: champions who are expected to lead AI adoption in addition to a full clinical or research schedule will burn out or quietly deprioritize the role. One to two hours per week of protected time for the champion function is not generous; it is the minimum investment required for the role to be sustainable. Second, a community of practice that connects champions across domains and service lines, giving them a venue to share what is working, flag emerging problems, and develop the translator skills that allow them to bridge between clinical reality and informatics infrastructure.

The champion program starter project in Section 15.9 describes the specific structure; the point here is that champion capacity is not a soft governance element. It is the primary mechanism through which governance reaches the point of care.

[3] https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page

4.8 Where to Start

4.8.1 Starter Project 1: Domain Implementation Roadmap

Select one specific use case — a clinical documentation tool, a research literature synthesis tool, or a scheduling optimization tool — and walk it through the staged lifecycle in Figure 4.1. Do not try to solve the whole institution at once. Pick a single department with a motivated leader, run the shadow deployment phase, document the integration tax, and complete a champion pilot with pre-registered success metrics. The roadmap you produce from this exercise, including what broke and what was harder than expected, becomes the institution's deployment playbook for the next tool.

Why now: The first implementation is always the most instructive. Every subsequent deployment benefits from having lived through the gap between a governance framework on paper and a governance framework tested against a real vendor, a real EHR integration, and real clinicians who have other things to do.

Buy vs. build: Process design and documentation work. The shadow deployment requires access to production data under governance approval; the infrastructure for logging model outputs against clinical outcomes may require a modest analytics build, but the governance process itself is documentation and meeting time.

4.8.2 Starter Project 2: Clinician Champion Cohort

Identify five to ten clinicians across departments who have expressed interest in AI and form a structured champion cohort. Provide focused AI literacy training, but spend the majority of cohort time on workflow analysis: where would AI realistically help, and where would it be a distraction or a safety risk? The cohort's assessments of specific tools under consideration — grounded in their own domain expertise — give governance decisions a clinical reality check that vendor performance data and literature reviews cannot provide.

Why now: Champion capacity needs to be built before it is urgently needed. An AMC that trains its first cohort while a pilot is already in crisis is late. An AMC that builds champion capacity as standing infrastructure can evaluate new tools, support ongoing deployments, and surface governance concerns from the frontline on an ongoing basis.

Buy vs. build: Curriculum and facilitation. The AAMC and AMA have both published AI literacy curricula that can be adapted without building from scratch. The cohort structure itself — protected time, community of practice, reporting relationship to the AISC — is a governance design and budget decision, not a technology purchase.

5 Barriers and How to Address Them

The standard narrative for why technology fails in academic medicine is as predictable as it is condescending. Clinicians are resistant to change. Faculty lack the expertise to adapt. The institution does not have the resources to see a project through. I have watched dozens of implementations over two decades, and I have rarely found these explanations to be true. The people on the front lines are often desperate for tools that actually work. They do not resist change because they are stubborn; they resist it because they have been burned by software that added three clicks to their workflow without removing a single one.

When an AI implementation fails in an academic medical center, it is almost never because of a lack of will. It is because of structural friction — the grit in the machine that no one wants to name during the kickoff meeting. We fail because governance structures are too slow for the pace of the technology, because electronic health records are built like fortresses rather than platforms, and because we treat AI deployment as a technological event rather than a sociotechnical integration. If we want to move past the pilot phase, we have to stop blaming users and start naming the structural barriers that make adoption genuinely hard. Some of those barriers are fixable. Some are not. This chapter tries to distinguish between them.

5.1 The Governance Vacuum

The most immediate barrier is the silence of the institution. At most AMCs, if a resident wants to use a spreadsheet to track their patients, they just do it. But if that same resident wants to use a language model to summarize a complex patient history, they find themselves in a gray zone — no policy that authorizes it, no policy that prohibits it, and no committee with the authority to make the call. This is the permission paradox, and it produces two outcomes, neither good.

In some institutions, the vacuum creates paralysis. Innovation stalls in the inbox of a risk officer who lacks the technical background to evaluate a tool but has the authority to say no. In others, the vacuum is filled by shadow AI: clinicians and researchers using consumer tools because the institutional alternative does not exist or is harder to access than ChatGPT on a personal phone. They are not trying to be reckless. They are trying to get through their shift. But operating outside any formal framework means their experience cannot be validated, shared, or built on — and it means institutional data is moving through systems with no audit trail, no BAA, and no governance oversight.

The absence of a formal AI Steering Committee is a structural failure of leadership, not a temporary gap to be filled by individual judgment. Without a central body to set the rules for data provenance, validation requirements, and deployment authorization, every department reinvents the wheel independently, and the institutional AI program fragments into a collection of unconnected pilots with incompatible governance. The AISC structure and intake process that addresses this vacuum are described in Chapter 18; the point here is that the vacuum itself is a choice the institution has made, whether or not anyone framed it that way.

5.2 EHR Integration Friction

Even with governance clarity, implementation runs into the technical debt of the electronic health record. Many capable AI tools have failed not because they did not work but because they could not be embedded in the clinician's natural workflow. The integration tax — the hidden labor required to make a modern tool communicate with a legacy system — is not just a matter of writing code. It means navigating proprietary APIs, paying app marketplace fees that can exceed the project budget, and resolving semantic mismatches where the AI's output does not map to any existing field in the clinical record.

Major EHR vendors have historically prioritized billing and documentation over interoperability. FHIR API support varies significantly across EHR versions and implementations, and the access required to embed AI into the clinical workflow is frequently incomplete. The result is brittle integrations: a team spends months getting a pilot working in a specific EHR configuration, and a scheduled vendor update breaks the connection. The clinical AI tool ends up as a sidecar application in a separate browser tab — and we know from years of informatics research that every additional context switch reduces adoption, often fatally.

This barrier is the primary reason why the infrastructure chapter (Chapter 14) emphasizes the institutional API gateway as the architectural foundation for everything else. A gateway that centralizes all AI traffic through a managed chokepoint insulates individual tools from EHR version changes and gives the institution governance leverage over the integration points it would otherwise negotiate tool by tool with each vendor.

5.3 The Alert Fatigue Trap

Clinical decision support has a documented failure mode so well-established it has its own name, and large language model deployment is at risk of reproducing it at scale. Alert fatigue — the clinical survival mechanism that develops when a system generates more low-specificity alerts than the clinician can meaningfully evaluate — has been measured at override rates exceeding 90 percent in traditional CDS deployments. A clinician in an intensive care unit where monitors chirp continuously reaches past the screen and silences the alarm without

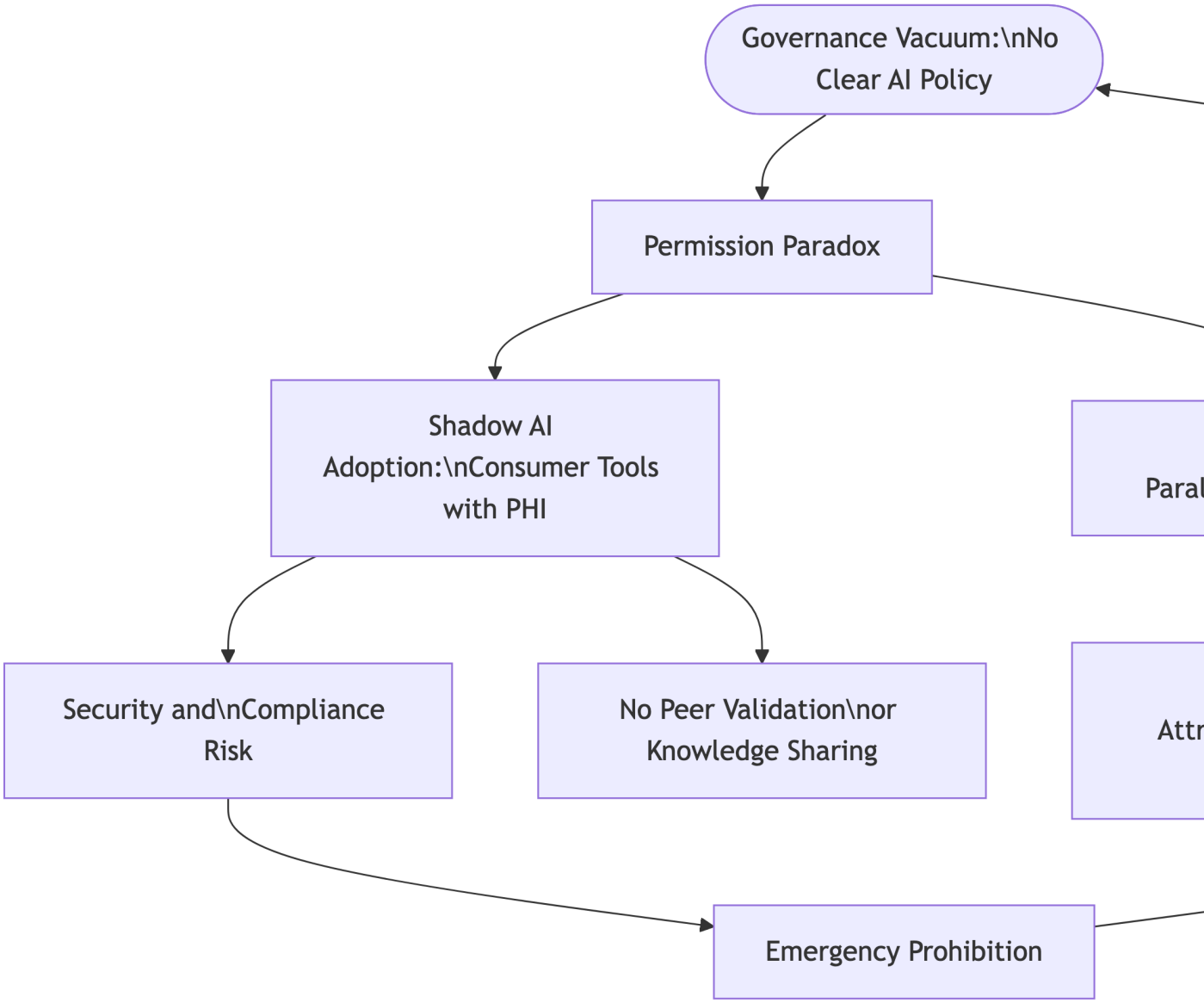


Figure 5.1: The governance vacuum and its downstream effects. Shadow AI adoption is a predictable response to institutional silence, not a security failure to be solved by network blocking alone.

looking at it. This is not negligence. It is rational adaptation to an environment where most alerts carry no useful information.

The external validation of the Epic Sepsis Model illustrates the problem precisely. When Wong and colleagues validated the model against a real patient population, they found that the tool's reported accuracy substantially overstated its performance in the specific context where it mattered most: early identification of patients not yet suspected of having sepsis (Wong et al. 2021). The model was generating alerts that mostly captured patients already on the clinical team's radar, contributing to alert burden without proportional clinical value.

The structural response is not to generate fewer alerts by being more conservative — that just means missing more cases. It is to rethink what AI outputs look like in clinical workflows. AI that surfaces one high-confidence recommendation in a relevant clinical context is more useful than AI that flags everything above a low threshold and lets the clinician decide. This is a design choice, not a model improvement. It requires the clinical informatics team and the clinical leadership to agree on what a "good" alert looks like before the tool is configured, not after the override rate accumulates. The clinical AI governance process in Chapter 6 addresses how to make these design decisions before deployment.

5.4 The Trust Calibration Problem

Trust is the most misunderstood variable in AI adoption. The goal is not to get clinicians to trust AI more. The goal is calibrated trust — where clinicians trust AI outputs in the contexts where the model is reliable and maintain appropriate skepticism in the contexts where it is not. The automation complacency literature documents clearly that consistent accuracy leads to reduced scrutiny, and that reduced scrutiny is precisely where errors compound (Parasuraman and Manzey 2010). A tool that is accurate 95 percent of the time creates a cognitive environment where the 5 percent of errors are less likely to be caught.

Over-trust is one failure mode. Under-trust is the other. A single high-profile error — a hallucinated medication dose, an incorrect lab interpretation, an AI-generated note that a subsequent clinician relied on without independent verification — can cause an entire department to reject a tool that was, on net, performing well. Neither failure mode is solved by more training alone. Both require training programs that address specific failure modes, not just aggregate accuracy, and operational designs that make scrutiny structurally easy rather than cognitively expensive. The training requirements for clinical human-in-the- loop practice are detailed in Chapter 15.

The trust calibration problem is also a transparency problem. When a vendor cannot or will not disclose how a model was trained, what its performance looks like across demographic subgroups, or what types of inputs cause it to fail, clinicians are being asked to trust a black box. In an academic environment built on evidence and accountability, that ask has limited shelf life. The transparency requirements at intake — the model card, the source attributes

required under ONC HTI-1, the demographic performance stratification required under HHS Section 1557 — exist precisely because trust without transparency is not a stable governance posture.

5.5 Budget Fragmentation

In AMC governance, budget is the most accurate statement of institutional priorities, and AI implementation routinely falls into the gap between the IT infrastructure budget and clinical departments' operational funds. I have watched projects die not because they were expensive but because no one could agree on who should pay for the API tokens.

IT views AI as an enterprise capability that clinical departments should fund as operational expense. Clinical departments view it as infrastructure that IT should provide. Research operates on grants that expire in three years. Education has no dedicated technology budget at all. This fragmentation means the institution never builds the shared infrastructure — managed API gateway, data enclave, model monitoring platform — that makes individual projects sustainable. Every project starts by solving the same infrastructure problems the previous project solved, usually in a slightly incompatible way.

Solving this requires a governance decision about AI infrastructure funding, not a clever business case. The right analogy is the institutional network: nobody argues about whether each department should fund its own network switches. The network is shared infrastructure, funded centrally, available to all. AI infrastructure has the same property. The project management chapter (Chapter 18) describes the central AI portfolio budget as a governance mechanism rather than an IT line item.

Table 5.1: AMC AI implementation barrier taxonomy. Each barrier traces to a root cause and a countermeasure described in a specific chapter. The pattern: structural problems require structural solutions.

Barrier	Affected Domains	Root Cause	Countermeasure	Addressed In
Governance vacuum	All	No AISC or policy	Establish AISC + intake process	Chapter 18
EHR integration friction	Clinical, Research	Closed vendor APIs	Institutional API gateway	Chapter 14
Alert fatigue	Clinical	High-volume low-precision CDS	Design-before-deploy, alert review	Chapter 6

Barrier	Affected Domains	Root Cause	Countermeasure	Addressed In
Trust miscalibration	All	Opacity, training deficit	Specific failure mode training, transparency at intake	Chapter 15
Budget fragmentation	All	Siloed funding	Central AI portfolio budget	Chapter 18
Shadow AI adoption	All	Governance vacuum + usability gap	Make sanctioned path easier	Chapter 14

5.6 The Barriers You Cannot Fix

I want to be honest about the constraints that are not within institutional control, because planning around them is different from solving them.

Regulatory lag is real. The FDA, ONC, CMS, and state legislatures are regulating a technology that moves faster than their meeting cycles. The rules in Chapter 10 represent the current state; they will change. The governance program that treats current compliance as sufficient will find itself out of compliance in eighteen months. The governance program built around documented evidence of responsible practice — validation records, monitoring logs, equity audits — is better positioned to adapt because it is building the institutional record that every new regulatory requirement will ask for.

Vendor opacity is a structural feature of the current AI market, not a problem any one institution can solve. Foundation model providers control model weights that AMCs cannot inspect, update models on schedules the institution does not control, and write terms of service that protect their liability rather than the institution's patients. The practical response is contractual: the no-training clause, zero-data-retention provisions, and algorithmic change notification requirements described in Chapter 17. These do not make the opacity go away, but they create contractual leverage that the institution can exercise when a vendor's model changes and clinical behavior changes with it.

The pace of the technology is the constraint that most consistently defeats institution- specific deployment strategies. A deployment built around a specific model version is likely to be disrupted by that model's successor within a year. The infrastructure investment that ages best is not model-specific — it is the governance layer: the API gateway, the audit logging, the validation framework, the monitoring dashboards. Those persist across model generations. The model configurations inside them change. Building for model interchangeability is not a technical aspiration; it is the only governance strategy with a shelf life longer than twelve months.

5.7 Where to Start

5.7.1 Starter Project 1: Barrier Audit and Governance Gap Analysis

Conduct a structured audit of any currently running AI pilots — not the model performance, but the governance experience. How long did it take to get authorization to start? Who paid for it? Who owns it now? Is there a monitoring plan? What happens when it breaks? The answers will map the institution's specific governance gaps more accurately than any framework assessment. Presented to leadership as a roadmap rather than a complaint — these are the hidden costs the institution is already paying, in wasted time and unmanaged risk — the audit becomes the business case for the governance program the institution is not yet running.

Why now: The governance gap is already costing something. Every pilot that runs without authorization is liability exposure. Every shadow AI user is potential PHI exposure. Every tool without a monitoring plan is a future safety event that has not happened yet. The audit surfaces these costs before they become incidents.

Buy vs. build: Analytical work using interviews and document review. No technology required. Frame the output as an internal report to the AISC or equivalent, not as a compliance audit — the goal is to understand the current state, not to assign blame.

5.7.2 Starter Project 2: Shadow AI Discovery and Triage

Rather than attempting to ban unsanctioned AI use — which will fail and will damage the institution's relationship with the clinicians and researchers it needs as partners — try to understand why it is happening. Anonymous surveys or focus groups with residents, junior faculty, and administrative staff will identify the specific tasks that are driving consumer AI adoption: summarizing long patient histories the EHR makes unreadable, drafting patient education materials in multiple languages, abstracting data from unstructured notes. These are the institution's most urgent AI use cases, identified by the people who know the work.

The output of this project is a prioritized list of sanctioned alternatives to develop or procure, grounded in actual use rather than assumed need. The most effective governance response to shadow AI is not the network block; it is the institutional tool that does the job better than the consumer alternative, with appropriate governance built in. The infrastructure chapter (Chapter 14) describes the internal AI assistant as the primary countermeasure to shadow AI. This discovery project is what tells you what that assistant needs to do.

Why now: Shadow AI is not a future risk. It is a present one. Most AMCs that survey honestly find that consumer AI is already in use with sensitive data across their institution. The discovery project defines the scope of the current exposure before an adverse event defines it for the institution.

Buy vs. build: Survey design and facilitation. Anonymous survey platforms are available institutionally. The analysis is qualitative coding of open-ended responses — two to three weeks of research team time, not a technology project.

Part II

Domain Resources

The four chapters in this section move from the general framework to the specific operational terrain of each AMC domain: clinical care, research, education, and business operations. Each domain chapter shares a common structure — what AI is actually being used for, what governance considerations are distinctive to this domain, what the regulatory obligations are, and what an institution can do in the next ninety days — but the content is not interchangeable. A clinical chapter written for clinicians and clinical informatics teams will not serve a research dean's office trying to set AI policy for investigators, and vice versa.

Readers who need the full governance framework before engaging with any specific domain should read the Framework, Values, and Implementation chapters first. Readers who arrived here because a specific domain problem is pressing — a clinical AI deployment under consideration, a research integrity policy to draft, a medical school accreditation review approaching — can start with the relevant domain chapter and treat the framework chapters as reference material. The cross-references in each domain chapter point to the workstream chapters (Data Governance, Infrastructure, Ethics, Workforce Development, and Project Management) for the shared capabilities that cut across all four domains.

6 Clinical Domain

The clinical deployment of AI has moved faster than most governance frameworks anticipated. Emergency departments use predictive sepsis models as standard workflow. Radiology reading rooms incorporate AI preliminary reads before the attending opens the queue. Primary care practices evaluate ambient scribes that transcribe and synthesize the clinical encounter without the physician touching a keyboard. In each setting, the governing question is no longer whether to deploy AI but how to deploy it in a way that extends rather than erodes the quality, equity, and safety of care.

By 2026, three-quarters of U.S. health systems report deploying at least one AI application, up from 59 percent the year before (Fierce Healthcare 2026). Cleveland Clinic's 2025 expanded rollout of Bayesian Health's sepsis detection platform shows what that adoption looks like when the validation work comes first. Applied across more than 760,000 patient encounters, the platform identified 46 percent more sepsis cases than legacy tools, reduced false alerts ten-fold, and generated alerts 5.7 hours before antibiotic treatment (Cleveland Clinic 2025). Those results required rigorous local validation before deployment and sustained monitoring after. The more common pattern is the reverse: tools deployed on vendor-reported accuracy that does not hold at the deploying institution, with governance questions deferred until after the gap shows up in patient outcomes.

This chapter builds on the foundational principles proposed by Badal and colleagues (2023) and extends them into the operational and regulatory terrain that now confronts AMC clinical and informatics leaders. The framework offers an eight-principle scaffold — from alleviating health disparities to facilitating shared decision-making — that remains the most clinically grounded normative structure available for this work. But principles require operational counterparts, and those counterparts now have regulatory teeth. The ONC's Health Data, Technology, and Interoperability (HTI-1) rule mandating algorithm transparency inside EHR workflows, the FDA's guidance on Predetermined Change Control Plans for adaptive AI-enabled devices, and the CMS Medicare Advantage rule prohibiting AI-only coverage denials took effect in 2024 and 2025. Institutions that have not yet mapped those rules onto their existing AI governance structures are already behind.

6.1 The Guiding Principles Framework

The table below summarizes the eight principles Badal and colleagues propose for evaluating AI in clinical and healthcare contexts. Principles 1–4 establish the ethical foundation; Principles

5–8 define the operational requirements for AI tools to be genuinely useful in the specific deployment context.

Table 6.1: Questions that can be used when considering each principle in the AI development process (Badal et al. 2023)

Principle	Questions
1. Alleviate healthcare disparities	What health disparities are reported for the present AI application? How can the AI tool be designed to be accessible to and improve outcomes for the disadvantaged population? What clinical interventions are needed to realize the benefit, and are these accessible? How can data collection be supported in underserved communities for tool retraining over time?
2. Report clinically meaningful outcomes	How is clinical benefit defined in this domain? What is the present threshold for the clinical benefit of existing tools, and how can the AI tool improve upon this threshold?
3. Reduce overdiagnosis and overtreatment	What disease state is an overdiagnosis? For every case of overdiagnosis, what are the downstream costs to the patient and healthcare system? How can this AI application reduce the number of overdiagnoses compared to existing approaches?
4. Have high healthcare value	Is this AI tool addressing a high-priority healthcare need? What would be the cost to the healthcare system in implementation, maintenance, and update? What would be the cost to the patient who does and does not benefit from this tool? Does this tool have high healthcare value, and if not, how can it be improved?
5. Incorporate biography	What biographical data can be collected or carefully coded for the intended population? How do these factors vary in the intended population? How can these factors be included when developing AI tools?

Principle	Questions
6. Be easily tailored to the local population	Can the training features be easily collected in different settings? Are these features reliable for training across different populations? Will the AI/ML workflow be made open-access?
7. Promote a learning healthcare system	How will this AI application be evaluated over time, and at what intervals? What are acceptable thresholds for performance? How will the evaluation results contribute to continuous improvement?
8. Facilitate shared decision-making	Have AI explainability tools been explored and utilized? Do clinicians and patients find the explainability results helpful? Have simpler, explainable algorithms been tried and compared to 'black-box' algorithms to determine if a simpler model performs just as well? How can patient values be easily integrated into the use of the AI tool?

6.1.1 Principle 1: AI Tools Should Aim to Alleviate Existing Health Disparities

Reaching health equity requires eliminating the disparities in health outcomes that are closely linked with social, economic, and environmental disadvantages. At their very core, AI tools require specialized and high-quality data, advanced computing infrastructure, capacity to purchase or partner models from commercial entities, and unique technical expertise — all of which are less available to healthcare systems that serve the most disadvantaged populations.

More careful training and model development that accounts for the unique needs of disadvantaged populations is needed to ensure that AI tools do not exacerbate existing health disparities. Creating equitable AI tools may require prioritizing simpler models for deployment, and the trade-off between balancing accuracy and equity can potentially be resolved by designing tools that can be tailored to the local population. AI tools designed to serve disadvantaged groups must not unnecessarily divert resources from higher-priority areas and more effective interventions (see Principle 4 below).

6.1.2 Principle 2: AI Tools Should Produce Clinically Meaningful Outcomes

AI tools should be evaluated based on their ability to improve clinically meaningful outcomes. The clinical benefit of AI tools should be defined in the context of the existing standard of care, and the AI tool should be evaluated against this standard. If AI practitioners do not define clinical metrics for clinical benefit *a priori*, they risk producing tools that clinicians cannot

evaluate or use. Clinician partners of AI researchers should evaluate accuracy, fairness, and risks of overdiagnosis and overtreatment, as well as the healthcare value and explainability of AI tools and models (see Table 6.1).

6.1.3 Principle 3: AI Tools Should Reduce Overdiagnosis and Overtreatment

Particularly in the United States, overdiagnosis and overtreatment are major drivers of healthcare costs and patient harm. Overdiagnosis occurs when a disease is diagnosed that would not have caused symptoms or death in a patient's lifetime. Overtreatment occurs when a patient is treated for a disease that would not have caused symptoms or death in a patient's lifetime. AI tools should be carefully constructed with attention to the full spectrum of disease and treatment burden, with the goal of reducing unnecessary interventions rather than simply maximizing detection rates.

6.1.4 Principle 4: AI Tools Should Have High Healthcare Value

AI tools applied in healthcare should result in the same outcomes for reduced cost, or better outcomes for comparable cost. Costs to gather inputs, build, maintain, update, interpret, and deploy in clinical practice must be estimated and included in weighing decisions around AI tool adoption. What is cost-effective in one setting may be extremely cost-ineffective in settings where resources are scarce — a point that becomes especially sharp when comparing deployment in well-resourced academic medical centers with the conditions facing safety-net or rural institutions.

6.2 Principles 5–8: From Ethics to Operations

Principles 1–4 address whether a clinical AI tool should be built at all — whether the intervention is equitable, meaningful, safe, and cost-effective. Principles 5–8 address a harder question: whether the tool will work for the specific patients an AMC actually serves, under the specific clinical workflows and social conditions of that institution. These four principles mark the transition from governance-as-ethics to governance-as-operations.

6.2.1 Principle 5: Incorporating Biographical and Structural Drivers of Health

The fifth principle calls for AI tools to account for the full range of social, structural, emotional, and psychological factors that shape health outcomes. In practice, this means models must incorporate social determinants of health (SDOH) as genuine features, not afterthoughts appended during a bias audit.

The problem is that SDOH documentation in the structured EHR is systemically undercaptured. ICD-10 Z-codes — the designated mechanism for recording housing instability, food insecurity, and transportation barriers — appear in a small fraction of encounters for patients known to experience these conditions (Wiens et al. 2019). The information often exists, but in unstructured text: clinician notes, social work assessments, telephone triage summaries. LLMs are proving to be effective extraction tools for exactly this kind of problem. Models applied to clinical narrative notes can identify SDOH indicators with substantially higher recall than structured coding alone, though performance varies across demographic groups and institutions — a disparity that creates its own equity risk and demands ongoing monitoring.

The operational implication for AMC leaders is direct: deploying an AI model without auditing SDOH feature coverage is equivalent to deploying a sepsis predictor on a population where the most at-risk patients have systematically missing data. That audit must happen before deployment, not after the first disparity finding surfaces in a quality review.

6.2.2 Principle 6: Local Calibration for the Local Population

A model trained and validated at a tertiary academic medical center does not automatically perform in a safety-net hospital, a rural affiliate, or a bilingual federally qualified health center. Population shift — the statistical mismatch between training distribution and deployment distribution — is one of the most documented causes of clinical AI failure after initial deployment (Wiens et al. 2019). The sepsis prediction model trained on a predominantly insured, English-speaking cohort may require substantially different decision thresholds when deployed in an emergency department serving a high proportion of recently incarcerated or housing-unstable patients.

Local calibration is the process of adjusting a pre-trained model's parameters or decision thresholds to match the local case mix, documentation practices, and outcome base rates. This is well-established methodology in the biostatistical literature under the name recalibration, and it applies with particular force to predictive models in clinical settings.

AMCs that procure commercial AI tools should expect vendors to provide validation data from populations that approximate their own, and should treat validation studies conducted exclusively in academic medical centers as a cautionary flag when deploying in different institutional contexts. Contract language for AI procurement should include explicit provisions for post-deployment performance monitoring and for model recalibration if performance degrades beyond agreed thresholds.

6.2.3 Principle 7: Promoting a Learning Healthcare System

The seventh principle asks whether an AI tool contributes to continuous improvement over time, or whether it is deployed, validated once, and then left to degrade silently. The learning

healthcare system (LHS) model holds that clinical care generates data, that data enables learning, and that learning improves care in an ongoing cycle.

Clinical AI fits naturally into this loop — but only if the monitoring infrastructure is built into the deployment plan rather than retrofitted after the fact. For AMC leaders, this means specifying performance metrics and acceptable degradation thresholds before deployment; building a mechanism to detect when those thresholds are crossed; and defining the institutional response (recalibration, temporary suspension, vendor escalation, or retirement) in advance of any incident.

The ONC HTI-1 rule, discussed in detail below, adds a regulatory dimension to this operational requirement. For AI tools qualifying as Decision Support Interventions under the rule, vendors are required to provide specific source attributes about the model — including performance characteristics and the populations on which the model was validated — that can be surfaced within the clinical workflow. Institutions should treat these attributes as inputs to ongoing LHS monitoring, not as one-time procurement documents.

6.2.4 Principle 8: Facilitating Shared Decision-Making Through Explainability

The final Badal principle asks whether clinicians and patients can understand why an AI tool produced a given output well enough to incorporate it meaningfully into a clinical decision. This is the explainability question, and it has both a technical and a human-factors dimension that are frequently conflated.

The technical question is which explainability method to apply. The two most widely deployed methods in clinical settings are SHAP (SHapley Additive exPlanations, grounded in cooperative game theory) and LIME (Local Interpretable Model-Agnostic Explanations, which fits a simpler surrogate model locally around each prediction). SHAP values are globally consistent — the same feature receives the same attribution across comparable patients — while LIME explanations can vary for patients with similar risk profiles. For clinical AI tools where consistency across patients matters (discharge planning, readmission prediction), SHAP has become the preferred method. For rapid bedside use where approximate feature attribution is sufficient, the lower computational cost of LIME may be acceptable.

The human-factors question is harder: does presenting explainability output to a clinician actually improve decision quality, or does it create a new form of over-reliance? The evidence is mixed. Clinicians who understand the top features driving a risk score may correctly identify when the model is pattern-matching to correlational rather than causal features (Jones et al. 2023). Clinicians under high cognitive load — which describes most of clinical practice — may instead anchor on the AI output, with the explainability display functioning as post-hoc rationalization rather than genuine scrutiny. Designing human review checkpoints that are genuinely scrutinized rather than rubber-stamped is an open problem addressed in Chapter 11.

For shared decision-making with patients, the bar is different. Patients rarely need SHAP waterfall plots. They need to know that an AI tool was involved in their care, what it was used for, and that their clinician evaluated and took responsibility for the recommendation. Disclosure language that meets this bar is not technically complex; it requires institutional will to adopt and apply consistently.

6.3 Clinical AI in Practice

The principles framework addresses the normative question: what should clinical AI accomplish and what risks should it avoid? The following sections address the operational question: what does clinical AI deployment actually look like in 2025–2026, what regulatory requirements govern it, and where do institutions most commonly stumble?

6.3.1 Ambient Documentation and the Documentation Burden

Physicians spend roughly a third of their working hours on documentation and administrative tasks (American Medical Association 2024b). That fraction has not materially changed since the EHR era began. The most widely deployed category of clinical AI at AMCs is now ambient documentation — AI systems that listen to the clinical encounter, generate a draft clinical note, and return it to the clinician for review and attestation. Abridge[1], Microsoft DAX[2] (Dragon Ambient eXperience), Nabla[3], and Suki[4] are the leading commercial systems; Epic[5] has introduced native ambient documentation for institutions already on its platform.

The evidence that ambient systems reduce documentation time is now substantial. Tierney and colleagues found that physicians using an ambient scribe system reported meaningful reductions in time spent on documentation, with the largest gains in primary care and outpatient specialties where note volume is highest (Tierney et al. 2024). The same work reported improvements in clinician well-being and in patients' perception that their physician was present and attentive during the encounter — because the physician was not looking at a screen. The professional wellness chapter of this book examines this evidence in greater depth (Chapter 13).

The safety concern that receives less attention in the promotional literature is omission error. Ambient systems are optimized to produce fluent, complete-sounding notes, but they can silently omit clinically significant findings mentioned during the encounter — a patient's reference to chest pain that the system categorizes as chronic rather than new, a medication allergy mentioned during the social history, a patient-reported symptom that the system filters out as conversational noise. These omissions are more dangerous than transcription errors

[1] https://www.abridge.com
[2] https://www.nuance.com/healthcare/ambient-clinical-intelligence.html
[3] https://www.nabla.com
[4] https://www.suki.ai
[5] https://www.epic.com

precisely because they produce notes that look complete. A clinician reviewing a well-formatted, plausible note is less likely to detect a missing item than one reviewing a note with obvious structural gaps.

The operational response is to treat ambient documentation as draft generation, not as finished note production — a tool that eliminates the blank-page problem and compresses documentation time, but not one that relieves the clinician of responsibility to verify every clinical fact before attestation. Institutions deploying ambient systems should document this expectation explicitly in clinical AI policy and should monitor attestation patterns: a clinician attesting in under sixty seconds is almost certainly not reviewing the draft.

Figure 6.1: Ambient AI scribe workflow with consent and verification checkpoints. Clinician review before attestation is the critical safety step.

6.3.2 The Federal Regulatory Landscape

Three regulatory actions in 2024–2025 significantly changed the compliance obligations for AMCs deploying clinical AI. Understanding their scope — and their limits — is a prerequisite for sound clinical AI governance.

Table 6.2: Federal regulatory actions affecting clinical AI deployment at AMCs, 2024–2025 (Office of the National Coordinator for Health Information Technology 2024; U.S. Food and Drug Administration 2024; Centers for Medicare and Medicaid Services 2024b; U.S. Department of Health and Human Services, Office for Civil Rights 2024b)

Agency	Rule / Guidance	Effective Date	Key AMC Obligation
ONC	HTI-1: Algorithm Transparency	January 2025	EHR vendors must surface algorithm source attributes (training data, performance, limitations) for Decision Support Interventions within the clinical workflow
FDA	Predetermined Change Control Plan Guidance	December 2024	AI-enabled medical devices may update within pre-specified bounds without new 510(k) or PMA submissions
CMS	Medicare Advantage 2025 Final Rule	January 2025	AI outputs may not serve as the sole basis for coverage denials; human review is required
HHS OCR	Section 1557 Final Rule	May 2025	Covered entities may not apply discriminatory algorithms in patient care decisions

The ONC HTI-1 rule is the most operationally significant for clinical informatics teams. It defines a new regulatory category — Decision Support Interventions — that includes EHR-based algorithms meeting specified criteria for automated decision-making. For each qualifying tool, EHR vendors are required to make accessible a structured set of source attributes: the model's training data sources, performance characteristics, known limitations, and the populations on which it was validated. These attributes must be surfaced within the clinical workflow, not buried in procurement contracts. The practical implication is that AMC clinical informatics leaders should be engaging their EHR vendor about DSI compliance now, and should verify that the source attributes being provided are substantive rather than boilerplate (Office of the National Coordinator for Health Information Technology 2024).

The FDA PCCP guidance addresses the governance of AI tools that learn and adapt after initial deployment. Traditional medical device regulation assumes a static device that functions the same way throughout its useful life. AI systems that update their parameters based on new data do not fit this model. The PCCP guidance establishes a pathway under which developers

specify in advance the types of changes they intend to make, the bounds of those changes, and the performance criteria that must be met before changes are implemented. An AMC that has developed an internally deployed predictive model and intends to update it over time should assess whether the tool qualifies as an FDA-regulated Software as a Medical Device (SaMD) and, if so, whether a PCCP would be the appropriate regulatory pathway (U.S. Food and Drug Administration 2024).

The CMS Medicare Advantage rule responds to documented cases of health insurers using AI-driven prior authorization systems to systematically deny claims with minimal human review. For AMC care management teams working with Medicare Advantage plans, this rule creates leverage: a plan that cites AI output as the basis for a clinical denial without demonstrable human review is operating out of compliance with CMS requirements (Centers for Medicare and Medicaid Services 2024b).

6.3.3 Governing Adaptive AI: FDA SaMD and the PCCP Framework

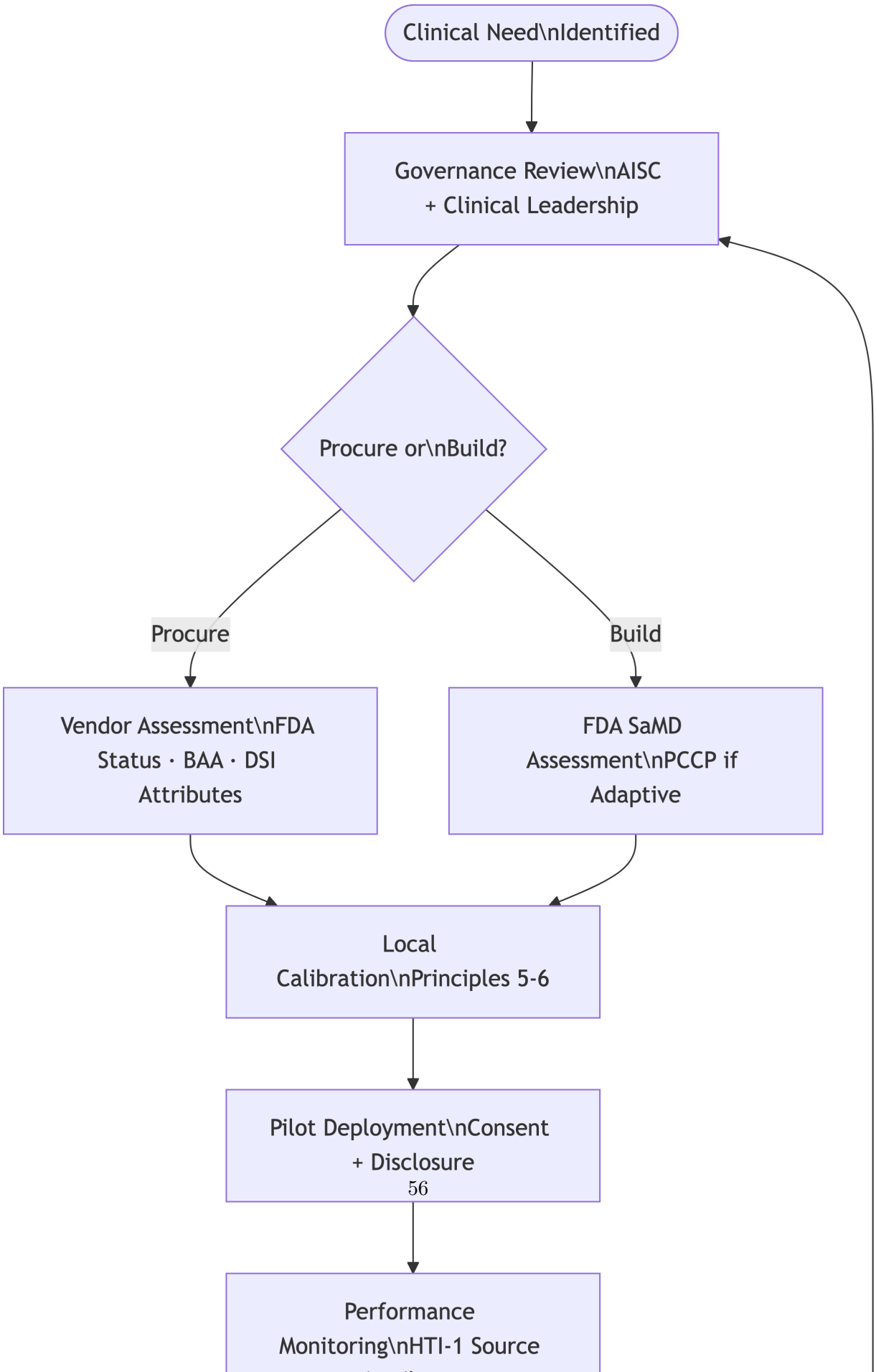
The boundary between a clinical AI tool and an FDA-regulated medical device is not as clear as many AMC leaders assume. The FDA's definition of Software as a Medical Device is broad: software that meets the device definition and is intended to diagnose, treat, cure, mitigate, or prevent disease or other conditions. Many clinical AI tools that AMCs use or develop — risk stratification models, diagnostic decision support, monitoring algorithms — may qualify.

The critical governance question is whether a tool was purchased from a vendor who has obtained FDA clearance or approval, or whether it was developed internally. Vendor-developed tools with FDA clearance carry the regulatory burden on the vendor side; AMC responsibility is largely limited to ensuring the tool is used within its cleared indication. Internally developed tools present a more complex picture. An institution that develops and deploys a predictive model for clinical decision support may be operating an uncleared medical device if the tool meets the SaMD definition — even if the institution does not market or sell the tool externally.

The practical governance recommendation is direct: any clinical AI tool that informs a treatment decision — as opposed to a purely administrative one — should be assessed against the FDA SaMD definition before deployment, and the assessment should be documented. Tools that qualify should be cleared through the appropriate regulatory pathway, structured to meet the clinical decision support exemption criteria, or restricted to research use until regulatory status is resolved (U.S. Food and Drug Administration 2024).

6.3.4 Explainability, Trust, and the Liability Frontier

The liability question in clinical AI is still developing, but the direction is clear. Legal commentary and early judicial precedent are examining two categories of AI-related clinical liability: the failure to exercise adequate oversight of an AI tool that produced a harmful recommendation, and — less intuitive but increasingly discussed — the failure to use an AI tool


Clinical Need\nIdentified
Governance Review\nAISC + Clinical Leadership
Procure or\nBuild?
Procure
Build
Vendor Assessment\nFDA Status · BAA · DSI Attributes
FDA SaMD Assessment\nPCCP if Adaptive
Local Calibration\nPrinciples 5-6
Pilot Deployment\nConsent + Disclosure
Performance Monitoring\nHTI-1 Source

that has become part of the standard of care (Jones et al. 2023). The second category is newer and more unsettling to many clinicians: the idea that using AI could become an obligation, not merely a permission.

The AMA survey data indicates that approximately 65% of physicians see potential value in AI for non-diagnostic tasks — documentation, prior authorization, administrative work — while confidence in AI for diagnostic support is substantially lower (American Medical Association 2023). This gap is clinically rational. The cost of a documentation error is generally lower and more recoverable than the cost of a diagnostic error. But as validated diagnostic AI tools accumulate peer-reviewed evidence of performance at or above attending-level accuracy in specific domains, the hesitancy will come under increasing scrutiny from patients and from liability counsel.

The prudent governance posture is to treat each deployed clinical AI tool as carrying its own liability profile: who is responsible for verifying the output, how that verification is documented in the medical record, and what the institutional protocol is when an AI recommendation is not followed. The clinician who overrides an AI recommendation and documents the clinical reasoning is in a defensible position. The clinician who follows an AI recommendation without review, or who overrides it without documentation, is not (Jones et al. 2023).

6.3.5 Patient-Facing AI and the Consent Gap

The clinical AI tools discussed above are largely invisible to patients: the ambient scribe listens during encounters, but the patient may not know a draft note is being generated; risk stratification models run in the background without patient-facing output. Patient-facing AI tools present a different problem: the patient is the primary user, and the risk of misunderstanding what the tool can and cannot do is concentrated in that interaction.

Epic's patient-facing messaging features, including AI-drafted responses to patient portal messages, are now deployed at dozens of major health systems. A pilot study at UC San Diego Health found that clinician-reviewed AI-drafted patient message responses maintained quality comparable to clinician-authored messages while substantially reducing the time required — though the study also raised open questions about whether patients were adequately informed that AI was involved in the drafting process (Polubriaginof et al. 2024).

California Assembly Bill 3030, effective January 2025, requires healthcare providers to include explicit disclosure when AI has generated patient communications (AB 3030 2024). The law covers written communications — letters, portal messages, discharge instructions — but does not extend to real-time verbal interactions. AB 3030 is the most specific AI disclosure mandate currently in effect at the state level, and several states are considering similar requirements. AMC compliance and legal teams should treat the California standard as a leading indicator of where disclosure requirements are heading nationally, and should evaluate adopting it as a default even for institutions serving patients primarily outside California.

The disclosure language itself does not need to be alarming. A statement that some communications may be drafted with AI assistance and reviewed by the clinical team before sending accomplishes the core transparency goal. Institutions that have adopted such language consistently report minimal patient concern when it is introduced alongside explanation of how the tool is used; the concern is substantially higher when patients discover the practice without prior notice.

6.4 Where to Start

Clinical AI governance can feel paralyzing when confronted simultaneously with FDA guidance documents, ONC rulemaking, the breadth of the Badal framework, and a vendor landscape that changes quarterly. The two projects below are scoped to produce tangible institutional value within six months, using infrastructure most AMCs already have.

6.4.1 Starter Project 1: Clinical AI Inventory and Regulatory Risk Assessment

What it is: A structured inventory of every AI tool currently deployed in clinical workflows — including EHR-embedded tools that clinical informatics teams may not have formally reviewed — with a brief assessment of FDA regulatory status, DSI qualification under HTI-1, and the existence or absence of active performance monitoring.

Why now: The ONC HTI-1 rule is in effect. The FDA PCCP guidance has been published. An institution that cannot identify which of its deployed clinical AI tools are regulated under which frameworks cannot meaningfully comply with either requirement, and cannot answer the liability question if an AI-related adverse event occurs.

How to execute: This is primarily a governance and documentation exercise, not a technical project. The clinical informatics team, the AI governance committee, and the EHR team collectively identify all deployed tools with a clinical workflow touchpoint. Each tool is assessed against a standardized template: FDA clearance status, DSI qualification, BAA coverage, active performance monitoring, and last validation date. The output is a register reviewed by the AI Steering Committee on a quarterly schedule.

Buy vs. build: This is governance work, not a technology purchase. The output is a document and a process. Software tools for AI governance inventory exist (Credo AI, IBM OpenScale) but are not prerequisites for starting; a shared spreadsheet with enforced schema is sufficient for a first-pass inventory at most institutions.

6.4.2 Starter Project 2: Ambient Documentation Pilot with Pre/Post Measurement

What it is: A structured pilot of an ambient AI scribe system in a high-volume outpatient specialty, with pre-specified measurement of after-hours EHR time, note quality, and clinician well-being scores before and after the intervention.

Why now: The evidence base for ambient documentation is strong enough to act on (Tierney et al. 2024). But the institution that deploys without measurement cannot know whether its specific deployment produced the expected benefit, or whether local workflow factors attenuated it — and cannot make the internal business case for broader rollout without outcome data.

How to execute: Select a single outpatient specialty with a measurable documentation burden and an identified clinical champion. Establish baseline metrics before deployment: after-hours EHR time (available from EHR audit logs at most institutions), note length, attestation turnaround, and a validated burnout or well-being survey. Deploy the ambient system with explicit workflow expectations — draft generation, not autonomous documentation — and with the consent and attestation disclosure language described above. Measure the same metrics at 60 and 120 days. Share results transparently with clinical leadership regardless of outcome; a null result has governance value.

Buy vs. build: Buy. The commercial ambient systems have existing EHR integrations, BAA coverage, and vendor-side regulatory responsibility for the AI component. Building an ambient documentation system from scratch is not a tractable project for most AMC informatics teams, and the commercial market has matured rapidly. The institutional work is in selecting the system, structuring the measurement framework, and managing the change management required for sustainable clinical adoption.

7 Research Domain

The academic medical center's research enterprise has never been easy to sustain. Investigators spend a growing fraction of their time on grant writing and manuscript administration rather than on investigation. The volume of published literature has outpaced any individual's ability to track it: PubMed indexed more than 37 million citations by mid-2024, with approximately 1.5 million new records added each year (National Library of Medicine 2024). The peer review system strains under submission pressure, and the reproducibility crisis — well documented across biomedical fields for more than a decade — continues to surface failures that call the research enterprise's integrity into question.

Generative AI arrived into this strained system and immediately found traction, not because investigators were looking for AI specifically but because they needed relief from exactly the tasks where AI is most useful: reading and synthesizing large volumes of text, drafting standard-form documents, and generating plausible starting points for complex writing tasks. The technology also introduced new failure modes — citation hallucination, data fabrication, peer review confidentiality violations — that are serious enough to warrant careful institutional guidance.

This chapter maps the real capabilities and real risks of generative AI in the research lifecycle. It does not argue that AI will transform research; the transformation, where it is happening, is more prosaic than that. It argues that an academic medical center that does not give its investigators clear guidance and appropriate infrastructure for AI use in 2025–2026 is leaving a significant productivity gain on the table and creating an integrity risk at the same time.

7.1 The Information Crisis in Biomedical Research

Understanding why AI has taken hold in research requires understanding the scale of the problem it is solving. Publication volume growth is not a recent phenomenon, but the rate has accelerated. A systematic reviewer completing a comprehensive literature search in oncology or cardiology today can expect to screen thousands of abstracts for a single review. The manual process — reading title and abstract, applying inclusion/exclusion criteria, extracting data from included papers — consumes months of researcher time for a single systematic review. The Cochrane Collaboration, which sets the standard for systematic review methodology, has documented that the time from literature search to publication for a systematic review averages more than two years.

Grant writing represents a comparable burden. Survey data collected by von Hippel and von Hippel found that principal investigators spend between 116 and 170 hours on a single NIH R01 application, with success rates hovering between 18 and 22 percent (Hippel and Hippel 2015). A substantial fraction of that time goes into writing that is formulaic: specific aims language, human subjects sections, data management plans, facilities descriptions. These are exactly the document types where AI drafting assistance has the clearest productivity case.

The reproducibility crisis adds a third dimension. Estimates of irreproducibility in preclinical biomedical research range widely, but a 2015 survey published in *Nature* found that more than 70 percent of researchers had tried and failed to reproduce another scientist's experiments (Baker 2016). The causes are multiple — inadequate methods reporting, insufficient sample sizes, selective publication — but inadequate documentation is a consistent theme. AI tools that help investigators write more precise and complete methods sections, generate analysis code that is transparent and version-controlled, and produce CONSORT- or PRISMA-compliant reporting checklists address a real problem.

7.2 Literature Discovery and Synthesis

The most mature and immediately useful AI applications in research are in literature discovery and synthesis. A new generation of "semantic search" tools has emerged that goes well beyond PubMed keyword search: they embed papers as dense vectors, retrieve semantically similar content, and allow natural-language queries over large corpora.

Elicit (elicit.com) allows investigators to upload a set of seed papers or pose a research question and receive a literature matrix — a table in which rows are papers and columns are user-specified attributes extracted by AI (sample size, outcome measures, key findings, population). This is meaningfully different from a list of search results: it is a first-pass data extraction that a reviewer can validate and extend. Elicit's performance on recall is imperfect — it does not replace a comprehensive database search — but it is a useful starting point for scoping reviews and rapid evidence syntheses.

Consensus (consensus.app) is optimized for answering specific empirical questions: "Does X intervention improve Y outcome?" It returns a meter indicating the degree of published consensus, a list of supporting and contradicting papers, and extracted quotations. It is less useful for exploratory synthesis than Elicit but more useful for answering bounded clinical questions.

SciSpace (scispace.com) and **Semantic Scholar** (semanticscholar.org) offer complementary capabilities — SciSpace for interactive paper reading with AI explanation, Semantic Scholar for citation network analysis and influence mapping.

Across all these tools, the empirical literature on performance is sobering. A 2024 study in the *Journal of Medical Internet Research* evaluated GPT-4 on systematic review tasks in orthopedic surgery and found approximately 13% precision in literature recall — the model retrieved

many irrelevant papers and missed many relevant ones — with substantial hallucination in the reported findings (Chelli et al. 2024). An earlier study found that AI-generated medical abstracts were indistinguishable from real abstracts by human reviewers more than 30% of the time, and by automated detectors more than 60% of the time (Gao et al. 2023). These findings do not argue against using AI tools for literature work; they argue for using them as a first pass that a human reviews, not as a substitute for comprehensive search.

The practical recommendation for an AMC research computing program is to provide investigators with access to two or three of these tools through an institutional account (most offer institutional pricing), include evaluation guidance in researcher training, and set the expectation that AI-assisted literature search supplements but does not replace structured database search (PubMed, Embase, CENTRAL) for systematic reviews.

7.3 Hypothesis Generation and Study Design

The use of AI in hypothesis generation is the most philosophically interesting application and the one most surrounded by hype. The honest account is more modest.

Language models have been used to generate candidate hypotheses in drug repurposing, protein function prediction, and epidemiological research. In each of these domains, the model's output reflects statistical patterns in the training corpus — it will propose hypotheses that are plausible given the existing literature, which means it is most useful for generating well-grounded starting points and least useful for generating genuinely novel ones. A model trained on the biomedical literature will not reliably propose hypotheses that contradict established findings, even when those findings are wrong. The value is in speed and breadth: an investigator exploring a new area can generate twenty candidate research questions in thirty minutes with AI assistance and spend their judgment on evaluating them rather than generating them.

For study design, AI tools are useful for generating draft statistical analysis plans, identifying potential confounders from the literature, drafting power calculations for common study designs, and checking draft methods sections against reporting standards (CONSORT for randomized trials, STROBE for observational studies, PRISMA for systematic reviews). These are tasks that currently require biostatistician or methodologist time; AI cannot replace that expertise, but it can reduce the number of iterations needed before a statistician review.

The buy-versus-build question here is clearly "use existing tools." No AMC should be building a hypothesis generation AI system. The tools that exist — general-purpose models like GPT-4 and Claude, domain-specific tools like Elicit — are adequate for the investigator-facing use cases. The institutional responsibility is to ensure investigators have access to these tools through secure, BAA-compliant channels when the research involves human subjects data.

7.4 Code Generation and Reproducible Analysis

For investigators who work with data — which is most investigators in clinical and translational research — AI code generation is among the highest-value applications in this chapter. The ability to describe a data manipulation task in natural language and receive working R or Python code is genuinely productivity-transforming for investigators who are competent researchers but not expert programmers.

The limitations are important. AI-generated code produces what looks like correct code; whether it is correct depends on whether the investigator can evaluate it. A model asked to perform a mixed-effects regression in R will produce syntactically valid code that often produces numerically plausible results, but it may apply the wrong random effects structure, use the wrong reference level, or fail to account for missing data in the way the investigator intended. An investigator who cannot read the code cannot catch these errors.

The institutional implication is that AI code generation in research requires a floor of computational literacy — the ability to read code, understand its structure, and evaluate whether it implements the intended analysis. This is an argument for including AI-assisted programming in the Training & Workforce Development chapter curriculum, not an argument against using AI for code generation.

Reproducibility is a distinct concern. Research code generated by AI should be committed to version control, documented, and included in data and code availability disclosures. The MarkLLM toolkit[1] and related watermarking approaches offer technical mechanisms for tracking AI-generated code, though institutional adoption of these tools is not yet widespread. At minimum, investigators should document in their methods sections whether AI tools were used for analysis code generation, consistent with the disclosure norms discussed below.

7.5 Manuscript Drafting, Authorship, and the Policy Landscape

The use of AI in manuscript preparation has generated more policy activity than any other AI application in research, and the policies from journals and funders converge on a small number of principles that are worth knowing precisely.

AI cannot be an author. The International Committee of Medical Journal Editors updated its authorship recommendations in 2023 to state explicitly that AI tools do not meet authorship criteria because they cannot take responsibility for the work, cannot consent to authorship, and cannot be held accountable for errors (International Committee of Medical Journal Editors 2023). *Nature*, *Science*, *JAMA*, *NEJM*, and virtually every major biomedical journal has adopted equivalent language. The responsible corresponding author is accountable for any AI-generated content in the manuscript, including any errors it contains.

[1] https://github.com/thu-coai/MarkLLM

AI use must be disclosed. The specificity of disclosure requirements varies by journal. *Cell Press* requires a dedicated "Declaration of Generative AI and AI-Assisted Technologies in the Writing Process" section placed before the references. *Nature* requires disclosure in the methods. *JAMA* requires disclosure of the tool, the extent of use, and confirmation that the author has verified the final content (JAMA Network 2023). An investigator drafting a manuscript should check the target journal's specific requirements; the safe default is to describe in the methods or a dedicated section which AI tools were used and for which tasks.

AI in peer review is prohibited. The NIH issued Notice NOT-OD-23-149 in June 2023 prohibiting peer reviewers from using AI tools to analyze or critique NIH grant applications, citing confidentiality requirements for the content of applications (National Institutes of Health 2023b). The NSF issued parallel guidance for its merit review process (National Science Foundation 2024). Springer Nature explicitly bars uploading manuscripts to AI tools for review purposes. The confidentiality rationale is sound: when a reviewer pastes an unpublished manuscript into a commercial AI service, the manuscript has been disclosed to a third party, potentially in violation of the reviewer's confidentiality agreement. This is true regardless of whether the AI provider claims not to train on the submitted content.

Grant applications are subject to disclosure requirements. NIH has not prohibited the use of AI in preparing grant applications, but has issued guidance on disclosure. The expectation is that the intellectual contribution to the science described in an application is the PI's own. An application in which the scientific narrative was substantially AI-generated would not meet that expectation. The practical boundary is between AI as an editing and drafting tool for standard-form sections (data management plans, biosketches, resource descriptions) and AI as a substitute for the investigator's scientific thinking.

Table 7.1 summarizes the policy landscape for major journals and funders.

Table 7.1: AI policy summary for major biomedical journals and funders. Policies are as of early 2026; investigators should check current journal instructions before submission.

Organization	AI authorship	Disclosure required?	AI in peer review	Policy date
ICMJE (all member journals)	Prohibited	Yes, extent of use	Not addressed	May 2023
Nature	Prohibited	Yes, in methods	Prohibited	Jan 2023
Science	Prohibited	Yes	Prohibited	Jan 2023
JAMA	Prohibited	Yes, tool + extent	Not explicitly addressed	2023
NEJM	Prohibited	Yes	Not explicitly addressed	2023
Cell Press	Prohibited	Yes, dedicated section	Not explicitly addressed	2023

Organization	AI authorship	Disclosure required?	AI in peer review	Policy date
NIH (grants)	N/A	Expected for substantial use	Prohibited (NOT-OD-23-149)	Jun 2023
NSF (grants)	N/A	Expected	Prohibited	2024

7.6 Research Integrity Risks

Two integrity risks in AI-assisted research deserve explicit attention: citation hallucination and data fabrication.

Citation hallucination — the generation of plausible-sounding but non-existent references — is the most widely documented AI failure mode in research contexts. Studies have found hallucination rates in AI-generated reference lists ranging from 30% to over 90%, depending on the model, the prompt, and whether the task explicitly required citation verification (Gao et al. 2023). The pattern is recognizable: a hallucinated citation typically has a plausible author list, a plausible journal, a plausible year, and a DOI that either does not exist or resolves to a different paper. An investigator who does not verify every AI-generated citation before submission will submit manuscripts with fabricated references, which constitutes research misconduct regardless of whether the fabrication was intentional.

The institutional response is a disclosure and verification norm: any AI tool used for literature-related tasks must have its output verified before it enters a manuscript or grant application. This is not a burden unique to AI — investigators are expected to have read the papers they cite regardless of how they found them. AI makes it easier to accumulate unchecked citations, so explicit verification practice is necessary.

Data fabrication through AI is a more serious concern that has attracted less attention in the policy literature. An investigator who asks an AI model to "fill in" missing data points, "smooth" irregularities in a dataset, or "generate" plausible results for an underpowered study is committing data fabrication in exactly the same way as manual fabrication. The fact that the fabrication was AI-assisted does not reduce the culpability. Research integrity training programs should explicitly address AI-assisted fabrication as a recognized misconduct category, not only the traditional forms.

7.7 Human Subjects, Privacy, and Secure Infrastructure

Research involving human subjects data requires particular care in AI tool selection and use. The Common Rule governs research involving identifiable private information; HIPAA governs

the use and disclosure of protected health information. Neither was written with AI in mind, and institutional interpretation of how they apply to AI tool use is still evolving.

The operative question for an investigator is: what data can be entered into which AI tool? A reasonable institutional framework distinguishes four scenarios.

First, **publicly available data or fully de-identified data** (under HIPAA Safe Harbor or Expert Determination) may generally be entered into enterprise AI tools — provided the institutional BAA covers the tool. Whether it can be entered into consumer AI tools (public API, no BAA) depends on whether the data remains genuinely non-identifiable at the population level. The re-identification literature is clear that combinations of demographic variables that appear non-identifying individually can uniquely identify individuals in sparse datasets. When in doubt, treat data from clinical sources as identifiable.

Second, **limited dataset** (HIPAA limited dataset, which retains dates and geographic data at the county level) requires a data use agreement and should be treated as potentially identifiable for AI tool purposes. Enterprise tools with BAAs are appropriate; public APIs are not.

Third, **the full EHR** — identifiable clinical data — requires a BAA with the AI tool provider and should not be entered into any tool that does not have that agreement in place. Research use of full EHR data through AI tools should go through the institution's IRB and through the honest broker process if the investigator is not the treating provider.

Fourth, **genomic and genetic data** is subject to additional constraints — the Genetic Information Nondiscrimination Act (GINA), NIH data sharing policies for GWAS data, and the special re-identification risk of genomic data that is well documented in the literature.

The infrastructure conclusion from this framework is that the research enterprise needs institutional API access to a language model that operates within an enterprise tenant with BAA, not just general permission to use consumer tools. This is a capital and contract decision that should go through the AI Steering Committee, not an investigator-by-investigator determination.

A distinct data governance issue that has emerged from NIH's 2023 Data Management and Sharing Policy is how model weights and training datasets from AI research projects are treated as "scientific data" subject to sharing requirements. The 2025 NIH guidance on AI-derived genomic data specifically designates trained models as data derivatives subject to the same controlled-access restrictions as the underlying genomic data. The practical implications — what must be shared, what may be shared, and what cannot be shared under DMS obligations — are addressed in Chapter 17.

Figure 7.1 illustrates the full research lifecycle with AI tool availability and constraints annotated at each stage.

Figure 7.1: The biomedical research lifecycle with generative AI tool applicability. Green nodes indicate mature, lower-risk AI applications. Yellow nodes indicate useful but higher-risk applications requiring human verification. Red indicates institutional or regulatory prohibition.

7.8 Where to Start: Two Starter Projects

7.8.1 Project 1: Institutional Literature Review Toolkit

What it is. Negotiate an institutional subscription to Elicit or a comparable semantic search tool and configure it with access for all investigators. Create a one-page usage guide that specifies: which tasks the tool is appropriate for (scoping reviews, rapid evidence maps, initial literature exploration), which tasks require supplementation with structured database search (systematic reviews intended for peer review, Cochrane-style meta-analyses), and how to document AI tool use in methods sections per target journal requirements.

What you need to start. A research computing or library liaison who can manage the institutional subscription, budget for a modest annual license (Elicit institutional pricing is public and reasonable), and someone in the research integrity or library team who can write the one-page guide. The guide should take one working day to draft and one meeting with a few senior investigators to validate.

Build or buy? Buy — specifically, subscribe to an existing tool. Do not build a literature search AI from scratch. The commercial tools have invested years in recall optimization over biomedical corpora; a locally built vector search over PubMed will not match them and will require ongoing infrastructure investment to maintain.

What done looks like. Within 90 days: institutional account active, usage guide published on the research computing or library website, at least 20 investigators have used the tool, and a short usage report has been delivered to the research dean showing which departments are using it and for what. The usage data is itself valuable: it shows where investigators are investing in literature work and where training is needed.

7.8.2 Project 2: Secure AI Gateway for Research Computing

What it is. Deploy an institutional API gateway that routes requests to a language model (GPT-4, Claude, or Gemini) through an enterprise-tenanted endpoint with a BAA, logs all

usage by user and project, and enforces data classification rules — specifically, rejecting or flagging requests that contain patterns matching PHI or genomic identifiers. Researchers submit queries through a simple web interface or through an API key issued against their institutional credentials.

Why this matters for research specifically. The alternative — investigators using personal OpenAI or Anthropic accounts for research tasks — creates two problems. It routes research data through personal accounts with no institutional oversight or logging. It makes it impossible to answer the question "what AI tools did you use on this project?" when a journal or funder asks, because there is no institutional record.

What you need to start. The core technical components are an Azure OpenAI or AWS Bedrock deployment (both offer BAAs; setup takes one to two weeks for a system administrator familiar with the platform), a simple API gateway (LiteLLM proxy is an open-source option that adds routing, logging, and cost tracking with minimal configuration), and a credential issuance process tied to the institution's identity provider. The entire system can be built by one competent platform engineer in two to four weeks.

Build or buy? Build the gateway, buy the model. The model capability comes from the commercial provider; the gateway is lightweight infrastructure that the institution controls. This is one of the cases where a modest engineering investment produces a durable capability: once the gateway is running, adding new model providers, new data classification rules, or new usage analytics is incremental work.

What done looks like. Investigators can make API requests to the institutional gateway from their preferred coding environment (R, Python, Jupyter) or from a simple web interface. Usage is logged by project and researcher. A data classification filter is running on prompt content. A quarterly usage report goes to the research dean and the AISC. When a journal asks "did you use AI on this project?", the investigator can answer based on their logged requests rather than on memory.

8 Education Domain

The conversation about generative AI in health professions education has been dominated, since late 2022, by two questions that are increasingly beside the point: Can students use AI to cheat? Can we detect it if they do? Both questions assume that the problem is a student behavior problem, and that better enforcement will solve it. Neither assumption holds, and the obsession with both has distracted education leaders from the more consequential question: what does it mean to teach and assess clinical reasoning when a language model can pass the USMLE?

In January 2023, ChatGPT passed all three steps of the United States Medical Licensing Examination at or above the passing threshold for human test-takers (Kung et al. 2023). That benchmark has since been exceeded: GPT-4 achieves approximately 86% accuracy on Step 1 questions and 87% on Step 2 Clinical Knowledge, performance that meets or exceeds the median medical student (Garabet et al. 2024; Dhakal et al. 2024). The boards were designed to certify that physicians had internalized a body of medical knowledge. They were not designed to certify that a physician could distinguish their own reasoning from AI-generated reasoning. Those are now different things.

This chapter is not an argument against rigor or assessment. It is an argument that the assessment infrastructure of health professions education needs to be rebuilt around what AI cannot yet do — authentic clinical reasoning, communication under uncertainty, integration of patient values — and that the institution's job is to support that rebuilding rather than defend the examination infrastructure that preceded it.

8.1 The Collapse of the Proxy

For decades, written assessments in health professions education served as proxies for the cognitive processes educators actually cared about. A well-written clinical case write-up indicated that the student had synthesized information, generated a differential, and reasoned through management options. A literature review demonstrated that the student could retrieve and critically evaluate evidence. A patient encounter note showed that the student could document a clinical interaction coherently.

These proxies worked because they were hard to produce without actually doing the thinking. They no longer work. A language model given the same case information a student is given can produce a write-up that is, on most dimensions, better than the average student's write-up

— better organized, more comprehensive, with fewer factual errors. It can generate a complete SOAP note from a brief encounter description. It can produce a literature review with a reasonable bibliography.

The problem is not that students are using AI to do these tasks. The problem is that educators designed the tasks to measure one thing and are now trying to use them to measure another. An assignment that genuinely measures clinical reasoning under AI-augmented conditions looks different from an assignment designed to measure unaided recall. The first step in adapting assessment is to be honest about which kind of measure you need.

8.2 The Detection Trap

The institutional response that gained the most traction in 2023 was enforcement: updated honor codes prohibiting AI use, mandatory disclosure requirements for AI assistance, and AI detection tools. The detection tools — Turnitin[1]'s AI writing detection, GPTZero[2], Copyleaks[3] — were marketed at a moment of institutional anxiety and adopted widely before their limitations were understood.

Those limitations are substantial. Detection accuracy varies significantly across writing styles, assignment types, and AI models. False-positive rates — flagging human-written work as AI-generated — disproportionately affect non-native English speakers, whose more formal and pattern-consistent prose is more likely to trigger detection thresholds (Liang et al. 2023). Students who have been falsely accused of AI use in high-stakes grading contexts have documented the experience as deeply damaging, and at least one institution faced legal exposure over a misapplied detection finding.

The Federal Trade Commission's enforcement posture on AI detection tools — including actions against companies that overstated their detection accuracy — has added a compliance dimension to adoption decisions (Federal Trade Commission 2024). An institution that implements a detection tool based on vendor accuracy claims without independent validation is making a factual representation it may not be able to defend.

Prohibition without a structural alternative produces the same behavior with added incentive to hide it. Students who are prohibited from using AI but who observe that their future clinical colleagues use it constantly experience the prohibition as arbitrary rather than principled. The honest conversation with students is not "AI use is prohibited" but "here is what you need to be able to do on your own, here is why, and here is how we assess that."

[1] https://www.turnitin.com
[2] https://gptzero.me
[3] https://copyleaks.com

8.3 A Workable Framing: Tiered Policies and Process Grading

The framing that has produced the most defensible and educationally coherent institutional policies is a tiered approach that distinguishes among assignment types rather than making a blanket determination about AI use.

Ethan Mollick and Lilach Mollick's work on assigning AI (Mollick and Mollick 2023) describes a spectrum of pedagogical stances: AI prohibited (where unaided performance is the learning objective), AI as a tool to be used transparently (where the process of working with AI is part of what is being taught), and AI use as contextually unrestricted (where the product quality matters more than the production process, as it often does in practice). The same framework applies in health professions education, where the three stances correspond roughly to high-stakes clinical reasoning assessments (unaided performance required), literature review and synthesis tasks (AI as tool, process documented), and clinical note drafting and patient communication practice (AI use mirrors professional reality).

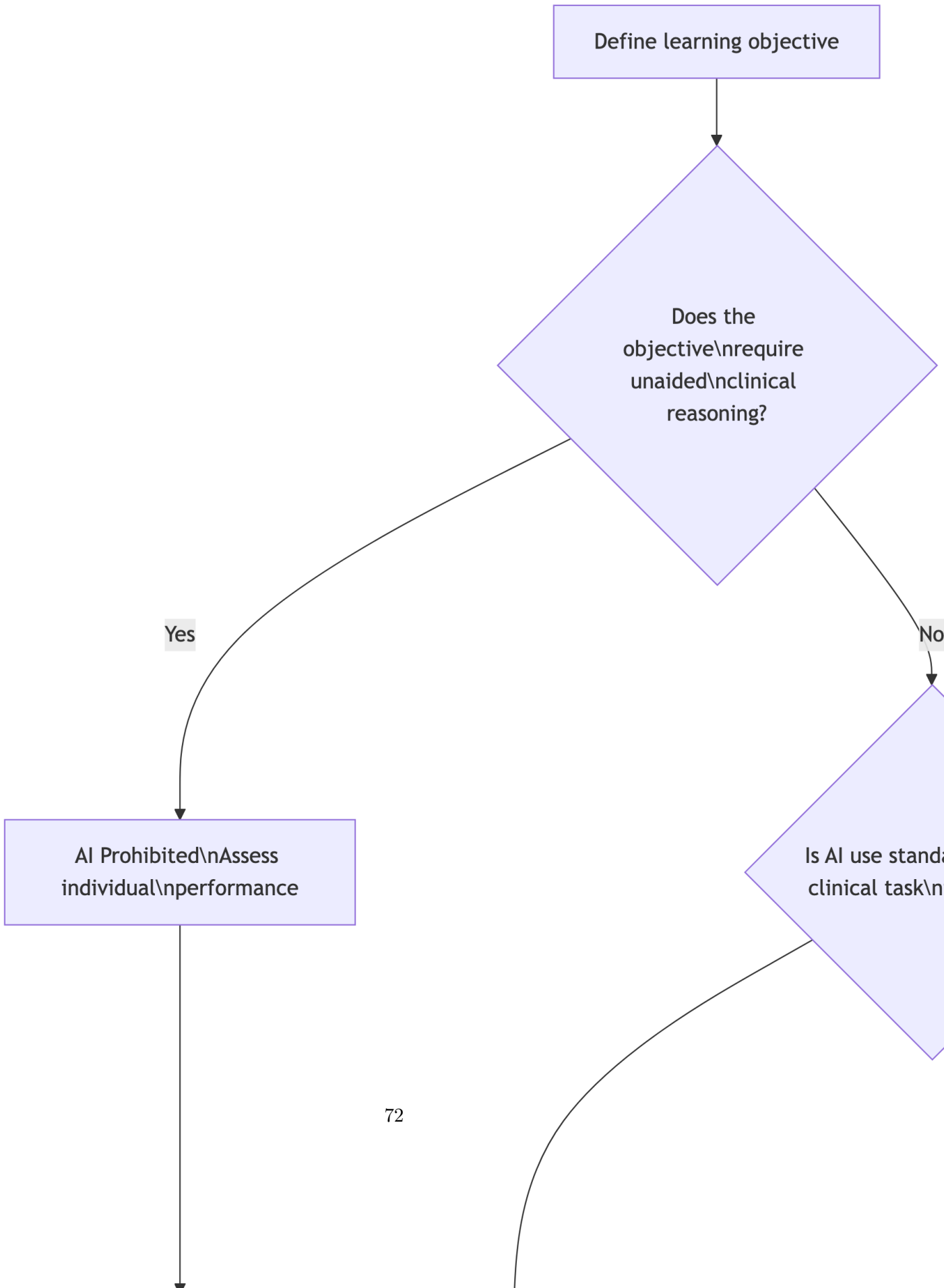
Figure 8.1 shows a decision tree an educator can use to assign assessment tier.

The shift from product grading to process grading is the most important structural change available to educators. Grading the process — the prompt log, the AI output, the student's verification and editing decisions, the written reflection on what the AI produced versus what the student changed — measures exactly the skills that matter in AI-augmented clinical practice: the ability to evaluate AI output critically, identify errors, and exercise professional judgment over a generated starting point. This approach is more labor-intensive to design and grade than a conventional written product; it is also considerably harder to outsource entirely to AI, because the reflection on AI output requires a human who actually did the work.

8.4 The USMLE and the Limits of Licensing Exam Reform

The licensing exam context deserves specific attention because it is the external constraint that shapes so much of undergraduate medical education. When AI passes Step 2 CK at 87%, the natural question is: what are we preparing students for?

The answer from the National Board of Medical Examiners — which has been studying AI performance on its exams since 2023 — is nuanced (Yaneva et al. 2024). High AI performance on multiple-choice clinical vignettes does not mean that AI can replicate authentic clinical reasoning. The vignettes test pattern recognition on structured information; clinical practice requires integrating incomplete, contradictory, and affectively complex information in real time, with a patient whose presentation does not fit the textbook pattern. AI is weaker on these tasks, and considerably weaker on the communicative and relational dimensions of clinical care that licensing exams do not currently measure.


Define learning objective
Does the objective\nrequire unaided\nclinical reasoning?
Yes
No
AI Prohibited\nAssess individual\nperformance
Is AI use standa
clinical task\n

The implication for medical education is not that the boards are irrelevant but that they are increasingly insufficient. Competency-based medical education (CBME) frameworks — the ACGME milestones, the entrustable professional activities — provide a structure for assessing the capabilities that board exams do not capture. Educators who redesign assessment to emphasize direct observation, structured clinical encounters, and oral examination of reasoning are building toward an assessment model that remains meaningful in an AI-augmented world. Educators who optimize for multiple-choice performance are not.

8.5 AI Literacy as Curriculum

The question of what medical students should know about AI was the subject of a foundational 2020 paper by McCoy et al. in *npj Digital Medicine* (McCoy et al. 2020). Their argument — that medical students need literacy in AI as users and evaluators rather than as developers — remains the right framework. The specific competencies have been elaborated since then, and as of 2025 the question has moved from "should we teach this?" to "what exactly are we required to teach and how do we demonstrate compliance?"

The AAMC's 2025 AI Competencies Across the Learning Continuum, developed through a formal Delphi process involving medical educators across the country, provides a national standard for AI competency across undergraduate, graduate, and continuing medical education (Association of American Medical Colleges 2025). The ACGME's July 2025 Common Program Requirements include explicit provisions on human-AI teamwork and require residency programs to have institutional policies on AI use — the first time ACGME has required programs to govern AI specifically (Accreditation Council for Graduate Medical Education 2025). Programs without documented AI governance are out of compliance starting with the 2025 to 2026 accreditation year. The American Medical Informatics Association has proposed a competency framework for health professionals that distinguishes among AI consumers (who use AI tools in clinical work), AI translators (who evaluate AI tools and advocate for appropriate use), and AI developers (who build and validate AI systems) (American Medical Informatics Association 2024).

The consumer/translator/developer distinction is useful for curriculum design because it implies that different learners need different things. Every medical student needs consumer-level AI literacy: how to evaluate AI-generated clinical information, how to recognize hallucinations and errors, how to document AI use in clinical settings, and how to have an informed conversation with a patient about AI's role in their care. Only some students will become translators or developers, and the curriculum should not require developer-level technical knowledge of everyone.

What this means concretely: a required AI literacy module in medical school should cover the basic capabilities and limitations of language models (without requiring programming), the specific failure modes relevant to clinical use (hallucination, bias, over-confidence), the disclosure and documentation standards expected by major journals and by the clinical record,

and how to evaluate an AI tool for fitness in a specific clinical task. This module is not optional enrichment — it is basic preparation for the clinical environment students are about to enter.

Table 8.1 summarizes the major competency frameworks proposed for health professions AI literacy.

Table 8.1: Major AI literacy competency frameworks for health professions education, with target audience and emphasis. The 2025 AAMC and ACGME entries reflect requirements that became effective for the 2025–2026 academic year; programs should verify current language directly with each organization.

Framework	Organization	Year	Target audience	Core emphasis
AI Literacy for Medical Students	McCoy et al., *npj Digital Med*	2020	Medical students	Evaluation, bias, communication
Principles for Responsible AI Use	AAMC	2024	Medical educators	Faculty development, equity
AI Competencies Across the Learning Continuum	AAMC	2025	UME/GME/CME programs	National standard; Delphi-developed competency set
AI Competencies for Health Professionals	AMIA	2024	Health professionals	Consumer/translator/developer tiers
Common Program Requirements (AI provisions)	ACGME	2025	Residency programs	Human-AI teamwork; institutional AI policies required
Preparing Clinicians for AI	James, Wachter, & Woolliscroft, *JAMA*	2022	All clinicians	Reasoning under AI uncertainty

8.6 Institutional Policies: What Has Actually Been Published

The most useful institutional AI policies in health professions education are the ones that specify which tools are permitted, under what data handling conditions, and with what disclosure requirements — not the ones that simply declare a principle of "responsible use." Several health professions schools have published policies worth examining.

UCSF's School of Medicine has implemented a tiered system (AI-1 through AI-4) that signals permitted AI use levels for each assessment, with AI-4 designating full integration of approved tools. The system uses an institutional AI environment ("Versa Chat") that keeps patient-related data within a HIPAA-compliant boundary. The explicit data-security specification — which tool, under what conditions — is the element most often absent from general university policies and most necessary for a clinical training environment.

Johns Hopkins University has published a red/yellow/green framework for courses, with green indicating that AI use consistent with professional norms is expected, yellow indicating that AI use is permitted with documentation, and red indicating that AI use is prohibited and the assessment tests unaided performance. The color system is simple enough for students to internalize without reading a policy document for every course.

The common element across the most effective policies is that they specify the institutional tool rather than deferring to student choice of tool. This is a data governance decision as much as an educational one: medical students on clinical rotations are surrounded by patient information, and a policy that permits "any AI tool the student finds useful" is a policy that permits HIPAA exposure.

Table 8.2 shows a comparison of selected health-professions AI policies.

Table 8.2: Comparison of health-professions institutional AI policies on key implementation dimensions. Policies were documented in early 2026; institutions update these frequently.

Institution	Tier/color system	Approved tools specified?	Data security requirement	Disclosure required?
UCSF School of Medicine	AI-1 to AI-4	Yes (Versa Chat)	HIPAA-compliant only	Yes
Johns Hopkins University	Red/Yellow/Green	Partial	Not specified	Yes
Mayo Clinic College	None (principled guidance)	No	"De-identified only"	Yes
Harvard Medical School	Case-by-case faculty decision	No	Not specified	Yes

8.7 Academic Integrity as a Patient Safety Issue

The stakes of academic integrity in health professions education are not equivalent to the stakes in a general undergraduate program. A student who outsources a history paper to AI has submitted fraudulent work, and the consequence is primarily to their own learning. A medical student who outsources clinical reasoning to AI during training — who presents AI-generated differentials and management plans as their own reasoning through repeated assessments — has learned to operate a clinical workflow they do not understand. The consequence is potentially borne by future patients.

James, Wachter, and Woolliscroft made this argument directly in their 2022 JAMA piece on preparing clinicians for an AI-influenced clinical world (James et al. 2022). The concern is not that AI will replace clinical reasoning but that clinicians who have not developed strong independent reasoning will not be equipped to recognize when AI reasoning is wrong. Automation bias — the tendency to accept automated system output without adequate critical evaluation — is well documented in aviation and other high-stakes domains, and there is no reason to believe clinical medicine is immune.

This reframes the academic integrity conversation from "students are cheating" to "we are responsible for training clinicians who can function safely when AI fails." That is a more honest and more motivating frame for curriculum redesign. It also makes the argument for genuine investment in assessment reform: not detection systems that catch cheaters, but assessment methods that require demonstrating the reasoning that makes a physician safe.

8.8 Where to Start: Two Starter Projects

8.8.1 Project 1: AI Literacy Module for All Health Professions Students

What it is. Develop and deploy a required, one-hour AI literacy module covering: how language models work at a conceptual level (no programming), what they are good for and where they fail, how to evaluate AI-generated clinical information critically, how to document AI use in clinical records per institutional policy, and how to discuss AI with patients. The module is assessed with a short practical exercise — evaluate a clinical AI output and document its errors — not a multiple-choice knowledge test.

What you need to start. A course director or clinical informatics faculty member willing to champion it, two to four weeks of faculty time to write and pilot the content, and an LMS slot in a required course. The AAMC's published principles and the McCoy et al. framework provide the content scaffold; you are writing the local application, not the theory.

Build or buy? Write locally. National-level modules are available (the AMA and AAMC have published resources) but they are not calibrated to your institutional tools and policies.

The module needs to name your approved AI tools, your data classification rules, and your disclosure requirements — none of which a generic national module will include.

What done looks like. Within one academic year: every student in the graduating class has completed the module, the practical exercise has been piloted and revised, at least one faculty group has reviewed it for currency, and the module is integrated into the orientation or first-year curriculum on a recurring basis.

8.8.2 Project 2: Assessment Redesign Workshop for Course Directors

What it is. A half-day faculty development workshop that walks course directors through the tiered assessment framework, gives them a structured process for auditing their current assignments against AI capability, and supports them in redesigning one high-stakes assignment per course to assess process rather than product. The output of the workshop is a concrete revised assignment, not an abstract policy commitment.

Why this matters more than policy. Policy says "use AI responsibly." A redesigned assignment with a prompt log requirement and a structured reflection makes "responsible use" measurable and teaches the skills that matter. Faculty who have redesigned one assignment understand the framework well enough to apply it to others.

What you need to start. A facilitator familiar with writing-studies pedagogy and clinical education (or one from each), three to five volunteers from courses that have explicit writing or reasoning components, and a half day in the academic calendar. The Mollick & Mollick SSRN paper (Mollick and Mollick 2023) and the UCSF tier system provide the conceptual scaffolding.

Build or buy? Design locally and run it yourself. National faculty development resources on AI exist (AAMC, AACME), but they are not structured to produce revised assignments. The workshop only produces value if it ends with a concrete artifact.

What done looks like. Five course directors have revised one assignment each, the revised assignments have been piloted for one academic cycle, and a follow-up session has been held to share what worked and what needed adjustment. The materials from the workshop — the assessment audit rubric, the redesign process — are available for any course director who wants to use them independently.

9 Business Operations

At most academic medical centers, the first place generative AI takes hold is not the clinic or the research lab. It is the finance department's revenue cycle team, the HR business partner drafting job descriptions after hours, the communications writer who discovered that a rough prompt produces a passable first draft in thirty seconds. Business operations — the finance, HR, supply chain, IT, marketing, philanthropy, facilities, and legal functions that keep an AMC running — are where AI adoption tends to move fastest and where institutional governance tends to lag furthest behind.

This gap is partly structural. Clinical AI requires FDA clearance or institutional validation before it can touch a patient. Research AI is constrained by IRB requirements and publication ethics. Business-operations AI faces none of these formal gatekeepers, which makes it both the easiest domain to start in and the one most likely to produce a compliance incident before the institution has thought through its policies.

The argument of this chapter is simple: business operations is the right place to build early institutional capability with generative AI, and the regulatory exposure is higher than most operations leaders assume. Getting both of those things right at the same time is possible — it requires a governance-first procurement approach, a clear data classification framework, and the discipline to deploy an enterprise-grade platform before employees have reason to route around it.

9.1 A Domain Unlike the Others

The four domains of this framework — education, research, clinical, and business operations — differ not just in what they do but in the regulatory regimes that govern what AI can do inside them. Clinical AI operates under the most demanding oversight: every tool that influences a clinical decision is potentially a medical device subject to FDA regulation, and deployment without local validation is a patient-safety risk. Research AI is constrained by the norms of research integrity, the requirements of IRBs for human-subjects data, and the intellectual-property rules of journals and funders. Educational AI sits inside a web of academic integrity policies and accreditation requirements.

Business operations has fewer of these formal guardrails, which creates a false sense of safety.

Median U.S. hospital operating margins reached 1.3 percent in 2025 and turned negative at the start of 2026, with labor accounting for 84 percent of total hospital expenses (Kaufman Hall 2026). The nursing workforce gap approaches 500,000 vacancies nationally, with 40 percent of currently employed nurses reporting plans to leave the profession by 2029 (American Hospital Association 2025). Operations leaders reaching for AI-assisted administrative tools against that backdrop are making a reasonable call. The compliance obligations attached to those tools do not diminish because no clinical oversight body is in the room.

The absence of FDA clearance requirements does not mean an absence of legal risk. Three areas carry particular exposure.

The first is employment. Automated systems used to screen, rank, or evaluate employees and candidates are subject to anti-discrimination law regardless of whether they involve AI. The EEOC's Strategic Enforcement Plan for fiscal years 2024–2028 explicitly names algorithmic discrimination in employment as an enforcement priority (U.S. Equal Employment Opportunity Commission 2024). New York City's Local Law 144, which took effect in July 2023, mandates annual independent bias audits for any automated employment decision tool used in hiring or promotion for New York City-based roles (Local Law 144 2021). Colorado's SB 24-205, signed in 2024, imposes "reasonable care" requirements on developers and deployers of high-risk AI systems, including those used in employment contexts (SB 24-205 2024). An AMC with employees or applicants in these jurisdictions — which includes most major academic medical centers — needs to account for these requirements before deploying any HR automation tool.

The second is consumer protection and marketing. The Federal Trade Commission's Operation AI Comply, announced in September 2024, documented enforcement actions against companies making deceptive claims about AI capabilities in consumer-facing contexts (Federal Trade Commission 2024). Patient-facing communications that overstate AI accuracy or imply diagnostic capability without adequate disclosure carry both FTC and state consumer protection risk.

The third is HIPAA — specifically, the risk that an employee uses a business-operations AI tool in a way that incidentally exposes protected health information. A revenue cycle coordinator who pastes a patient name and insurance ID into a public AI chatbot to draft a denial appeal letter has created a potential HIPAA breach, regardless of how routine the business context feels. The HHS Office for Civil Rights has stated that nondiscrimination requirements extend to algorithmic decision support tools in administrative contexts under the Section 1557 final rule (U.S. Department of Health and Human Services, Office for Civil Rights 2024b). The business case for an enterprise-grade AI platform is partly a compliance case.

9.2 What Business Operations AI Actually Does

Generative AI's footprint in AMC business operations is already broad. The use cases vary considerably in maturity and risk, and a realistic function-by-function account is more useful than a general endorsement.

Finance and revenue cycle. The revenue cycle is among the most labor-intensive administrative functions in any hospital, and also among the most data-structured — claims have standard formats, denial reasons follow defined codes, and appeal letters follow recognizable templates. AI tools are most mature here. Denial management, prior authorization drafting, and coding review are all areas where commercial AI-enabled platforms have demonstrated measurable improvement in denial overturn rates and coder productivity. The risk is vendor lock-in and the assumption that an AI platform's performance in another institution's EHR will replicate in yours. Local validation against your payer mix and documentation patterns is necessary before attributing cost savings to the tool.

Financial planning and analysis benefits more from AI-assisted modeling and scenario generation than from automation. An FP&A analyst who can ask a language model to explain the variance in a budget line, generate competing budget scenarios from a shared set of assumptions, or draft the narrative for a board presentation will work faster and produce better prose — but the model will not catch a misclassified expense or a projection built on an invalid assumption. AI augments the analyst; it does not replace the financial judgment.

Human resources. The clearest near-term opportunity is in drafting: job descriptions, offer letters, standard policies, and employee Q&A responses. These are tasks that consume significant HR staff time and produce documents that are largely similar from one iteration to the next. An enterprise AI tool configured to follow the institution's style, terminology, and legal requirements (pay equity language, for example) can reduce the time from request to first draft dramatically.

The clearest risk is in screening. Any tool that ranks candidates, filters applications, or produces scores used in hiring decisions is potentially an automated employment decision tool under the legal definitions of NYC Local Law 144 and related state laws. The bias-audit requirement applies. An institution that deploys an AI-based screening tool without an independent audit and a documented remediation process is exposed. The safer path in 2025–2026 is to use AI for drafting and research tasks — generating interview question banks, summarizing employment law updates, drafting performance review templates — while keeping screening and ranking decisions human-driven.

Supply chain and contracting. Contract review and vendor risk assessment are natural fits for generative AI: the task requires reading large volumes of structured text, extracting specific terms, and comparing them against standards. AI tools that summarize contract terms, flag non-standard clauses, and generate redlines against a baseline template are available from multiple vendors and are being used in healthcare procurement. The risk here is primarily one of accuracy and audit trail. A contract reviewed by AI and approved by a human who did not re-read the AI's summary is a contract reviewed by no one in any legally meaningful sense. The tool produces a starting point; a qualified human closes it.

IT operations. Ticket triage, incident summarization, code review, and internal documentation generation are all areas where IT teams at major AMCs are already using AI, often through GitHub Copilot, Microsoft 365 Copilot, or custom deployments on internal codebases. The

productivity gains for experienced developers and system administrators are real and well-documented. The risk for IT operations specifically is prompt injection — the vulnerability by which malicious input to an AI system can cause it to execute unintended actions. Any AI tool that can invoke shell commands, API calls, or database queries based on natural-language input must be treated as a security surface, not just a productivity tool.

Marketing and communications. Content drafting, translation, accessibility reformatting, and social media response generation are areas where the productivity case is straightforward and the risk is manageable. The FTC's enforcement posture means that AI-generated content making clinical or research claims needs a human review step before publication. The deeper issue is brand and voice consistency: a communications team that uses AI to draft a hundred pieces of content without a style framework will produce a hundred pieces that sound subtly generic. A deliberate style guide and a review process that treats AI output as a first draft, not a final product, addresses most of this.

Philanthropy and development. Donor research summarization, draft cultivation letters, and grant prospect identification are all tasks that consume significant staff time at AMC development offices and are tractable for AI. The risk is relationship sensitivity: a donor who receives a letter that was obviously AI-generated from a public database — especially one that contains a factual error about their history with the institution — is a donor at risk of disengagement. AI-generated development communications need a personal review step before they leave the institution.

Facilities and legal. Work-order routing, maintenance request summarization, and facilities policy Q&A are low-complexity AI use cases with low risk. Legal is more sensitive: AI tools for contract drafting and policy research are useful, but anything that constitutes legal advice or professional judgment remains the responsibility of licensed counsel.

Table 9.1 summarizes the function-by-function landscape.

Table 9.1: Business operations AI use cases by function, risk level, and key constraint.

Function	Mature use cases	Risk level	Key constraint
Revenue cycle	Denial drafting, coding review	Medium	Local validation required
HR	JD drafting, policy Q&A	Medium–High	Bias audit if screening
Supply chain	Contract review, redlining	Medium	Accuracy + audit trail
IT operations	Ticket triage, code review	Medium	Prompt injection risk
Marketing	Content drafting, translation	Low–Medium	FTC and brand review
Philanthropy	Donor research, letter drafting	Low	Relationship sensitivity
Facilities	Work-order routing, policy Q&A	Low	Minimal

Function	Mature use cases	Risk level	Key constraint
Legal	Contract research, drafting	Medium	Professional judgment stays human

9.3 The Regulatory Layer

Most AMC operations leaders have read about HIPAA. Fewer have encountered the employment AI laws, the FTC's AI enforcement posture, or the specific nondiscrimination requirements that apply to algorithmic decision tools. A useful organizing frame is that the regulatory exposure in business operations AI falls into three areas: employment decisions, patient-facing communications, and data handling.

The employment area is the most actively litigated. The EEOC has been explicit that Title VII, the ADA, and the ADEA apply to automated employment decisions, and that employers cannot shift liability to a third-party AI vendor (U.S. Equal Employment Opportunity Commission 2024). NYC Local Law 144 requires employers using automated employment decision tools for NYC-based roles to conduct and publish annual independent bias audits and to notify candidates before the tool is used (Local Law 144 2021). Colorado's SB 24-205 requires "deployers" of high-risk AI systems (which include employment AI) to implement risk management programs, conduct impact assessments, and notify consumers when they are subject to an algorithmic decision (SB 24-205 2024). An AMC that operates in New York, Colorado, or is a federal contractor faces compliance obligations that are already in effect. Several other states — Illinois, Texas, Utah — have enacted or proposed similar legislation.

The Federal Trade Commission's "Operation AI Comply" matters for AMC marketing teams in particular. The FTC's enforcement actions targeted companies that made false or unsubstantiated AI capability claims in customer-facing contexts, including claims that AI had reviewed or validated health-related content (Federal Trade Commission 2024). Patient communications that describe AI tools as "reviewed by clinical AI" or "AI-verified" without substantiated validation claims invite FTC scrutiny. The safer approach is to describe what the AI actually did (drafted, organized, translated) rather than making accuracy claims.

The HIPAA exposure in business operations is subtler than the exposure in clinical domains, but it is real. Revenue cycle, HR (benefit eligibility, medical leave records), and even marketing (patient testimonials, program enrollment data) are all areas where patient information can appear in business-operations workflows. The HHS Section 1557 nondiscrimination rule makes explicit that the nondiscrimination requirements extend to algorithmic tools used in the administration of health programs, not just to clinical decisions (U.S. Department of Health and Human Services, Office for Civil Rights 2024b). A revenue cycle AI tool that systematically produces less aggressive appeal letters for Medicaid patients than for commercially insured patients is potentially a Section 1557 violation, regardless of whether that disparity was intentional.

The OMB's Memorandum M-24-10 on advancing AI governance, while directed at federal agencies, establishes the federal government's expectations for institutions receiving significant federal funding — which describes every AMC (Office of Management and Budget 2024). Its requirements around rights-impacting and safety-impacting AI systems, and its emphasis on impact assessments and transparency, set a standard that AMC AI governance programs should be prepared to meet or justify deviating from.

9.4 The Shadow-IT Problem

The single most common business-operations AI event at an AMC today is not a deliberate deployment: it is an employee using a consumer AI tool — ChatGPT, Claude.ai, Gemini — with institutional data, without authorization, because no better option is readily available.

The dynamics are predictable. Administrative staff are under pressure to produce more documentation with fewer resources. Consumer AI tools are free, fast, and effective for exactly the drafting, summarizing, and reformatting tasks that dominate their workload. The institutional alternatives — if they exist — require tickets, approvals, and onboarding processes that take weeks. The employee uses the consumer tool once, finds it dramatically faster, and makes it a habit.

The problem is not that employees are careless. The problem is that the institution has not given them a compliant, convenient alternative. The solution is not a prohibition — prohibition without an alternative produces the same behavior plus an incentive to hide it. The solution is a gateway: an institutionally managed AI platform that is as easy to reach as the consumer tools, covered by appropriate data handling agreements, and integrated into the workflows employees actually use.

This is the core argument for enterprise AI tenancy, which is addressed in the next section. The point here is that the governance problem and the security problem in business operations AI are, at bottom, the same problem: the institution has not built a path of least resistance that is also a path of least risk.

9.5 Enterprise AI Platforms

The practical procurement question for most AMCs in 2025–2026 is not whether to build a large language model, but which enterprise platform to deploy and how to configure it. Three platforms dominate the healthcare market: Microsoft 365 Copilot, Google Workspace AI, and ChatGPT Enterprise. Anthropic's Claude for Enterprise is a smaller but growing option. Each offers a different trade-off between integration depth, data handling controls, and administrative overhead.

Microsoft 365 Copilot is the obvious choice for institutions already running Microsoft 365, which describes most AMCs. It runs inside the existing Microsoft tenant, inherits existing Entra ID (Azure Active Directory) identity and access management, and operates under Microsoft's existing HIPAA Business Associate Agreement. The data does not leave the tenant for model training. The limitation is that Copilot's performance is tightly coupled to the quality of the institution's Microsoft 365 data hygiene — it surfaces documents, emails, and Teams conversations that are accessible to the user, which means poorly organized SharePoint sites and unclassified document libraries produce low-quality responses. Copilot is a tool that amplifies whatever information architecture the institution already has.

Google Workspace AI integrates Gemini models into Gmail, Docs, Sheets, and Meet for institutions on Google Workspace. Google offers a BAA for covered entities and expressly excludes customer data from training for Workspace customers. Like Copilot, it inherits the existing identity infrastructure. Its strongest use cases in AMC business operations are in drafting and document summarization.

ChatGPT Enterprise gives users access to GPT-4 models within an isolated organizational tenant, with a commitment that data is not used for OpenAI model training. It does not currently offer a HIPAA BAA, which limits its use to data that does not constitute protected health information. For many business-operations tasks this is not a constraint — job descriptions, vendor correspondence, and internal policy drafts rarely contain PHI — but revenue cycle and HR use cases that might touch patient data require a compliant alternative.

Claude for Enterprise (Anthropic) offers similar data handling commitments and organizational controls. It has been adopted at a smaller number of healthcare institutions, but its comparative strength in long-context tasks (reading and summarizing long contracts or policy documents) makes it worth evaluating for legal and supply chain applications specifically.

Table 9.2 compares the platforms on the dimensions that matter most for an AMC deployment decision.

Table 9.2: Enterprise AI platform comparison on dimensions relevant to AMC deployment. BAA availability has expanded since 2023; verify current BAA scope, data residency options, and sub-processor disclosures directly with each vendor before deployment involving PHI.

Platform	HIPAA BAA	Data residency	No training on customer data	Identity integration	Best fit
Microsoft 365 Copilot	Yes	US available	Yes	Entra ID (SAML/OIDC)	M365-native AMCs
Google Workspace AI	Yes	US available	Yes (Workspace customers)	Google Workspace SSO	Google Workspace AMCs

Platform	HIPAA BAA	Data residency	No training on customer data	Identity integration	Best fit
ChatGPT Enterprise (OpenAI)	Available (2024+)	US	Yes	SAML SSO	PHI-containing workflows with signed BAA
Claude for Enterprise (Anthropic)	Available	US	Yes	SAML/OIDC	Long-context tasks; contract review; policy Q&A

The buy-versus-build question in business operations has a clearer answer than in clinical or research domains. Building a custom large language model for revenue cycle or HR tasks is not a reasonable investment for any AMC. Even fine-tuning an open-weight model on institutional data requires infrastructure, ML engineering capacity, and an ongoing maintenance commitment that very few AMC IT organizations have. The right question is not build-versus-buy, but which enterprise platform to configure and how deeply to integrate it into existing workflows. The exception is retrieval-augmented generation (RAG) over internal document repositories — this is closer to configuration than model development, it runs on commodity infrastructure, and it produces meaningful value for internal policy Q&A, contract search, and HR knowledge bases.

9.6 Governance as Procurement

The NIST AI Risk Management Framework (National Institute of Standards and Technology 2023) organizes AI risk management into four functions: Govern (establish policies and roles), Map (identify and categorize risks), Measure (analyze and assess), and Manage (prioritize and treat). In the context of business operations AI procurement, these functions translate to a procurement process that is itself the governance mechanism.

The key insight is that most AMCs already have vendor risk management processes: security reviews, legal review of contract terms, privacy impact assessments for tools handling patient data. The task is not to build new infrastructure from scratch but to add AI-specific steps to existing workflows. The additions are modest: a bias audit requirement for any tool used

in employment screening, a PHI-exposure assessment for any tool handling revenue cycle or HR data, a data-training prohibition clause for any AI vendor agreement, and an audit log requirement for all enterprise AI tool usage.

Figure 9.1 illustrates a workable procurement flow for business-operations AI.

A few principles govern this flow. First, the preference for existing enterprise platforms over new vendors is explicit. The institution already has a Microsoft or Google tenant, already has SSO configured, already has data handling agreements in place. Using those tenants for a new use case is not a procurement event; it is a configuration decision. The governance overhead is lower, the data handling is already contracted, and the identity integration is already done. Second, any tool used in employment decisions requires a bias audit regardless of where in the institution it is used and regardless of whether the vendor claims their product is "bias-free." Third, every deployment ends with a defined success metric and a quarterly review cadence — not because every business-operations AI tool is high-stakes, but because without review, tools that are not working accumulate rather than being replaced.

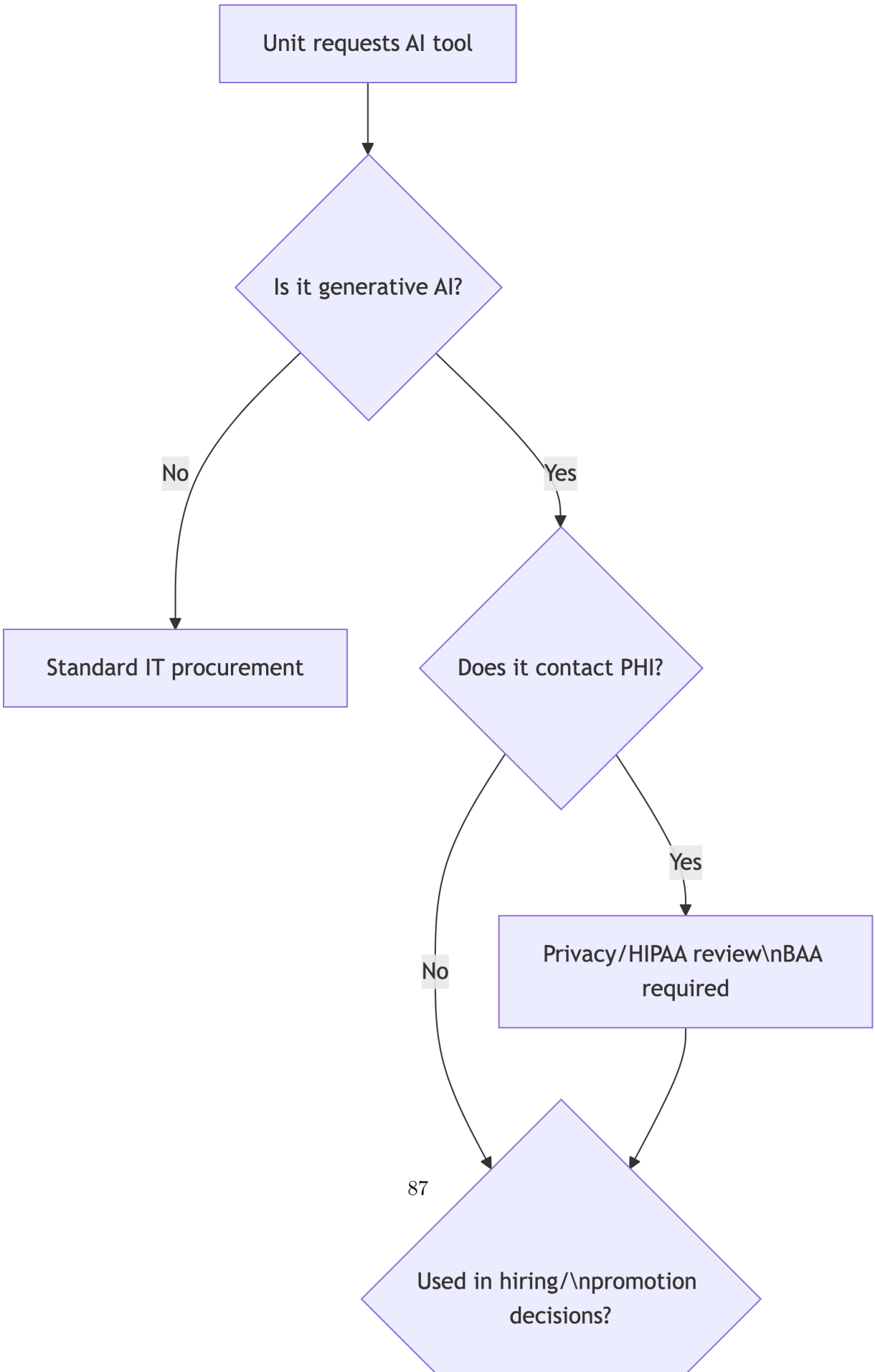
The ISO/IEC 42001 Artificial Intelligence Management System standard (ISO/IEC 42001 2023) provides an international governance framework for AI management that complements the NIST AI RMF. Institutions seeking to demonstrate AI governance maturity to partners, accreditors, or regulators will find that a 42001-aligned management system maps well to NIST and requires no duplication of effort.

9.7 What Success Looks Like

Measuring the value of business-operations AI is harder than it sounds, not because the tools produce no value but because the value shows up in forms that are difficult to isolate. Time savings are rarely captured systematically; productivity improvements appear as caseload increases, not headcount reductions; quality improvements in documents produced are hard to quantify. The AMC that waits for a clean ROI calculation before deploying AI will still be waiting when its administrative staff have long since found their own tools.

A workable measurement approach focuses on three levels. At the output level, track task-level time-to-completion for the specific workflows where AI is deployed — time from ticket receipt to denial letter, time from job requisition to posted description, time from contract receipt to first redline. At the quality level, track human review override rates (what fraction of AI outputs are substantially changed by the reviewing human) and error rates. At the compliance level, track the number of unauthorized AI tool uses detected and the trend over time.

None of these metrics requires a formal ROI model. Each requires only that the deployment was scoped, that baseline measurements were taken before AI was introduced, and that someone is responsible for tracking the numbers after deployment. The institutional discipline to take the baseline measurement is, in practice, the hardest part.


Unit requests AI tool
Is it generative AI?
No
Yes
Standard IT procurement
Does it contact PHI?
Yes
No
Privacy/HIPAA review\nBAA required
Used in hiring/\npromotion decisions?

9.8 Where to Start: Two Starter Projects

The preceding sections describe a governance framework and a set of use cases. This section describes what to actually do in the next ninety days. The goal is not to deploy everything — it is to deploy something well, learn from it, and build institutional capability that the next deployment can inherit.

9.8.1 Project 1: Enterprise AI Gateway Deployment

What it is. Deploy the AI capabilities already included in your existing Microsoft 365 or Google Workspace contract to a defined group of business-operations users — a pilot cohort of 20–50 people across finance, HR, and communications. Configure it within the existing tenant (no new vendor, no new security review), establish a simple usage policy, enable audit logging, and run for 90 days.

What you need to start. A Microsoft 365 E3/E5 or Google Workspace Business license already in effect, an IT administrator with tenant configuration access, a data classification policy that tells users what data they can put into the tool, and a willing domain lead in one of the three pilot functions. Nothing else is required.

What done looks like. After 90 days: audit logs are running, the pilot cohort has a shared understanding of what the tool can and cannot be used for, at least three specific workflow improvements have been documented with before/after time measurements, and no unauthorized PHI events have occurred. The 90-day report goes to the AISC with a recommendation on whether to expand.

Build or buy? Configure. This is not a procurement event. It is turning on a capability already licensed and paying for a governance and usage framework around it.

9.8.2 Project 2: Internal Policy Q&A Chatbot (RAG over Internal Documents)

What it is. Build a retrieval-augmented generation system over a defined corpus of institutional policies — HR policies, revenue cycle coding guidelines, facilities maintenance procedures, or any high-traffic internal knowledge base. Users ask questions in natural language; the system retrieves the relevant policy text and generates an answer, with the source document cited.

Why this one. Policy Q&A is high volume, low stakes, and naturally auditable. The current alternative — searching a SharePoint site or calling HR — is time-consuming and produces inconsistent answers. The AI system can be evaluated straightforwardly: does it retrieve the right policy section, and does its answer match what the policy actually says?

What you need to start. A curated set of no more than 50–100 policy documents in a shared drive, access to an embedding API (Azure OpenAI embeddings or Google Vertex AI

embeddings are available within the existing enterprise tenant), a vector store (pgvector on an existing PostgreSQL instance, or a cloud-native option), and one developer with two to four weeks of focused time. No specialized ML expertise is required; RAG over a small document corpus is well within the capability of a competent generalist developer following published patterns.

Build or buy? Build. This is one case in business operations where building is clearly preferable: the corpus is institution-specific, the commercial alternatives for internal policy Q&A are expensive, and the technical complexity is low. A developer who builds this system gains skills and understanding that transfer directly to the next RAG project — research literature summarization, clinical protocol Q&A, EHR documentation search.

What done looks like. A working system that answers 80% of test questions correctly against the policy corpus, with cited sources, running at acceptable latency. A defined process for adding new policies and reindexing. Usage logs showing which questions are being asked most frequently (which is itself valuable institutional intelligence about where policy documentation is unclear).

10 Regulatory and Policy Landscape

An AMC operating in 2026 does not face a single AI regulatory framework. It faces a patchwork of federal transparency mandates, state anti-discrimination laws, professional society accreditation standards, and international requirements for any institution with global research partnerships — each with different effective dates, different enforcement mechanisms, and different definitions of which AI tools they cover. The temptation is to treat this as a compliance problem: assemble a checklist, check each box, and move on. That approach will fail, because the regulatory landscape is not static. Two major federal rules took effect in early 2025, two more significant state laws take effect in 2026, and the EU AI Act's high-risk provisions come into force at the same time. Compliance today does not guarantee compliance in eighteen months.

The more durable approach is to treat AI governance as a risk management function that responds to a shifting grid rather than a fixed list. This chapter gives AMC legal, compliance, and informatics leaders the current state of that grid, with emphasis on the provisions that most directly affect clinical and research operations.

10.1 The Federal Regulatory Baseline

The foundational federal regulatory layer consists of rules from four agencies: ONC[1], FDA, CMS, and HHS OCR. Each has moved from guidance to enforcement-capable rule since 2024.

The ONC HTI-1 rule, published January 2024 with key provisions effective January 2025, is the most operationally significant for clinical informatics teams (Office of the National Coordinator for Health Information Technology 2024). It creates a new regulatory category — Decision Support Interventions — that covers EHR-based predictive algorithms meeting specified criteria for automated clinical guidance. For each qualifying tool, certified EHR vendors must make accessible a structured set of source attributes: training data sources, performance characteristics on the populations validated, known limitations, and update history. The rule does not require AMCs to build new infrastructure; it requires AMCs to demand that their EHR vendors fulfill their existing compliance obligations, and to incorporate the provided source attributes into their ongoing governance processes.

[1] https://www.healthit.gov

The FDA's Predetermined Change Control Plan guidance, finalized December 2024, provides the regulatory pathway for adaptive AI medical devices — systems that update their parameters based on new data after initial clearance or approval (U.S. Food and Drug Administration 2024). Traditional device regulation assumes a static design; a device that functions differently after deployment requires a new submission. The PCCP pathway allows a developer to specify in advance the types and bounds of permitted changes, the performance criteria that must be met before changes are implemented, and the monitoring required to detect unintended effects. AMCs that have developed or licensed AI-enabled SaMD (Software as a Medical Device) and intend to update those tools over time should assess whether a PCCP is the appropriate regulatory pathway.

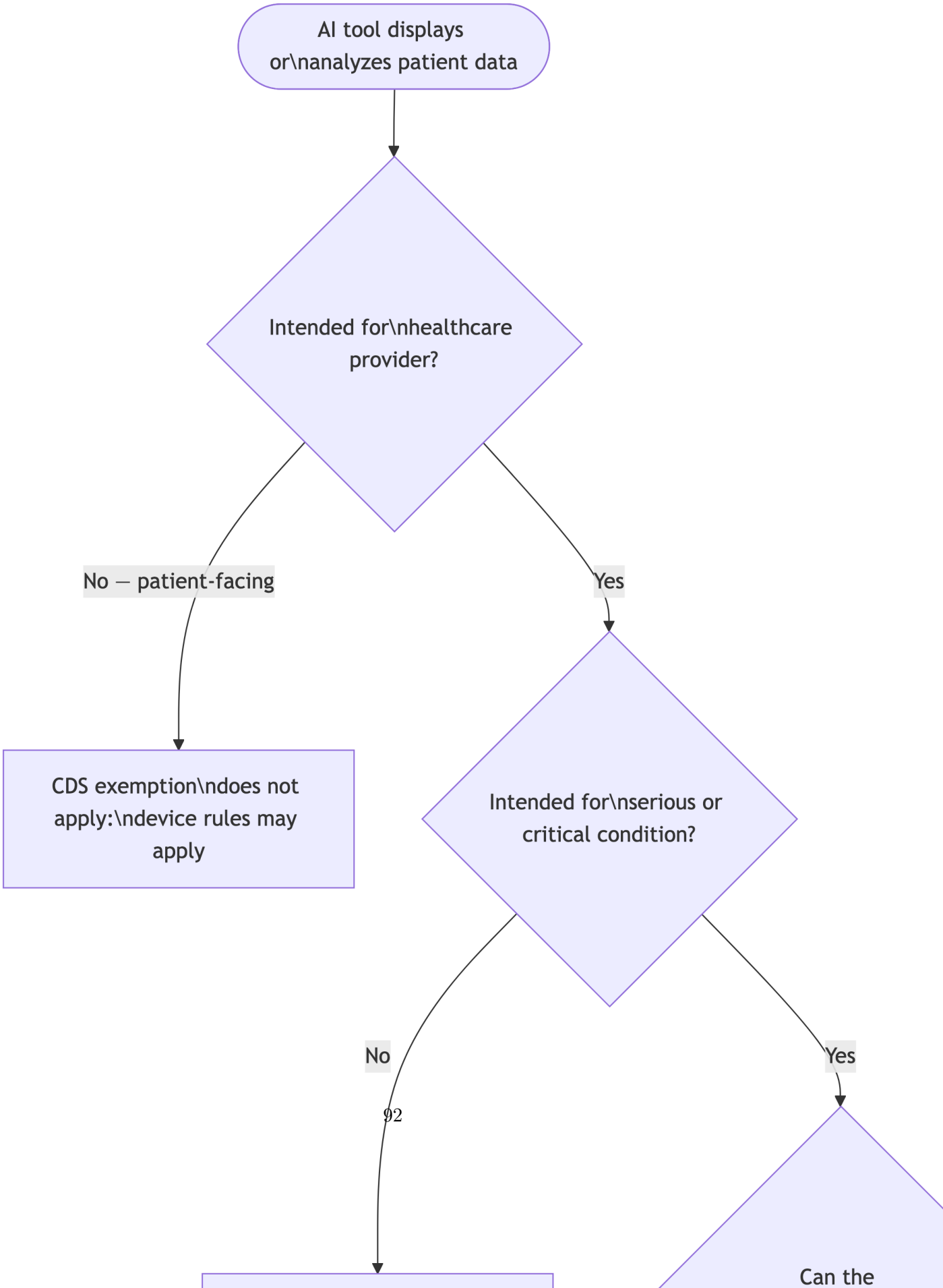
HHS OCR's Section 1557 final rule, effective May 2025, extended the nondiscrimination provisions of the Affordable Care Act explicitly to clinical AI (U.S. Department of Health and Human Services, Office for Civil Rights 2024b). Covered entities — including most AMCs — may not use patient care decision support tools that result in discriminatory treatment based on race, color, national origin, sex, age, or disability. The rule does not define which algorithms are covered with surgical precision; it places the burden on covered entities to assess whether tools they use could produce discriminatory outputs and to maintain documentation of that assessment. Institutions that have not audited deployed clinical AI for demographic performance parity are not in compliance with the spirit of the rule.

The CMS Medicare Advantage 2025 final rule addressed a narrower but important point: AI outputs may not serve as the sole basis for coverage denials under Medicare Advantage (Centers for Medicare and Medicaid Services 2024b). The rule responds to documented cases of insurers using AI to systematically deny claims with minimal human review, and it establishes that AI cannot substitute for the human clinical judgment required under existing utilization review standards.

10.2 The Executive Pivot: Federal Deregulation and State Divergence

The Biden administration's Executive Order 14110 (October 2023) directed federal agencies to develop AI risk assessments, transparency requirements, and accountability frameworks for government AI use — and signaled an expansive federal regulatory posture. The Trump administration revoked EO 14110 in January 2025, replacing it with Executive Order 14179, which directs agencies to remove regulatory barriers to AI development and prioritize American AI leadership over risk-based oversight frameworks (Executive Office of the President 2025).

For AMCs, the practical implications are specific rather than abstract. Federal agencies with AI oversight roles have received signals to reduce enforcement activity and defer to industry self-governance where possible. The rules already in effect — ONC HTI-1, FDA PCCP, HHS 1557 — remain law and do not require executive guidance to enforce. But the regulatory


AI tool displays or\nanalyzes patient data
Intended for\nhealthcare provider?
No — patient-facing
Yes
CDS exemption\ndoes not apply:\ndevice rules may apply
Intended for\nserious or critical condition?
No
Yes
Can the

pipeline of proposed rules that would have expanded algorithmic accountability requirements has slowed significantly.

The consequence is a growing divergence between federal deregulation and state-level legislative activity. Colorado and California have enacted, and other states are considering, AI transparency and accountability requirements that go beyond the current federal floor. For national AMCs serving patients across multiple states, the practical compliance standard is increasingly set by the most demanding state in which they operate.

10.3 The State Legislative Wave

Colorado Senate Bill 24-205 — the Colorado Artificial Intelligence Act — is the most comprehensive U.S. state AI law currently in effect, with key provisions taking effect June 30, 2026 (SB 24-205 2024). It covers "high-risk AI systems" used in consequential decisions, a category that includes healthcare. Covered entities — AMCs included — must conduct and document annual impact assessments of their high-risk AI tools, provide patients a meaningful notice of AI use in consequential decisions, and establish a process for patients to appeal or seek correction of AI-influenced decisions. The law places the compliance burden on the "deployer" — the entity that uses the AI tool in its operations — not solely on the developer.

California's regulatory posture is more fragmented but cumulatively significant. AB 3030, effective January 2025, requires disclosure on AI-generated patient communications (AB 3030 2024). The FTC's Operation AI Comply enforcement actions in 2024 targeted deceptive AI claims in healthcare marketing and patient-facing tools (Federal Trade Commission 2024). New York City's Local Law 144 requires independent bias audits for AI-assisted hiring decisions — a provision directly relevant to AMCs using algorithmic screening for clinical staff recruitment (Local Law 144 2021).

The compliance challenge for national AMCs is that these laws have different definitions of covered AI, different disclosure requirements, and different enforcement mechanisms. The recommended institutional posture is to adopt the most demanding applicable standard as the default — which in most cases means Colorado's annual impact assessment and California's patient disclosure requirements — rather than attempting to maintain state-specific compliance workflows.

Table 10.1: Key regulatory milestones for AMC AI governance, 2024–2026 (Office of the National Coordinator for Health Information Technology 2024; U.S. Food and Drug Administration 2024; U.S. Department of Health and Human Services, Office for Civil Rights 2024b; Centers for Medicare and Medicaid Services 2024b; SB 24-205 2024, AB 3030 2024, Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (Artificial Intelligence Act) 2024)

Agency / Body	Rule / Law	Effective Date	Key Provision for AMCs
ONC	HTI-1 Algorithm Transparency	January 2025	EHR vendors must surface DSI source attributes within clinical workflow
FDA	PCCP Guidance	December 2024	Adaptive AI-SaMD may update within pre-specified bounds without new filing
HHS OCR	Section 1557 Final Rule	May 2025	Covered entities may not use discriminatory patient care decision-support tools
CMS	Medicare Advantage 2025 Rule	January 2025	AI alone cannot be basis for coverage denial; human review required
Colorado	SB 24-205 (CAIA)	June 2026	Annual impact assessments; patient notice and appeal rights for high-risk AI
California	AB 3030	January 2025	Disclosure on AI-generated patient communications
EU	AI Act (high-risk provisions)	August 2026	High-risk systems in clinical use require conformity assessment and human oversight

10.4 Professional Sovereignty and Accreditation Standards

Alongside the legislative layer, professional societies and accreditation bodies have moved to codify AI governance expectations in standards that carry their own enforcement mechanisms — credentialing, accreditation, and membership. For AMCs, these standards often have more immediate operational effect than distant federal rules, because their consequences for day-to-day operations are more direct.

The NIST AI Risk Management Framework[2] (National Institute of Standards and Technology 2023) has become the de facto organizational scaffold for institutional AI governance in

[2] https://www.nist.gov/itl/ai-risk-management-framework

U.S. healthcare. Its Govern/Map/Measure/Manage functions provide the structure for AI program design, and multiple state laws and professional society standards have incorporated its terminology. The companion Generative AI Profile (National Institute of Standards and Technology 2024) extends the framework to address risks specific to large language models — hallucination, data memorization, and bias amplification — that the original framework did not fully anticipate.

The AMA's policy on augmented intelligence (American Medical Association 2024a) asserts that physicians must be able to interpret and act upon AI outputs, and advocates for independent verification of developer performance claims. The policy's practical significance is in its framing of the physician relationship to AI tools: the physician remains the final arbiter of clinical decisions and bears professional accountability for actions taken with or without AI input. This is not merely an ethical position; it maps directly to the liability landscape discussed below.

The ISO/IEC 42001:2023 standard for AI management systems (ISO/IEC 42001 2023) provides a certification pathway for institutions that want to demonstrate systematic AI governance to regulators, accreditors, and payers. It is the closest analogue in AI to ISO 27001 for information security — a voluntary standard that is increasingly required or expected by procurement and regulatory processes. AMCs that have built NIST RMF-aligned governance programs are well positioned to seek ISO 42001 certification with targeted additional work.

10.5 International Requirements for Global Research Partners

For AMCs with international research partnerships, clinical trial enrollment in Europe, or data sharing with EU-based institutions, the EU AI Act[3] represents a material compliance obligation. The Act's high-risk provisions — applicable from August 2026 — cover AI systems used in clinical decision support, medical devices, and management of critical infrastructure (Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (Artificial Intelligence Act) 2024). High-risk systems must undergo a conformity assessment before deployment, maintain technical documentation, implement human oversight measures, and register in the EU AI database.

The practical implication for AMC research operations is that any AI tool used in a clinical study enrolling EU subjects, or any EHR-integrated AI system at an EU affiliate, must be assessed against EU AI Act requirements before August 2026. AMC general counsel should treat this as a research contracts and technology transfer issue, not solely an IT compliance matter.

The "Brussels Effect" — the tendency for EU regulation to become a de facto global standard because multinational companies find it easier to apply the highest standard uniformly — may accelerate EU AI Act adoption even among AMCs without current European operations.

[3] https://artificialintelligenceact.eu/

Vendors seeking to maintain EU market access will increasingly build their products to EU standards, and AMCs that adopt those products will inherit the compliance posture.

10.6 Where to Start

10.6.1 Starter Project 1: AI Regulatory Compliance Mapping

What it is: A mapping of the institution's current AI tool inventory (see Section 6.4) against the applicable regulatory frameworks in Table 10.1, identifying which tools are covered by which rules and where compliance gaps exist.

Why now: Several rules are already in effect; others take effect in 2025 and 2026. An institution that has not completed this mapping cannot certify compliance to its board, its accreditors, or its patients.

How to execute: Use the clinical AI inventory as the starting point. For each tool, determine: Does it qualify as a DSI under HTI-1? Does it qualify as SaMD under FDA regulations? Is it used in decisions that qualify as "consequential" under Colorado SB 24-205? Does it generate patient communications covered by California AB 3030? The output is a compliance matrix that the legal and compliance teams can use to prioritize remediation.

Buy vs. build: Legal analysis and governance work. Some commercial GRC (Governance, Risk, and Compliance) platforms have begun adding AI regulatory mapping capabilities, but the analysis itself requires legal judgment that cannot be fully automated.

10.6.2 Starter Project 2: Annual AI Governance Report to the Board

What it is: A structured annual report to the institutional board of trustees on the state of the AI governance program — deployed tools, regulatory compliance status, adverse events, and strategic priorities.

Why now: Colorado SB 24-205 requires annual impact assessments for high-risk AI. Beyond the legal requirement, board-level visibility into AI governance is increasingly expected by accreditors and institutional investors. An AMC that cannot report coherently to its board on its AI risk posture is behind the governance standard the field is converging toward.

How to execute: Define a standard reporting template that includes: inventory of deployed AI tools with risk tiers, compliance status against applicable regulations, any adverse events involving AI in the reporting period, performance monitoring findings, and planned additions or retirements. Present to the board annually, with quarterly updates to the relevant board committee if the portfolio is large. The report format should be adapted from the NIST AI RMF Govern function documentation requirements.

Part III

Agentic AI in Practice

11 Agentic Safety and Guardrails

There is a meaningful safety boundary between an AI system that suggests a course of action and one that executes it. On one side of that boundary, a clinician reads the output and decides what to do next; the human is the actor, and the AI is a source of input. On the other side, the AI plans and executes a sequence of steps — placing a message in a patient's portal, submitting a prior authorization request, updating a medication list — while a human nominally supervises but does not approve each individual action. Most AMC governance frameworks were designed for advisory AI. They ask how to validate a model's output and how to surface it appropriately in the clinical workflow. They were not designed to govern systems that initiate actions in the world.

That gap matters now because the agentic threshold is being crossed — not in research settings, but in production deployments. Epic's "Chart with Art" can autonomously query clinical data and prepopulate order sets. Prior authorization agents submit benefit determinations to payers without manual triggering. AI-enabled inbox management tools route, draft, and in some configurations send patient responses. The safety failures that arise in these systems are qualitatively different from those that arise when a clinician misreads an AI recommendation. They are faster, harder to trace, and more likely to cascade into downstream harm before anyone notices. This chapter gives AMC clinical and operational leaders the framework for governing these systems before the first incident, not after.

11.1 From Advisory to Agentic: The Autonomy Spectrum

It is more useful to think of AI autonomy as a spectrum than as a binary distinction. At the advisory end, the AI generates text and stops — it drafts a clinical note, and the clinician reviews every word before attesting. The human is the sole actor in any consequential sense. Moving along the spectrum, systems begin to initiate actions without per-step human approval: routing a message to a triage queue, flagging a result for follow-up, prefilling a prior authorization form. At the far end, systems take actions that write to authoritative records — the medication list, the order queue, the problem list — based on autonomous reasoning about what the clinical situation requires.

The critical observation in Figure 11.1 is that risk is not monotonically increasing along the spectrum. A fully agentic system that writes to the medication list is obviously high-risk. But an autonomous prior authorization agent operating at the "autonomous administrative"

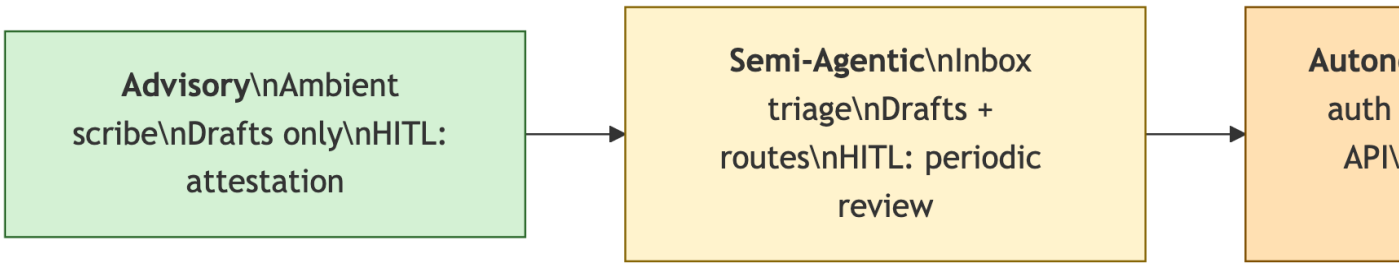


Figure 11.1: The advisory-to-agentic spectrum with current deployed examples and required oversight tier at each level. HITL = human-in-the-loop (approves each action). HOTL = human-on-the-loop (reviews logs after the fact).

level also carries substantial risk: it interprets clinical data, matches it to insurance criteria, and submits a determination on behalf of the institution — all without a clinician reviewing the specific case. When the agent makes an error, the error is already in an external system before anyone looks at it. The delay between the action and its discovery is the relevant safety variable, and it scales with autonomy.

11.2 Agentic Failure Modes

Advisory AI fails in familiar ways: a model produces an incorrect output, the clinician reads it, and ideally the clinician catches the error. The human is the circuit breaker. Agentic AI introduces failure modes where the circuit breaker either is not in position to act, or has been trained by experience to defer.

Cascading errors. In a multi-agent architecture — a diagnostic agent feeds a recommendation to a scheduling agent, which triggers a prior authorization agent — an error in the first agent propagates through the chain. Each downstream agent treats its input as authoritative, because that is how the system was designed. By the time a human reviews the output, the consequences of the initial error may have ramified across multiple systems and cannot be easily undone.

Scope creep. An agent authorized to perform task A may, when the instructions are ambiguous, take action on task B because it appears adjacent or necessary. An inbox triage agent authorized to route messages may, in the course of routing, infer that a result requires follow-up and draft an outreach message the clinician never requested. The agent is not malfunctioning; it is being helpful in a way the governance design did not anticipate.

Action on stale data. An agent's knowledge of the patient context — diagnosis, current medications, recent labs — is as fresh as its last synchronization with the EHR. If that synchronization is minutes or hours old, and the clinical situation has changed in the interval, the agent may take actions that were appropriate for a patient who no longer exists. This is not

a theoretical risk; it is a timing problem that arises whenever an agent operates asynchronously against a live clinical record.

Automation surprise. The clinician discovers that something happened — a message was sent, a form was submitted, a list was updated — without their explicit direction. The surprise itself is a safety event, independent of whether the action was correct. A clinician who does not understand what actions an agent may take cannot be a reliable overseer of those actions (Parasuraman and Manzey 2010). Automation surprise is the signal that the human oversight model was not designed correctly for the system's actual autonomy level.

Sycophancy. Research on agentic AI in reasoning tasks has documented a tendency for models to converge on the perspective of the most recent authoritative input rather than reason independently. In clinical terms, an agent asked to verify a clinician's tentative diagnosis may confirm that diagnosis because the clinician proposed it — not because the evidence supports it. This is particularly dangerous in contexts where the agent's role is explicitly supervisory, such as medication reconciliation or prior authorization review.

11.3 Human-in-the-Loop Architecture

"Human-in-the-loop" has become a compliance phrase in healthcare AI governance. It appears in policies, vendor contracts, and regulatory guidance as a shorthand for adequate oversight. The problem is that it describes an architectural design choice without specifying what that choice requires. A human who clicks "approve" on an AI-generated order in under five seconds is technically in the loop; they are not, in any meaningful sense, exercising oversight.

The automation bias literature is unambiguous on this point. Operators consistently over-trust automated systems, especially under cognitive load, and consistently under-correct for errors that the system presents in a confident, well-formatted output (Parasuraman and Manzey 2010). The more accurate the system generally is, the worse this effect becomes: a clinician who has approved 200 consecutive AI recommendations without incident will review the 201st less carefully than the first. This is not a character failure; it is a predictable consequence of human attention operating on probabilistic feedback.

The governance implication is that the question is not "is there a human in the loop?" but "is the HITL checkpoint likely to catch errors, given the cognitive conditions under which it operates?" A high-stakes HITL checkpoint — a human reviewing an autonomous medication order — should be designed for scrutiny, not for speed. That means surfacing the specific clinical reasoning the agent used, not just the recommended action; setting an expected review time that is long enough to be meaningful; and tracking override rates as a quality metric. A near-zero override rate is not a sign that the AI is performing well; it may be a sign that the checkpoint has become a rubber stamp.

For lower-stakes actions, a human-on-the-loop (HOTL) architecture may be appropriate: the agent acts, and the human reviews a log of actions after the fact. The distinction is consequential.

HOTL is only acceptable when the actions are readily reversible, the volume is too high for per-action review, and the exception-detection mechanism is robust enough to surface anomalous actions before they cause harm.

11.4 Kill Switches and Circuit Breakers

Every agentic AI system deployed in a clinical environment requires a mechanism to pause or terminate its operation in response to anomalous behavior, and that mechanism must be tested before deployment, not designed after an incident.

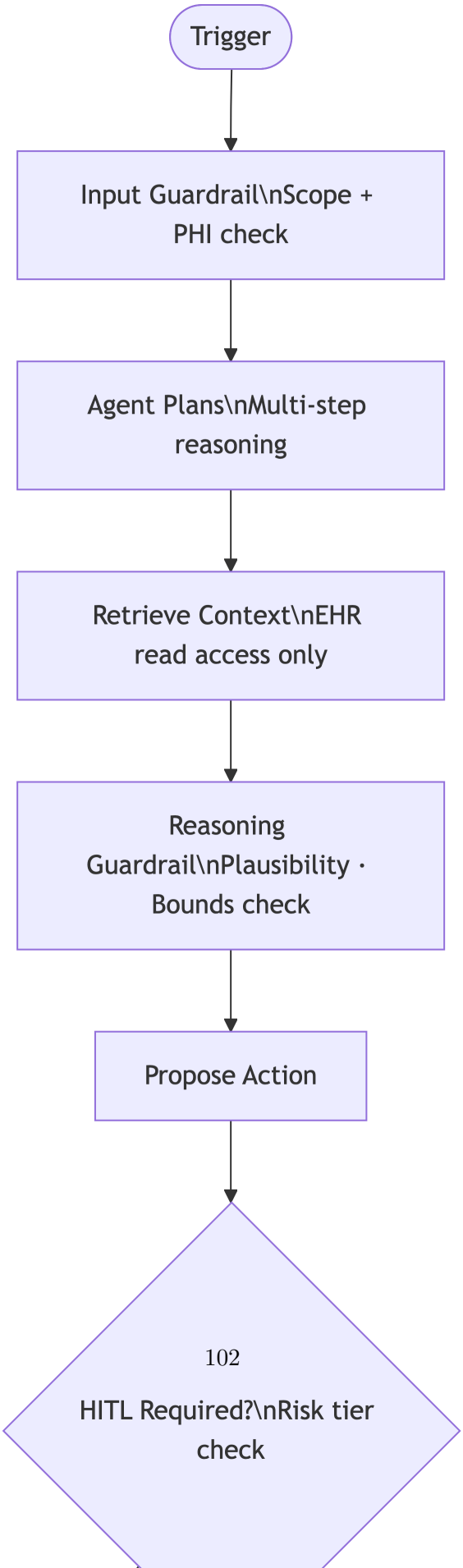
The aviation analogy is instructive. Aircraft autopilot systems do not operate without limits; they operate within a flight envelope, and they disengage automatically when the aircraft departs that envelope. The pilot has a manual disconnect that takes effect immediately. Neither feature requires the pilot to monitor every output of the autopilot to remain safe; the system enforces its own limits and returns control explicitly when those limits are exceeded.

Clinical AI governance needs the same architectural discipline. For each deployed agentic system, the governance design should specify: (1) the conditions under which the system will automatically pause, (2) who has authority to manually pause and resume, (3) what the fallback workflow is while the system is paused, and (4) how often the kill switch is tested. An untested circuit breaker is not a safety feature.

The specific trigger conditions depend on the system. An inbox management agent might pause automatically if more than a certain percentage of messages in a defined time window are routed to an unusual destination — an indicator that something in the routing logic has drifted or been compromised. A prior authorization agent might pause if the approval rate for a specific payer departs significantly from the institutional baseline. A medication reconciliation agent might pause if it encounters a clinical context it has not seen in training. The triggers should be specified by the clinical team and the informatics team together, not left to the vendor.

11.5 Least Privilege and Scope Limitation

The principle of least privilege — a system should have access only to the resources it needs to perform its specified function, and no more — is a foundational concept in information security that applies with particular force to clinical AI agents. An agent that needs to read a patient's problem list and draft a prior authorization request does not need write access to the medication list. An agent that needs to route inbox messages does not need access to the full longitudinal record.


Trigger
Input Guardrail\nScope + PHI check
Agent Plans\nMulti-step reasoning
Retrieve Context\nEHR read access only
Reasoning Guardrail\nPlausibility · Bounds check
Propose Action
HITL Required?\nRisk tier check

The technical mechanism for enforcing least-privilege access in the EHR context is the SMART on FHIR[1] authorization framework, which supports granular, parameterized access scopes. Rather than granting an agent broad access to a patient record, the institution can specify that the agent may read specific resource types (e.g., Observation, Condition) for specific patients, during specific session windows. The agent cannot read or write outside that scope, because the authorization framework physically prevents it.

Operationally, this requires a change in how clinical AI tools are procured and configured. Many commercial tools request broad EHR access during initial setup because it simplifies integration. AMC informatics teams should treat this as a negotiating point in the procurement process: what is the minimum access scope required for this tool to function as advertised? Any access beyond that minimum is a risk surface that the institution is accepting without commensurate benefit.

Audit logging for agent actions should be as granular as audit logging for human actions. If a clinician querying the EHR generates a log entry, an agent querying the EHR should generate an equivalent log entry. The log should capture: which agent, on behalf of which patient, accessed which resource, at what time, and what action resulted. This is not primarily a compliance requirement; it is the mechanism by which automation surprises are investigated and understood.

11.6 The Regulatory Picture for Agentic AI

The regulatory landscape for agentic clinical AI is unsettled, and AMC leaders should approach it with the same expectation of ambiguity that currently governs AI medical devices generally. Several regulatory frameworks bear directly on agentic systems, even where they do not address agenticity explicitly.

The FDA's Predetermined Change Control Plan guidance, discussed in Chapter 6, applies with special force to agentic systems that learn and adapt as they operate. A prior authorization agent that updates its matching criteria based on approval outcomes is an adaptive system; if it meets the Software as a Medical Device definition, it requires either a PCCP or a new regulatory filing for each adaptation (U.S. Food and Drug Administration 2024). The PCCP pathway is available but requires pre-specifying the bounds of the adaptation — which is precisely the kind of disciplined governance that agentic systems require in any case.

The CMS Interoperability and Prior Authorization Final Rule (CMS-0057-F) mandates that payers respond to prior authorization requests within 72 hours for urgent requests and seven days for standard requests (Centers for Medicare and Medicaid Services 2024a). This regulatory pressure is a primary driver of agentic PA deployment: health systems need to submit faster, and automation is the only technically feasible path at volume. It does not, however, change the safety requirements for those systems; it increases the operational pressure to skip them.

[1] https://smarthealthit.org

California AB 3030's disclosure requirement for AI-generated patient communications (AB 3030 2024) becomes more complex in agentic contexts. When a human clinician reviews and attests to an AI-drafted message, the disclosure is straightforward. When an agent routes and sends a communication without clinician review, the disclosure architecture requires more thought: what is being disclosed, to whom, and when? AMCs deploying agentic communication tools should consult legal counsel on whether AB 3030 and similar state laws require the disclosure to appear before the communication is sent, not after.

The NIST AI Risk Management Framework's Govern/Map/Measure/Manage structure (National Institute of Standards and Technology 2024) provides the most practical scaffold for AMC agentic AI governance. Specifically, the Measure function — developing metrics for agentic AI performance, error rates, and override patterns — is where most institutions are currently underprepared. You cannot govern what you are not measuring.

11.7 The Action Risk Authorization Matrix

The table below provides a working framework for assigning governance controls to categories of agent action. The tiers are illustrative; specific institutions will need to calibrate thresholds based on their risk tolerance, patient population, and clinical context.

Table 11.1: Risk-level and authorization matrix for agentic AI actions in the clinical environment. Rows with "Required" HITL tiers should have override rate monitoring built into the governance framework.

Action Type	Authorization Model	Audit Logging	HITL Tier	Kill-Switch
Read EHR data	Agent identity + scoped token	Summary level	Not required	Not required
Route message to clinical queue	Per-session scope	Granular	Optional	Recommended
Draft patient communication	Per-session scope	Granular	Required: clinician attestation	Recommended
Submit prior authorization	Periodic institutional review	Granular	Optional for low-risk PA	Required
Update problem list or medication	Per-action explicit approval	Granular	Required: per-action	Required
Place clinical order	Per-action explicit approval	Granular	Required: per-action	Required

11.8 Where to Start

The governance gap for agentic AI is real, but it does not require a complete rebuild of existing frameworks. The two projects below are tractable starting points for institutions that have already begun deploying semi-agentic tools and need to formalize oversight before the autonomy level increases.

11.8.1 Starter Project 1: Agentic AI Audit and Tiering

What it is: An inventory and autonomy-tiering of every deployed AI tool that takes any action — routes, submits, updates, or schedules — without per-action human approval. For each tool, assign an autonomy tier using the spectrum in Figure 11.1, assess whether the current governance controls match the tier requirements in Table 11.1, and document the gaps.

Why now: Many institutions have deployed ambient scribes, inbox triage tools, and PA assistants without formally categorizing them as agentic systems. The governance gap is often not a deliberate decision; it is an artifact of tools being piloted and scaled before the governance framework caught up.

How to execute: This is an extension of the clinical AI inventory recommended in Section 6.4. The additional work is adding an autonomy tier column, auditing the access scopes currently granted to each tool (via the EHR's OAuth/SMART application registry), and reviewing audit log coverage. The output is a tiered register with documented gaps and a prioritized remediation plan.

Buy vs. build: Governance exercise. The tiering framework and audit process are institutional work; the only technology question is whether the EHR's application registry surfaces the information needed for an access scope audit.

11.8.2 Starter Project 2: Kill-Switch Design and Testing

What it is: For each agentic tool at the "autonomous administrative" tier or above, define the circuit-breaker trigger conditions, pause mechanism, fallback workflow, and test schedule. Run at least one tabletop exercise in which the kill switch is activated and the fallback workflow is executed.

Why now: Kill switches that have never been tested do not work reliably under pressure. The tabletop exercise surfaces both technical gaps (the pause mechanism does not propagate to all system components) and operational gaps (the fallback workflow requires a staff role that is not adequately covered on nights and weekends).

How to execute: Work with the vendor, the EHR team, and the clinical operations team to document the pause mechanism for each tool. Define trigger conditions with the clinical team

— what anomaly pattern, at what threshold, triggers a pause? Schedule a tabletop exercise with the on-call clinical informatics lead and relevant clinical leadership. Use the exercise findings to update the governance documentation before the next deployment expansion.

Buy vs. build: Design and process work, not a technology purchase. Most commercial vendors have a pause mechanism; the gap is typically in the institutional design of when and how to use it.

12 Patient and Community Trust

There is a tempting shorthand for the case for transparency in clinical AI: institutions disclose how AI is used in patient care because they are required to, or because it reduces liability exposure. Both motivations are real. Neither is sufficient to explain why transparency actually matters for the safety and effectiveness of care.

The more important reason is that patient trust is a clinical variable. A patient who distrusts an AI tool used in their care will engage with it differently than one who understands and accepts it. They may withhold information, seek care elsewhere, or disregard recommendations they believe were generated by a system optimized for cost reduction rather than their welfare. Trust does not just affect how patients feel about their care; it affects whether the care is effective. An AI system that produces accurate outputs but is deployed in a population that distrusts it will produce worse population-level outcomes than a less accurate system that has earned confidence.

This chapter addresses the social license for AI in healthcare — the degree to which patients and communities actually believe that an AMC is deploying AI on their behalf. It is distinct from the regulatory compliance chapter, which addresses what institutions are required to do, and from the agentic safety chapter (Chapter 11), which addresses how autonomous systems should be governed. Social license addresses a prior question: do the people served by this institution believe it should be deploying AI at all, and in what ways? Institutions that skip this question and proceed directly to deployment governance are building on ground they have not tested.

12.1 The Empirical Trust Landscape

The consistent finding across surveys of patient attitudes toward AI in healthcare is that trust is lower than many technology optimists assume, and more variable across demographic groups than most institutional communications acknowledge.

Pew Research[1] surveys on AI in healthcare have found that most U.S. adults are uncomfortable with their provider relying on AI for their medical care, and that the large majority would prefer a human provider when it comes to accuracy-sensitive decisions (Pew Research Center 2023). The discomfort is not evenly distributed: younger adults, higher-income adults, and

[1] https://www.pewresearch.org

those with more experience using digital health tools consistently report more comfort with AI-assisted care than older adults, lower-income adults, and those with less digital access. Insurance status, chronic condition burden, and prior negative experiences with the health system all independently predict lower trust in AI.

The type of AI application matters as much as the population. Patients who are skeptical of AI-assisted diagnosis are often more accepting of AI in administrative functions — scheduling, prescription refill processing, billing. The concern is concentrated in contexts where the patient perceives the AI as making clinical judgments that might otherwise be made by a physician with whom they have a relationship. Ambient documentation, in which an AI listens during a clinical encounter, occupies a particularly sensitive position: surveys consistently find that patients who receive an explanation of what the ambient system does and how the data is used are substantially more accepting than those who are not told.

There is also a gap between what patients know and what they prefer. Many patients do not know that AI tools are already embedded in their care — in the risk scores that determine their appointment priority, in the routing of their portal messages, in the sepsis alerts that flag their deterioration. Studies that ask patients how they would feel about AI use in care that they already receive without knowing it frequently find that awareness increases concern, at least initially. Institutions that interpret patient acceptance of current care as patient acceptance of AI in care are drawing an inference that the data does not support.

12.2 The Historical Roots of Differential Trust

Understanding why trust in healthcare AI is lower in Black, Latino, and Indigenous communities than in white communities requires engaging with history rather than treating differential trust as an artifact to be corrected through better communication. The medical research system has a documented record of using marginalized communities as subjects without genuine consent, without equitable distribution of benefits, and without recourse when harm resulted. The legacy of the Tuskegee syphilis study — in which Black men with syphilis were observed without treatment for decades after effective treatment became available — produced measurable multigenerational effects on participation in medical research and on trust in medical institutions in Black communities that persist decades later.

These are not historical curiosities that have been corrected by better ethical oversight. Contemporary algorithmic systems in healthcare have replicated structural inequities in ways that researchers have documented repeatedly. A widely cited study found that a commercial algorithm used to allocate healthcare resources to high-need patients systematically underestimated the needs of Black patients, because the algorithm used healthcare utilization as a proxy for health need — and Black patients with equivalent illness burden utilized less care, due to structural barriers in access. The algorithm was not explicitly programmed to be discriminatory; it was trained on data that encoded existing inequities, and it reproduced them at scale.

For AMC leaders deploying clinical AI, the practical implication is that the communities with the most to gain from AI-enabled improvements in care access and quality are also the communities with the best-documented reasons to distrust the institutions deploying it. Community engagement for AI governance in these communities is not a communication exercise. It is not "informing" the community about a decision that has already been made. It is bringing the community into the governance process before deployment decisions are finalized — which means starting earlier, accepting that community input may change deployment decisions, and being transparent about that possibility from the outset.

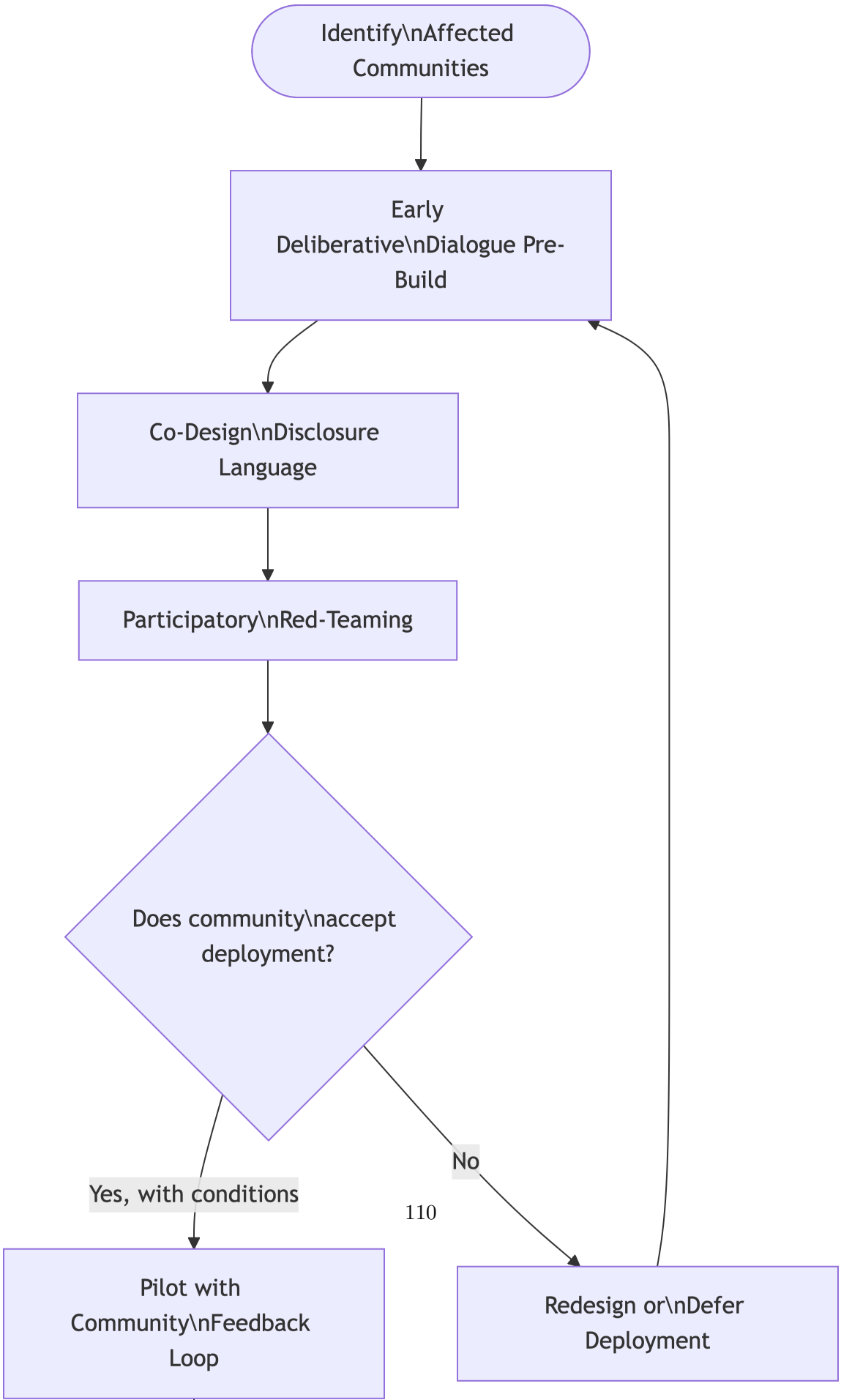
12.3 Meaningful Disclosure vs. Boilerplate

The regulatory minimum for AI disclosure — satisfying California AB 3030 (AB 3030 2024), for example — requires that patients be told when AI has been used to generate health communications. The regulatory minimum and the trust-building minimum are not the same thing.

Boilerplate disclosure — "This message was generated using AI technology" — satisfies the letter of the disclosure requirement. It does not tell the patient which AI system was used, whether a clinician reviewed the output, what the system is and is not capable of, or what the patient can do if they have concerns. Research on patient responses to AI disclosure consistently finds that boilerplate language increases concern without providing the context needed to resolve it. Patients who receive boilerplate disclosure are not better informed than those who received no disclosure; they are more anxious without a framework for understanding what their anxiety is about.

Meaningful disclosure has three components. It identifies what the AI did specifically (drafted this message, analyzed this image, flagged this result). It clarifies the human role (a physician reviewed this draft before sending; a radiologist reviewed the AI findings). And it provides an actionable recourse (if you have questions about how this was produced, you can ask your care team). None of this requires technical language or a long disclosure form. The most effective disclosure language found in pilot programs is specific and brief: "Dr. [Name] asked our AI system to draft this response based on your message and your recent visit. Dr. [Name] reviewed and approved it before sending."

The workflow in Figure 12.1 reflects a principle that is easy to state and difficult to practice: community engagement for AI governance is most valuable before deployment decisions are made, not after. An institution that presents a completed AI deployment to a community advisory group and asks for feedback is not practicing participatory governance; it is practicing consultation theater. The community members who participate in that process will know the difference, and the trust consequences will follow.


Identify\nAffected Communities
Early Deliberative\nDialogue Pre-Build
Co-Design\nDisclosure Language
Participatory\nRed-Teaming
Does community\naccept deployment?
Yes, with conditions
No
Pilot with Community\nFeedback Loop
Redesign or\nDefer Deployment

12.4 Informed Consent and Its Gaps

Existing patient consent frameworks in most AMCs were designed for interventions that are visible, discrete, and time-bounded: a surgery, a medication, an imaging procedure. Clinical AI tools do not fit this model. A predictive risk score that runs continuously in the background of the EHR, influencing which patients get proactive outreach and which do not, is not an intervention in the usual sense. There is no natural moment at which a patient is informed and asked to consent, because the tool is not doing anything to the patient directly at a specific moment. It is shaping institutional decisions about the patient in ways the patient may never observe.

Ambient documentation is the most visible gap. A patient who arrives for a clinic visit where an ambient scribe is running is being recorded and having their conversation synthesized by an AI system. They may or may not know this before the encounter begins. California AB 3030 requires disclosure in written communications; it does not address the consent architecture for the encounter itself. The institutions that have managed this most effectively have adopted a verbal consent practice — the clinician explains the ambient system at the start of the encounter, tells the patient they can ask for it to be turned off, and documents the consent in the visit record. This is not complicated; it is a thirty-second conversation that most patients appreciate.

Predictive risk scoring presents a different consent challenge. The patient does not know whether they scored high or low on the readmission risk model, or whether that score influenced which patients received a follow-up call and which did not. Whether patients should be told their AI risk scores is a question that different institutions have answered differently, and the ethics literature does not offer a clear consensus. What the literature does support is that patients who are told their scores and given an explanation of what the score means and what follows from it are more accepting of AI-influenced care than those who discover after the fact that such scoring was occurring (Jones et al. 2023).

12.5 Patient AI Advisory Councils

A small number of health systems have established Patient AI Advisory Councils — governance bodies distinct from standard patient advisory boards, with a specific mandate to advise on AI tool selection, deployment, and oversight. The Vanderbilt University Medical Center's ADVANCE Center created an AI Patient and Family Advisory Group that participates in "red-teaming" AI tools during the ideation phase — reviewing tools for potential biases and unintended consequences before they reach the pilot stage. The Digital Medicine Society[2] sets engagement with patients as active partners on AI governance bodies as the gold standard for AI program maturity.

[2] https://www.dimesociety.org

The structural distinction between a Patient AI Advisory Council and a standard patient advisory board matters. Standard advisory boards typically advise on service delivery and patient experience; their members are selected to represent satisfied patients. A Patient AI Advisory Council needs members who can engage with questions of algorithmic fairness, data use, and automated decision-making — which means it should deliberately recruit from communities with high stakes in those questions, including communities with documented reasons to distrust institutional AI. It should have access to technical staff who can explain how specific tools work, and it should have a defined relationship to the AI Steering Committee, including clarity about which decisions it is consulted on versus which decisions it has authority over.

Table 12.1: Structural comparison of a standard patient advisory board and a Patient AI Advisory Council. The key distinction is that the AI Advisory Council engages before deployment decisions are finalized.

Dimension	Standard Patient Advisory Board	Patient AI Advisory Council
Membership basis	Satisfied patient representatives	Diverse community members; includes historically marginalized voices
Subject matter	Service delivery, patient experience	AI tool selection, deployment, bias audits, disclosure language
Phase of engagement	Post-deployment feedback	Pre-deployment ideation and red-teaming
Relationship to AISC	Informational only	Formal consultative role; documented input on pilot approvals
Authority	Advisory	Consultative; specific veto scope may be defined for high-risk pilots
Meeting frequency	Quarterly	Monthly (active deployment periods); quarterly (steady state)

12.6 The Regulatory Landscape of Disclosure

The legislative landscape for AI disclosure in healthcare is still forming, but two state laws enacted in 2024 are now in effect and should be treated as the leading edge of a national trend.

California AB 3030, effective January 2025, requires healthcare providers to include explicit disclosure when AI has been used to generate patient communications, unless a licensed human reviews the content first (AB 3030 2024). The law is narrow in scope — it covers written communications, not real-time verbal interactions — but it is the most specific AI disclosure

mandate currently in effect at the state level, and AMCs with any California patient population must comply. Legal teams at national AMCs should treat the California standard as the floor for their disclosure practices regardless of the states in which they primarily operate.

Colorado Senate Bill 24-205, effective June 2026, goes further: it requires that any entity using AI in a "consequential decision" about health care provide a "pre-decision notice" that explains the purpose of the AI, the nature of the consequential decision, and how to appeal or opt out (SB 24-205 2024). The Colorado law covers a broader range of AI applications than California's, including predictive scoring and automated triage, and it places the disclosure obligation on the institution using the AI, not just on the party that communicates the output to the patient.

The FTC's Operation AI Comply enforcement action in 2024 targeted companies making deceptive claims about AI capabilities in healthcare contexts — claiming human-level accuracy for tools that had not been validated, implying physician oversight for tools that had none (Federal Trade Commission 2024). The enforcement record makes clear that the FTC treats healthcare AI interfaces as covered by existing consumer protection law, and that institutions should expect regulatory attention to the gap between what they claim about AI tools and what the tools actually do.

12.7 Trust Recovery After Adverse AI Events

An AI tool will eventually produce an outcome that harms or nearly harms a patient at any AMC that deploys AI at scale. How the institution communicates about that event — to the patient, to the clinical staff involved, and to the public — will matter more for long-term trust than whether the event occurred at all. Institutions that respond to adverse AI events with transparency and accountability recover trust more effectively than those that minimize or deflect (Jones et al. 2023).

The trust recovery research in other high-stakes domains — aviation, nuclear power, food safety — is consistent on two principles. First, prompt and specific disclosure of what happened and what is being done in response is more effective than delayed disclosure, even when the complete picture is not yet known. "We are investigating an error in our AI system that may have affected your care, and we will contact you with our findings by [date]" is more trust-preserving than silence followed by a detailed disclosure weeks later. Second, apology without accountability — "we are sorry this happened" without a commitment to specific changes — does not restore trust. Patients want to know that the institution understands what went wrong and has taken concrete steps to prevent recurrence.

For institutions with an AI transparency report program — a public annual or semi-annual report that describes deployed AI tools, their performance metrics, and any documented adverse events — the adverse event entry provides evidence that the reporting is genuine. A transparency report that reports zero adverse events in a large AI deployment is not credible. A

report that describes an adverse event, the investigation findings, and the remediation measures is evidence that the institution is actually monitoring.

12.8 Where to Start

12.8.1 Starter Project 1: Patient AI Disclosure Audit and Language Standardization

What it is: An audit of every current patient-facing AI interaction to assess whether disclosure is occurring, what language is being used, and whether that language meets the substantive standard described above — specific, human role clarified, recourse identified.

Why now: California AB 3030 is in effect, and Colorado SB24-205 takes effect in 2026. An institution that has not audited its disclosure practices is not in a position to certify compliance, and more importantly, is not in a position to know whether patients understand what AI is being used in their care.

How to execute: Map every patient-facing AI touchpoint: portal messages, discharge instructions, appointment communications, chatbot interactions, ambient documentation encounters. For each touchpoint, assess: is disclosure occurring? Is it meaningful or boilerplate? Does it meet AB 3030's requirement for communications not reviewed by a licensed human? The output is a prioritized remediation plan and a set of standardized disclosure language options for each touchpoint type.

Buy vs. build: Process and language work, not a technology purchase. Standardized disclosure language templates require legal review, not a vendor relationship.

12.8.2 Starter Project 2: Establish a Patient AI Advisory Council

What it is: A standing advisory council with defined membership, mandate, and relationship to the AI Steering Committee, recruited specifically to advise on AI tool selection and deployment from the patient and community perspective.

Why now: The institutions that deploy AI most sustainably are the ones that have community confidence before adverse events occur, not the ones scrambling to build community relationships afterward. The DiMe Society clinical AI maturity model places patient advisory engagement as a gold-standard governance indicator. An institution without a Patient AI Advisory Council is not yet at the governance standard the field is converging toward.

How to execute: Define the charter before recruiting members: what does the council advise on, how does its input reach the AISC, and what decisions (if any) require council consultation before proceeding? Recruit at least half the membership from communities with documented reasons to be concerned about algorithmic bias in healthcare. Build in access to a technical

liaison — a clinical informatics team member who can explain, in accessible terms, how specific AI tools work and what the known risks are. The council should meet monthly during active deployment periods and quarterly during steady-state operations.

Buy vs. build: Program design and staff time. No technology purchase required. The significant investment is in the quality of recruitment and the integrity of the governance relationship — specifically, whether the council's input is genuinely considered or merely documented.

13 Professional Wellness and Reducing Cognitive Burden

The case for AI in healthcare governance documents is typically framed around what AI can do for patients — more accurate diagnoses, more consistent treatment recommendations, earlier identification of deterioration. That framing is appropriate for those purposes. But there is a second case for clinical AI that rarely receives the same sustained attention, and that is the case for what it can do for the people providing care.

Clinician burnout in the United States has reached a level that the U.S. Department of Health and Human Services has characterized as a public health threat (U.S. Department of Health and Human Services, Office of the Surgeon General 2022). More than half of physicians in some specialties report symptoms of burnout, and the rates among nurses and advanced practice providers are comparable. This is not a resilience deficit, and it is not a recent development. It is the predictable consequence of a documentation and administrative burden that has grown continuously for two decades while clinical staffing has not kept pace. The introduction of the electronic health record improved data availability and reduced certain categories of error; it also created what physicians call "pajama time" — the hours spent completing documentation at home, after clinic hours, in the margins of what used to be personal time.

The empirical evidence that AI tools — specifically ambient documentation systems and AI-assisted inbox management — can meaningfully reduce this burden has now reached the level required to act on. This chapter makes that case with the data and tells AMC clinical and operational leaders what they need to know to deploy these tools responsibly and at scale. The argument is not that AI will solve burnout; burnout is a structural problem that requires structural solutions, including staffing ratios, workload limits, and EHR simplification. The argument is that ambient documentation and inbox AI produce measurable, clinically meaningful time savings that represent one of the most concrete near-term levers available to AMC leaders who are watching experienced clinicians exit the profession.

13.1 The Structural Crisis of the Administrative Burden

The EHR documentation burden did not originate with the EHR. It grew with it. As electronic records made clinical data more accessible for billing, quality measurement, and regulatory reporting, the number of required data fields, structured documentation elements, and attestation requirements expanded to fill the new capacity. What began as a data capture

tool became a documentation obligation that, by some estimates, requires two hours of EHR work for every hour of direct patient care.

The AMA's annual Prior Authorization Physician Survey documents one slice of this burden in stark terms. Physicians and their staff spend an average of thirteen hours per week on prior authorization activities alone — not including other administrative tasks, inbox management, documentation, and billing requirements. Ninety-five percent of physicians in the survey report that prior authorization contributes to burnout (American Medical Association 2023). The combined administrative load means that physicians in high-volume specialties may spend more of their working hours on tasks that are not direct patient care than on tasks that are.

The burnout crisis has a financial dimension that is often underappreciated by AMC leaders who focus on the subscription cost of AI tools. The AMA estimates that replacing a single physician — recruiting, credentialing, onboarding, and covering lost productivity during the transition — costs between five hundred thousand and one million dollars, depending on the specialty (American Medical Association 2023). Nurse turnover costs are lower per position but occur at higher frequency. For an institution with a hundred-physician practice that experiences a five percent annual attrition rate, the turnover cost in a single year is between two and five million dollars. An ambient documentation subscription that prevents even a fraction of that attrition generates a return on investment that most technology investments cannot approach.

13.2 Pajama Time: The Mechanism of Burnout

The concept of "pajama time" — named for the clothes physicians are wearing when they complete documentation after clinic hours — describes a specific phenomenon that the burnout literature has linked particularly strongly to exit intent. It is not simply that physicians work long hours; many physicians accept long hours as inherent to the profession. It is that documentation work bleeds into non-work time in a way that eliminates the psychological recovery that even modest off-hours downtime provides.

Tait Shanafelt's longitudinal studies of physician well-being, conducted through the Mayo Clinic program on physician well-being, have consistently found that EHR-related frustration is among the strongest independent predictors of burnout and exit intent — stronger than compensation, call schedule, and practice setting in some analyses. The specific mechanism is not EHR use during clinic hours; it is EHR use during hours that clinicians experience as personal time. The documentation does not get shorter when it moves to after hours; the psychological cost is higher because the context shift makes the burden more salient.

Specialty variation is substantial. Emergency medicine physicians report some of the highest burnout rates, with documentation demands that follow an acute encounter model in which notes cannot be completed until after the patient has left and the next patient has arrived. Primary care physicians face note volume that can exceed twenty encounters per session; even

modest per-note time savings aggregate to meaningful weekly hour reductions. Mental health providers have among the most complex documentation requirements given regulatory and billing requirements specific to behavioral health. The uniform finding across specialties is that documentation burden correlates with burnout more consistently than almost any other measured variable.

13.3 Ambient AI Documentation: The Evidence Base

Ambient AI documentation systems — which listen to the clinical encounter, transcribe and synthesize the conversation, and generate a structured draft clinical note — entered serious clinical deployment in 2023 and have accumulated a meaningful evidence base in the following two years. The headline findings are consistent enough across sites and vendors to warrant treatment as established rather than emerging.

Tierney and colleagues conducted one of the most rigorous early evaluations of an ambient scribe deployment across a large multispecialty practice. They found that physicians reported meaningful reductions in documentation time, with the largest effects in primary care and outpatient mental health — the two specialties with the highest note volume and the highest burnout rates (Tierney et al. 2024). Notes generated with AI assistance were rated comparable in quality to physician-authored notes on blinded review. Clinician well-being scores improved, and — notably — patients reported that their physicians seemed more attentive during encounters, because the physician was no longer managing a keyboard.

The patient attention finding deserves emphasis because it connects the wellness value proposition to the care quality value proposition. An encounter in which the clinician maintains eye contact, responds to nonverbal communication, and does not divide attention between the patient and a screen is a different clinical interaction from one in which documentation competes with the patient for the physician's attention. A 2025 study found that patients who received care from physicians using an ambient scribe system rated their clinician's attentiveness significantly higher than patients in a control group — not because the physician was doing anything differently in terms of clinical care, but because removing the documentation burden from the room changed the quality of the relationship.

Table Table 13.1 summarizes the time-savings findings from major evaluations of the leading commercial ambient AI systems. Figures should be treated as directional rather than precise: study designs vary, comparison conditions vary, and the systems themselves have continued to evolve.

Table 13.1: Summary of time-savings and clinician satisfaction findings from published and reported evaluations of commercial ambient AI scribe systems, 2024–2025. Note variation in study designs; figures are not directly comparable across rows.

System	Study / Source	Specialty	After-Hours EHR Reduction	Time Saved per Encounter	Clinician Satisfaction
Abridge	Tierney et al. 2024	Multi-specialty	Substantial reduction reported	~1 hour/day	Improved (quality score 48/50)
DAX (Nuance/Microsoft)	Atrium Health deployment	Multi-specialty	47% of users saw decreased home EHR time	Significant	High adoption rates
Nabla	Primary care RCT, 2025	Primary care	Not separately reported	9.5% reduction in time-in-note	Positive
Suki	Blinded comparison, 2025	Multi-specialty	Not separately reported	Notes rated more thorough	Mixed (complex cases: human slightly better)

The honest accounting of ambient AI limitations is as important as the positive findings. Ambient systems produce draft notes, not finished notes. The clinical content requires review before attestation, and the review is not costless: detecting a subtle omission — a negative finding the system did not capture, a medication allergy mentioned in passing — requires active clinical attention to a document that is formatted to look complete. The omission risk was discussed in Chapter 6; the wellness implication is that the "time savings" claimed by vendors may be partially offset by the cognitive effort of editing rather than writing, particularly for clinicians who are careful reviewers.

The appropriate institutional posture is to measure this tradeoff directly. A pre/post measurement design that captures not just total EHR time but after-hours EHR time, note completion latency, and clinician-reported task load provides the data needed to assess whether the specific deployment at the specific institution is producing the wellness benefit the vendor evidence supports.

13.4 The Ambient Consent Architecture

The fact that ambient systems listen to clinical encounters — by definition, in the same room as the patient — creates a consent and privacy obligation that institutions deploying these tools must address explicitly. The patient is being recorded; the patient's words are being

processed by an AI system operated by a third party; and in many states, the patient has a legal right to know this before it occurs.

The consent architecture should have four components. First, the patient should be told, at the start of the encounter, that an ambient AI system is active, what it does, and that they can ask for it to be turned off. This does not require a lengthy explanation; a brief verbal explanation plus an opportunity to decline is sufficient for most clinical contexts. Second, the vendor contract should specify that audio data is processed in transit and is not retained after note generation — a "zero-retention" policy that reduces the privacy risk surface. Third, the generated transcript and the note should be treated as protected health information under HIPAA from the moment of creation. Fourth, the attestation statement that the clinician signs should acknowledge AI involvement in the note's preparation, consistent with California AB 3030 (AB 3030 2024) and the general transparency standards described in Chapter 12.

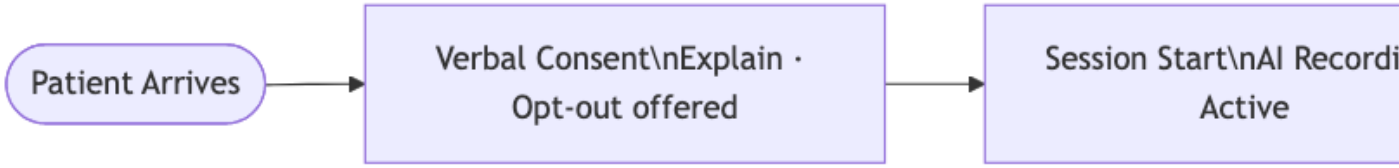


Figure 13.1: Ambient AI documentation workflow with consent and privacy checkpoints. Zero-retention audio processing limits the PHI exposure window to the encounter itself.

13.5 AI-Assisted Inbox Management

The clinical inbox — the stream of patient messages, abnormal results, prescription refills, referral requests, and administrative tasks that arrives continuously in the EHR — is a second major source of after-hours burden that ambient documentation does not address. Studies of EHR inbox volume suggest that physicians in primary care and internal medicine spend two to three hours per day on inbox-related tasks, with a significant fraction of that time occurring outside scheduled clinic hours.

AI-assisted inbox management addresses this burden in two ways. Message triage tools classify incoming messages by urgency and type, routing refill requests, administrative queries, and low-urgency informational messages to appropriate staff or queues rather than to the attending physician's personal inbox. Message drafting tools generate suggested responses to patient messages — responses that a clinician reviews and sends, or modifies and sends, rather than composing from scratch.

The Mayo Clinic's deployment of AI-drafted patient message responses across nursing staff reported savings of approximately thirty seconds per message; across the organization's message volume, this aggregated to roughly fifteen hundred hours of nursing time per month. At an institution where nursing turnover is expensive and inbox-related burnout is a retention risk,

that represents meaningful institutional value. A pilot at NYU Langone found that AI-drafted patient messages scored higher on patient-rated empathy than clinician-authored responses — a counterintuitive finding that the researchers attributed to the AI drafts being more consistently warm in tone, without the variation that comes from a clinician responding to their twenty-fifth message of the day at eleven at night.

The safety question for inbox AI is parallel to the safety question for ambient documentation: the AI draft is a starting point, not a final answer, and the clinician review is not optional. A patient message about a new chest pain symptom requires a different response than a message about a refill request, and the AI may not reliably distinguish between them in every case. The appropriate governance design is to treat AI inbox drafts as productivity tools that reduce the blank-page problem, not as autonomous response generators.

13.6 Automation Complacency and the Vigilance Gap

The single most important safety warning for ambient documentation and inbox AI is also the most predictable: the tools will work well enough, often enough, that clinicians will stop reviewing them carefully. This is not a hypothetical; it is the documented behavior of humans using any consistently accurate automated system (Parasuraman and Manzey 2010).

The "automation complacency trap" in clinical AI takes a specific form. A clinician who has reviewed two hundred AI-generated notes without finding a significant error will review the two hundred and first note less carefully than the first. Their eyes will pass over the text; their cognitive attention will be elsewhere. The error that slips through is not a random error — it is the subtle, plausible-sounding error that requires active clinical attention to catch. The system that makes large, obvious errors is less dangerous than the system that makes small, fluent-sounding errors, because the former will be caught and the latter will be signed.

The governance response is to build in mechanisms that maintain the clinician's engagement with the review task. Some vendors have experimented with displaying "confidence scores" or highlighting sections of the note that were reconstructed from partial information — visual signals that certain passages require more careful scrutiny. Some institutions have implemented periodic "vigilance audits" in which a clinical informaticist reviews a random sample of signed AI-generated notes against the original encounter recording to assess whether clinicians are catching the errors the system makes. The goal is not to create an adversarial relationship between the clinician and the tool; it is to preserve the review function that makes the system safe.

13.7 Nursing and Advanced Practice Burden

The burnout and documentation burden literature is weighted heavily toward physicians, partly because physicians have been more systematically surveyed and partly because physician attrition is more economically visible. The burden on nursing staff and advanced practice providers is equally real and somewhat distinct in character.

Nurses in inpatient settings face documentation requirements that are simultaneous with direct patient care responsibilities — they are documenting vital signs, medication administration, and patient assessments while also providing care, often in conditions where the documentation system and the care environment are in physical conflict. Ambient documentation systems were designed primarily for the physician encounter model — a structured conversation between a physician and a patient — and do not yet map well to the parallel, distributed, and physically active documentation patterns of bedside nursing.

One nursing-focused deployment study found that ambient AI could reduce nursing documentation time by a substantial margin in simulated conditions, and by a smaller but still meaningful margin in actual clinical practice. The gap between simulated and actual savings is instructive: ambient systems perform better in conditions where the conversational signal is clean and the documentation task is structured, and nursing documentation in many inpatient settings is neither. The honest characterization of ambient AI for nursing is that the potential benefit is real but the peer-reviewed evidence is thinner than for physicians, and the implementation design requires more adaptation to nursing workflows than most vendors have undertaken.

Advanced practice providers — nurse practitioners, physician assistants, certified nurse midwives, and clinical pharmacists — share both the physician's note-volume challenge and the nursing context of frequent task-switching. Burnout surveys of APPs have documented rates comparable to physicians in similar specialties, and the documentation burden is a consistent contributor. The current evidence base for ambient AI among APPs is largely derived from studies that enrolled APPs as part of multispecialty cohorts; dedicated APP studies are limited.

13.8 Where to Start

The wellness ROI case for ambient documentation is strong enough that most AMCs should be piloting at least one ambient system. The two projects below are designed to maximize the probability that the pilot produces usable evidence and, if successful, a foundation for responsible scale.

13.8.1 Starter Project 1: Ambient Documentation Pilot with Pre/Post Wellness Measurement

What it is: A structured twelve-week pilot of a single ambient AI scribe in a high- burnout outpatient specialty — primary care, mental health, or internal medicine are the highest-yield targets — with validated pre/post measurement of after-hours EHR time, physician well-being, and note quality.

Why now: The commercial systems (Abridge[1], DAX[2], Nabla[3]) have BAA-ready EHR integrations and vendor-side regulatory responsibility for the AI component. The institutional work is structuring the measurement framework, not building the technology. An institution that waits for more evidence before piloting is deferring a decision that the evidence already supports.

How to execute: Identify a willing specialty lead and a clinical champion. Pull baseline after-hours EHR time from EHR audit logs for two months before deployment. Administer a validated burnout survey (the two-item Maslach Emotional Exhaustion subscale or the Stanford Professional Fulfillment Index) to participating clinicians before and after. Establish note quality review with a blinded clinician reviewer on a random ten-percent sample throughout the pilot. Deploy with the consent and attestation design described in Figure 13.1. Measure at 60 and 90 days. Report results to clinical leadership and the AI Steering Committee whether the outcome is positive, null, or negative.

Buy vs. build: Buy. Building an ambient documentation system from scratch is not a tractable project for clinical informatics teams, and the commercial market has matured. The institutional investment is in measurement design and change management.

13.8.2 Starter Project 2: Prior Authorization Workflow Automation Assessment

What it is: An analysis of the institution's current prior authorization burden — total hours, denial rate, appeal rate, and staff time by specialty — and a structured evaluation of commercial PA automation solutions against that baseline.

Why now: The CMS Interoperability and Prior Authorization Final Rule (Centers for Medicare and Medicaid Services 2024a) mandates that payers accelerate PA response timelines beginning in 2026. Providers who are currently absorbing PA burden manually will face either increased volume or worse service levels unless they automate some portion of the workflow. The regulatory mandate creates a forcing function that makes PA automation evaluation overdue for most AMCs.

[1] https://www.abridge.com
[2] https://www.nuance.com/healthcare/ambient-clinical-intelligence.html
[3] https://www.nabla.com

How to execute: Work with the revenue cycle team to pull current PA metrics: volume by specialty, average staff time per PA request, denial rate by payer, appeal rate, and cost per PA in staff hours. Map the portions of the workflow that are currently manual and that a commercial PA automation tool could address. Evaluate two or three commercial vendors against the measured baseline, with particular attention to their accuracy documentation, their EHR integration path, and their error-handling design when the AI is uncertain or wrong. Use the evaluation to inform a business case for the CFO that uses the actual institutional cost data rather than vendor-supplied estimates.

Buy vs. build: Evaluate and likely buy. PA automation requires deep payer integration that is impractical to build de novo. The institutional differentiation is in the governance design — specifically, the HITL architecture for cases where the AI recommendation should not be submitted without human review.

Part IV

Workstream Resources

14 IT Infrastructure and Security

Every AMC CIO faces a version of the same question: should we build our own AI infrastructure, buy commercial AI platforms, or connect existing systems to foundation models through APIs? The framing of "build vs. buy" is familiar from decades of enterprise IT decisions, but it maps poorly onto the current AI landscape. Building a frontier foundation model is not a realistic option for any AMC; the compute and data requirements are measured in thousands of GPUs and hundreds of millions of dollars. Buying a turnkey clinical AI platform means accepting the vendor's model choices, governance design, and update cadence. The third option — connecting existing institutional systems to foundation models through a managed API gateway — is the approach that gives most AMCs the best combination of capability, control, and cost at the current state of the technology.

This chapter describes what that "connect" architecture looks like, how to secure it, and how to use it to defeat the most common AMC AI security failure: clinicians and researchers bypassing institutional governance by using consumer AI tools with patient data.

14.1 The Buy/Build/Connect Trilemma

Training a proprietary foundation model requires computational resources and data volumes that exist at a handful of commercial research labs. Most AMC AI programs that have pursued model training — fine-tuning a clinical model on institutional EHR data, for example — are doing so not to build a new foundation model but to adapt an existing one to their specific clinical context. This is a meaningful distinction. Fine-tuning a 7-billion-parameter open model on OMOP-formatted clinical notes is a tractable project for an AMC with a GPU cluster. Training a 70-billion-parameter foundation model from scratch is not.

The "buy" option has improved rapidly. Microsoft Azure OpenAI, AWS Bedrock, and Google Vertex AI all offer HIPAA-eligible foundation model APIs with signed BAAs, US-only data residency options, and zero-data-retention configurations for prompt content. For most clinical and administrative AI use cases, the question is not whether to connect to a foundation model but which one and how to govern the connection.

The "connect" approach means building institutional infrastructure around the API connection rather than around the model itself. The critical institutional asset is not the model — which the vendor manages and updates — but the governance layer: the access controls, audit logging,

PHI filtering, and use policies that determine who can connect to what model with what data under what oversight conditions.

14.2 The Institutional API Gateway

The central architectural element of a sound AMC AI infrastructure is a managed API gateway — a system that routes all AI API calls from institutional users and applications through a single, governed chokepoint. Rather than allowing each clinical application, research tool, and administrative system to connect directly to AI model APIs with its own credentials and its own security posture, the gateway enforces institutional policy uniformly across all connections.

A purpose-built AI gateway (LiteLLM[1], Kong AI Gateway[2], or equivalent) provides several capabilities that direct API connections lack. It enforces authentication and authorization — only institutional users with appropriate roles can access specific model endpoints, and the gateway can enforce role-based access control tied to the existing institutional identity provider. It performs real-time PHI scanning on outbound prompts, flagging or blocking requests that contain identifiable patient information before they leave the institutional network boundary. It enforces model-specific policies — certain models may be authorized for clinical use but not for research, or for internal communications but not for patient-facing outputs. And it maintains an immutable audit log of every API call: who made it, from which application, to which model, at what time, with what response latency.

The gateway approach also solves the cost management problem that enterprise AI programs frequently underestimate. Foundation model API calls are priced per token, and usage can grow dramatically when a new tool is adopted at scale. Without centralized cost tracking, an AMC has no visibility into which applications are driving spend, no mechanism to enforce usage limits on individual applications or users, and no early warning when an anomalously large request — a sign of either misconfigured behavior or adversarial input — is consuming unexpected resources.

14.3 Clinical RAG Architecture

The dominant architectural pattern for grounding clinical AI outputs in institutional knowledge is Retrieval-Augmented Generation (RAG). Rather than relying on a foundation model's training-time knowledge — which has a cutoff date and does not include institution-specific clinical protocols, formularies, or guidelines — a RAG system retrieves relevant documents from a curated knowledge base and includes them in the prompt context at inference time. The model generates its response grounded in the retrieved documents rather than in parametric memory alone.

[1] https://www.litellm.ai/
[2] https://konghq.com/products/kong-ai-gateway

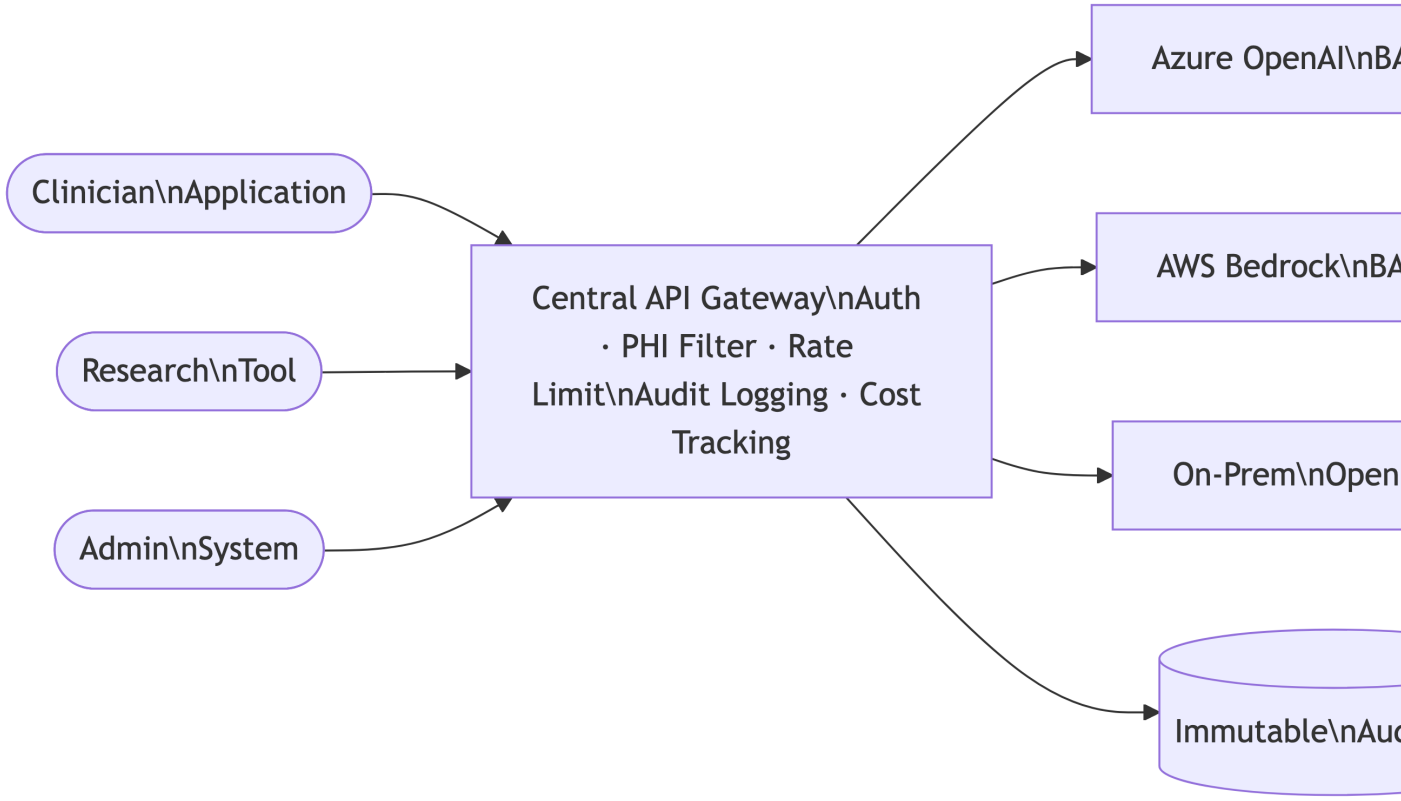


Figure 14.1: Institutional AI API gateway architecture. The gateway is the single chokepoint through which all AI traffic flows, enabling centralized security enforcement, audit logging, and cost management.

For clinical applications, the RAG knowledge base typically contains institutional clinical guidelines, formulary information, drug interaction databases, local evidence-based protocols, and relevant peer-reviewed literature. The quality of the RAG output is bounded by the quality of the knowledge base: a model that retrieves an outdated clinical protocol will produce an outdated recommendation, and there is nothing in the model itself that will flag the content as outdated.

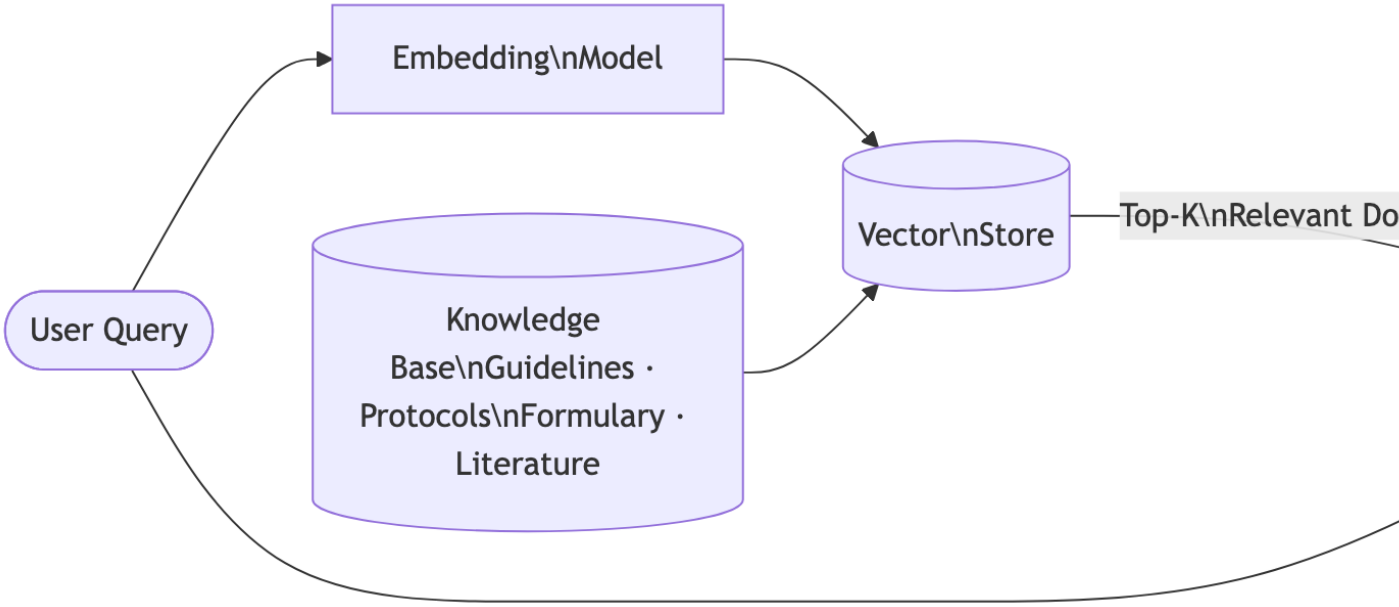


Figure 14.2: Clinical RAG pipeline. The retrieval step grounds the model's response in current institutional knowledge, reducing hallucination risk for institution-specific content.

The security implications of RAG deserve explicit attention. The retrieved documents are included in the prompt context, which means they are transmitted to the model API. If the knowledge base contains PHI — patient-specific clinical notes, for example — those notes will be transmitted in every query that retrieves them. For clinical RAG systems, the knowledge base should consist of de-identified or non-PHI content where possible; patient-specific RAG requires patient-specific access controls at the retrieval layer and should be treated as a regulated data use case under the governance framework in Chapter 17.

Indirect prompt injection — an attack in which malicious content embedded in retrieved documents manipulates the model's behavior — is a genuine security risk for RAG systems (Greshake et al. 2023). A clinical RAG system that retrieves documents from external sources (web pages, external literature databases) is more vulnerable than one that retrieves from a curated, internally controlled knowledge base. For clinical use, the knowledge base should be controlled, versioned, and audited, with external content ingested only after review.

14.4 The Security Stack

The NIST AI Risk Management Framework provides the governance structure for AMC AI security; the concrete security controls map the abstract framework functions to operational implementations (National Institute of Standards and Technology 2023, 2024).

The Govern function requires documented policies for who can use what AI tools with what data, and a governance body with authority to approve or restrict deployments. The gateway architecture is the technical implementation of this function.

The Map function requires identifying which AI tools are deployed, what risks they pose, and which organizational functions they affect. The AI tool inventory (see Section 6.4) is the primary Map artifact.

The Measure function requires performance monitoring — not just accuracy metrics, but security metrics: PHI scan false-negative rate, prompt injection attempt detection rate, anomalous usage patterns. These metrics should be reviewed by the CISO and the AI governance committee on a regular cadence.

The Manage function requires documented incident response procedures for AI security events. What is the response when a PHI scan failure allows patient data to reach an external model? What is the response when prompt injection is detected? The incident response plan for AI security events should be tested annually, like any other security incident response plan.

14.5 Ambient AI and the EHR Integration Layer

The gateway architecture assumes a clean separation between the AI system and the EHR. By 2025 that assumption no longer holds for clinical documentation. Ambient AI tools — Nuance DAX Copilot[3], Abridge[4], Suki[5], Nabla[6] — are now embedded directly inside EHR workflows rather than accessed through a separate browser tab or institutional portal. DAX Copilot operates within Epic's clinical narrative module via Epic's native integration. Abridge's note generation writes directly into Epic's documentation framework. The ambient AI is not connecting to the institution's gateway; it is connecting to the EHR vendor's infrastructure, under the EHR vendor's BAA, through the EHR vendor's integration channel.

This matters for governance in ways that a standard gateway deployment does not. When the ambient AI is an Epic partner product operating inside Epic's infrastructure, the audit logging may live in Epic's system rather than the institution's SIEM. The PHI flow — patient audio, transcript, structured note — may never touch the institutional gateway at all. The

[3] https://www.nuance.com/healthcare/ambient-clinical-intelligence.html
[4] https://www.abridge.com
[5] https://www.suki.ai
[6] https://www.nabla.com

institution has traded a single governed chokepoint for a shared-responsibility model with the EHR vendor, and the terms of that responsibility are defined by a contract most institutions have not specifically negotiated for AI use.

The practical governance response has three components. First, the ambient AI vendor contract must explicitly address audio retention policy, transcript handling, and who holds the BAA obligations when the tool is operating inside the EHR vendor's infrastructure — because in a sub-processor arrangement the primary BAA with the EHR vendor may or may not cover the ambient AI module's data flows. Second, the institution needs audit logging at the output layer even if it cannot capture the full ambient AI session: every AI-generated note that is attested in the EHR creates an audit record, and those attestation records are the primary evidence chain for governance oversight of ambient AI use. Third, the institutional AI governance committee needs a seat in EHR AI integration decisions, not just in standalone tool deployments. When Epic enables a new ambient AI partner integration, that is a deployment event. It requires the same governance review as any other tool that influences clinical documentation.

14.6 Agentic Infrastructure: Beyond the Advisory Model

The governance architecture described above — gateway, PHI scan, audit log, model API — was designed for advisory AI: systems that generate outputs a human evaluates and acts on. The AI recommends; the clinician decides. That model is giving way to agentic AI: systems that can take autonomous multi-step actions — querying the EHR for lab results, drafting a referral order, routing a message to a specialist, scheduling a follow-up appointment.

Epic's Digital Workforce, Oracle Health[7]'s Clinical AI Agent, and the first generation of clinical AI copilots built on general-purpose agent frameworks are already operating in production at some institutions. The infrastructure requirements for these systems are fundamentally different from advisory AI. Four controls matter most.

Non-Human Identity management. An agentic system that can query and write to the EHR needs a service account with defined permissions. The SMART on FHIR backend services authorization flow provides the mechanism: the agent authenticates with a private key registered against an institutional client credential, and its authorized scopes determine exactly which FHIR resources it can read and write. The governance principle is least-privilege: an agentic scheduling system should have FHIR scopes limited to the Appointment and Schedule resources it actually needs, not a blanket administrative credential (HL7 International 2024).

Semantic circuit breakers. Agentic systems can enter reasoning loops — taking an action, observing the result, taking another action — that compound errors in ways that advisory AI cannot. A circuit breaker is a preconfigured rule that halts autonomous action when a defined threshold is reached: a maximum number of consecutive actions without human confirmation, a rate limit on write operations, a flag on any action outside the agent's validated action space.

[7] https://www.oracle.com/health/

These are engineering controls, not policy statements, and they need to be specified before a clinical agent is deployed, not after it takes an unexpected action.

Action rollback capability. When an agentic system takes an incorrect action — drafts an order for the wrong patient, routes a message to the wrong provider — the governance question is not just how to prevent this but how to undo it cleanly. Reversible actions (draft orders, unsent messages) need a defined rollback path. Irreversible actions (sent messages, confirmed appointments) require a higher pre-authorization bar. The boundary between reversible and irreversible should be explicit in the system design, not discovered after the fact.

Action-level audit logging. Prompt logging records what the user asked. Action logging records what the system did. For an advisory AI, prompt logging is the primary audit record. For an agentic AI, action logging is what regulators, risk managers, and clinical safety officers will want when something goes wrong. The immutable audit log in the gateway architecture needs to be extended to capture discrete agent actions with timestamps, the user or trigger that initiated the action, and whether a human reviewed the action before it was executed.

14.7 Sovereign Cloud and On-Premises Deployment

Most AMC AI will run in enterprise cloud tenants — Azure OpenAI, AWS Bedrock, Google Vertex AI — because the economics are right and BAA availability has improved substantially since 2022. But there is a class of data for which cloud transmission is not an appropriate governance posture regardless of BAA provisions.

Genomic sequence data, psychiatric treatment records, substance use treatment records under 42 CFR Part 2, and data subject to specific research consent restrictions all fall into a category where institutional data stewardship obligations exceed what a commercial vendor agreement can satisfy. For these use cases, the architecture is either a sovereign cloud environment — Azure Government, AWS GovCloud, or Google Assured Workloads, all of which offer HIPAA-eligible compute with US-person controls and FedRAMP authorization — or on-premises model serving.

On-premises LLM serving has become more tractable than it was three years ago. A 70-billion-parameter model running on A100 or H100 GPU hardware produces inference throughput sufficient for most research and limited clinical use cases. Open-weight models under licenses that permit institutional deployment — Llama 3.1, Mistral[8], Phi-4 — allow fine-tuning on local clinical data without the data leaving the institutional network boundary. The operational overhead is real: the institution bears the maintenance burden for serving infrastructure, model updates, and performance monitoring that commercial providers handle invisibly. For restricted data tiers, that overhead is the price of the required data sovereignty.

[8] https://mistral.ai/

The practical framework is not "cloud or on-prem" but which data tier requires which infrastructure tier. The institutional data classification framework in Chapter 17 is the input to this decision. Regulated data (PHI, FERPA, Common Rule) goes through enterprise cloud tenants with verified BAAs. Restricted data (genomic, Part 2, specific consent restrictions) goes through sovereign cloud or on-premises. Public and internal data can use any enterprise-grade service. The governance committee's job is to enforce the mapping consistently, not adjudicate it case by case.

14.8 Defeating Shadow AI

Shadow AI — the use of consumer AI tools (ChatGPT, Gemini, Claude.ai) with institutional data, in violation of institutional policy — is the most common AMC AI security failure. It happens because the sanctioned alternatives are inconvenient, the consumer tools are capable, and the clinicians and researchers using them do not understand the specific risks they are creating. A "zero tolerance" approach — blocking all consumer AI sites at the network boundary — is both technically difficult and likely to fail: motivated users find workarounds, and overly aggressive blocking damages the institutional relationship with clinicians who are trying to do their jobs.

The more effective approach is to make the sanctioned path easier than the unsanctioned path. An internal AI assistant — accessible through the institutional portal, pre-authorized for appropriate data types, and connected to the institutional knowledge base through RAG — eliminates the most common reason for shadow AI use: the institutional tool either doesn't exist or is harder to access than the consumer alternative. This is the "infrastructure as policy" principle: the security control is the product quality, not the network block (Thirunavukarasu et al. 2023).

14.9 Where to Start

14.9.1 Starter Project 1: Institutional API Gateway Deployment

What it is: Deployment of a centralized AI API gateway that routes all institutional AI traffic through a single managed system with authentication, PHI scanning, audit logging, and cost tracking.

Why now: Every day that institutional users connect directly to AI model APIs without a gateway is a day without audit logging, without PHI scanning on outbound prompts, and without visibility into institutional AI usage patterns. The gateway is the foundational security control that everything else depends on.

How to execute: LiteLLM and Kong AI Gateway are the leading open-source options; Azure API Management can serve the same function for Azure OpenAI-heavy deployments. The implementation work involves: (1) provisioning the gateway instance; (2) configuring connections to each approved model provider; (3) integrating with the institutional IdP for authentication; (4) configuring PHI scanning rules; (5) connecting audit log output to the SIEM. This is a four-to-eight week engineering project for a team with cloud infrastructure experience.

Buy vs. build: The gateway software is open-source or commercial off-the-shelf. The institutional work is configuration, integration, and governance policy definition.

14.9.2 Starter Project 2: Internal AI Assistant with Clinical RAG

What it is: An institution-facing AI assistant that clinical and administrative staff can use for productivity tasks — drafting, summarization, question answering — with access controls appropriate to their role and grounding in institutional knowledge through a curated RAG knowledge base.

Why now: This is the primary countermeasure to shadow AI. An internal tool that works well for the most common productivity use cases eliminates the primary motivation for using consumer tools with institutional data.

How to execute: Build on top of the API gateway. Define the knowledge base content for the initial deployment — clinical guidelines, formulary, institutional policies. Implement FHIR-based access controls for role-specific knowledge access. Deploy through the institutional intranet or EHR patient context for clinical staff. Measure shadow AI usage (via network telemetry) before and after deployment to assess effectiveness.

Buy vs. build: Mixed. The gateway and RAG infrastructure is a build project. Several vendors (Nabla, Abridge, Microsoft Copilot for Healthcare, Google Workspace AI) offer institutional assistant products with existing EHR integrations that can reduce the build burden for common clinical use cases.

15 Training and Workforce Development

The most sophisticated AI infrastructure an AMC can build is only as effective as the workforce that uses it. A clinical AI tool that clinicians do not trust will not be used. An inbox management system that administrative staff route around will not reduce the documentation burden. An AI platform that no one in the institution can evaluate for bias or validate for accuracy is a governance fiction. AI literacy — the practical capacity to understand, evaluate, and deploy AI tools appropriately in a specific role — is not a peripheral technical skill. It is a prerequisite for the institutional AI strategy to work at all.

The gap between what AMC AI programs assume about workforce readiness and actual current readiness is substantial. Multiple validated assessment instruments have found that medical students, residents, nurses, and faculty consistently overestimate their AI knowledge on self-report measures and score substantially lower on objective assessments. This Dunning- Kruger pattern in AI literacy is a governance risk: clinicians who believe they understand AI well enough to evaluate it will not recognize when they need help doing so, and will not escalate concerns that a more calibrated colleague would flag.

This chapter describes the tiered competency model that matches training requirements to role-specific AI use, the national frameworks that define the content of those competencies, and the practical governance interventions that make AI literacy sustainable rather than a one-time onboarding event.

15.1 The Four-Tier Competency Model

The most durable framework for AMC AI workforce development is a four-tier model adapted from the evidence-based medicine approach proposed by Ng and colleagues (Ng et al. 2023). The four tiers map to different relationships with AI tools, different risk profiles, and different training requirements.

Consumers use AI tools for their intended purpose without modifying them — a clinician using an ambient scribe, an administrator using an AI drafting tool, a nurse using an AI inbox triage system. Consumer literacy requires understanding what the tool does, what its known limitations are, how to identify outputs that warrant skepticism, and how to document AI-assisted work in the institutional record. This is the baseline literacy that every AMC employee who touches an AI tool needs.

Translators bridge between AI capabilities and clinical or operational applications — a clinical informatics specialist evaluating a vendor's predictive model, a quality improvement lead using AI to analyze patient safety data, a department champion who helps colleagues adopt a new AI workflow. Translator literacy requires the ability to read and interpret model performance reports, evaluate bias audit findings, understand the difference between validation and calibration, and communicate AI limitations to non-technical audiences.

Developers build and deploy AI tools — data scientists training models, engineers integrating AI APIs into clinical systems, informatics fellows creating RAG pipelines. Developer literacy requires formal training in machine learning, statistical inference, data governance, and the regulatory landscape for AI-enabled medical devices.

Governors oversee institutional AI strategy and assume accountability for AI risk — CMIOs, CIOs, CMOs, CFOs, and board members. Governor literacy is not technical; it is risk-management literacy. Governors need to understand how to ask the right questions of the people who build and deploy AI tools, how to evaluate vendor claims, how to structure accountability across the institution, and how to recognize when an AI governance failure is occurring before it becomes a patient safety event.

15.2 National Competency Frameworks

The AAMC[1] published national AI competency standards for medical education in 2024, defining what medical students and residents need to know about AI foundational concepts, ethical and legal implications, data literacy, and collaborative practice (Association of American Medical Colleges 2024). These competencies are organized across the learning continuum — from pre-clinical students through practicing faculty — and are the closest thing to a national standard that medical education currently has.

The AMIA[2] Informatics Workforce Roadmap defines competencies for clinical informatics specialists, extending the consumer and translator tiers into the technical domains required for production AI deployment: prompt engineering, retrieval-augmented generation architecture, model drift management, and regulatory compliance for AI-enabled devices (American Medical Informatics Association 2024).

The WHO ethics and governance guidance specifically addresses workforce reskilling as a prerequisite for responsible AI in healthcare, framing AI literacy as a component of the human oversight obligation (World Health Organization 2024). The ANA[3] has published a position statement on ethical AI in nursing practice that updates the nursing Code of Ethics to include professional accountability for AI-influenced care decisions.

[1] https://www.aamc.org
[2] https://www.amia.org
[3] https://www.nursingworld.org

Table 15.1: Summary of national AI competency frameworks for healthcare workforces. Each framework targets a distinct audience and emphasizes different competency domains. AMC workforce programs should map their training content against all relevant frameworks for the roles they serve.

Framework	Organization	Target Audience	Core Emphasis
National AI Competencies	AAMC (2024)	Medical students, residents, faculty	Foundational literacy, ethics, collaborative practice
Informatics Workforce Roadmap	AMIA (2024)	Clinical informatics specialists	Technical deployment, governance, regulatory compliance
AI Competencies for Health Professionals	AMIA (2024)	All clinical staff	Critical appraisal, safe use, documentation
Ethics and Governance of AI for Health	WHO (2024)	Healthcare institutions	Human oversight, transparency, reskilling
Ethical Use of AI in Nursing Practice	ANA (2025)	Nurses and APPs	Professional accountability, patient safety, documentation

15.3 The Clinical Human-in-the-Loop Mandate

For clinicians at the consumer tier, the most critical training component is not conceptual — it is behavioral. Clinicians need to develop and sustain the habit of genuine review, as opposed to passive acceptance, of AI-generated outputs. As discussed in Chapter 13 and Chapter 11, the automation complacency literature documents clearly that consistent accuracy leads to reduced scrutiny, and reduced scrutiny is precisely when errors slip through (Parasuraman and Manzey 2010).

Clinical AI training should include explicit instruction on the error types that AI tools make — not just their aggregate accuracy, but the specific failure modes. Ambient scribes are more likely to omit negative findings than to hallucinate findings that did not occur. Diagnostic AI in radiology tends to underperform at rare presentations and at the boundaries of its training distribution. Prior authorization agents may match coverage criteria to clinical data in ways that differ subtly from clinical judgment. A clinician who knows the specific failure modes of the tools they use is a better human-in-the-loop than one who knows only the overall accuracy figure.

Training on clinical AI should also address documentation of AI use in the medical record. The attestation of an AI-generated note is a professional claim; clinicians need to understand what they are attesting to and why the integrity of that attestation matters for both patient safety and institutional liability.

15.4 Shadow AI as a Training Priority

Shadow AI — the use of consumer AI tools with institutional data — is primarily a training failure, not a security failure. Most clinicians and administrators who route clinical data through personal ChatGPT or Gmail AI accounts do so because they do not know the risks, not because they have decided the risks are acceptable. The solution is not exclusively a security intervention (blocking consumer AI sites) but a training intervention (making the risks specific and tangible) combined with an infrastructure intervention (making the institutional alternative easier to use than the consumer alternative, as discussed in Chapter 14).

The training component of shadow AI prevention is the same consumer-tier literacy that all AMC staff need: specific understanding of what happens to data entered into consumer AI tools (it may be logged, reviewed for safety, and potentially used for model training unless enterprise terms apply), what the HIPAA implications of that exposure are, and what the institutional consequences of a PHI exposure event are for the individual employee and the institution.

Administrative staff are the highest-risk population for shadow AI not because they are less trustworthy than clinical staff, but because their tasks — summarizing reports, drafting communications, analyzing spreadsheets — are exactly the tasks for which consumer AI tools are most immediately useful, and because administrative staff frequently have access to data that is sensitive without being obviously PHI-labeled.

15.5 Measuring the Gap: The Evidence the Chapter's Claims Rest On

The assertion that clinicians overestimate their AI knowledge is not an impression — it is a documented empirical pattern with a specific shape. Studies applying validated AI literacy instruments to healthcare workers consistently find gaps of 30 to 54 percentage points between self-reported competence and objectively measured performance. The instruments that have been developed for this purpose — the Meta AI Literacy Scale (MAILS), the Scale for Non-Expert AI Literacy (SNAIL), and the AI Readiness Scale for medical students — are not perfect assessments, but they are specific enough to distinguish among competency domains and reveal where overconfidence concentrates.

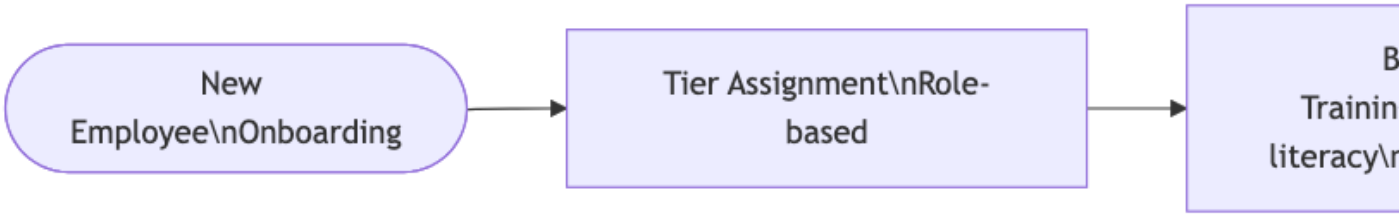


Figure 15.1: Workforce AI literacy development lifecycle. The loop structure reflects the reality that AI capabilities evolve continuously and training cannot be a one-time event.

The pattern is consistent across domains: clinicians score closer to their self-assessments on questions about AI tools they use daily, and diverge most sharply on questions about model mechanics, data governance, and failure mode recognition. The Dunning-Kruger structure matters for governance because the highest-confidence wrong answers tend to come from the consumer tier — the people using AI tools at the point of care — rather than from developers or governors. A consumer who is confident that they understand when an ambient scribe is likely to make an error is less likely to apply the verification discipline that the tool's actual failure modes require. The training program that addresses this cannot be reassurance that AI is generally reliable. It has to name the specific failure modes and require demonstrated ability to recognize them.

Survey data from 2025 has added a shadow AI dimension to this picture. A Wolters Kluwer[4] survey found that 57 percent of healthcare professionals had encountered or used unauthorized AI tools in their work (Wolters Kluwer 2025). A separate market research report found that 17 percent of front-line staff admitted to entering identifiable patient data into consumer AI tools (Black Book Market Research 2025). These numbers are almost certainly underestimates — surveys of unauthorized behavior depend on self-reporting by people who may recognize the behavior as problematic. The shadow AI problem is not primarily about bad actors. It is about people solving real workflow problems with available tools in the absence of a sanctioned alternative that works as well.

[4] https://www.wolterskluwer.com

15.6 The Accreditor Mandate: 2025 and Beyond

For the first five years of the generative AI era, AI literacy in healthcare education was largely voluntary — something institutions could pursue if they believed in it, skip if they were busy. That has changed. The 2025 to 2026 accreditation cycle has moved AI literacy from an optional enrichment to a formal program requirement in ways that are specific enough to enforce.

The ACGME[5]'s July 2025 Common Program Requirements include new language on human-AI teamwork and require programs to have institutional policies on the use of generative AI for academic work (Accreditation Council for Graduate Medical Education 2025). The specificity is meaningful: it is not "programs should be aware of AI" but "programs must have policies." Programs without documented AI use governance are out of compliance with the new requirements starting with the 2025 to 2026 accreditation year.

The LCME[6]'s current standards for medical school accreditation interpret Standards 7.6 and 8.2 — covering bias and self-directed learning — to require critical appraisal of AI outputs as a competency that medical students must demonstrate. The AAMC's 2025 AI Competencies Across the Learning Continuum, developed through a formal Delphi process, provides the national standard for what those competencies should include across undergraduate, graduate, and continuing medical education (Association of American Medical Colleges 2025).

For nursing programs, the American Nurses Association's 2025 update to the Code of Ethics explicitly addresses machine learning: Provision 7.5 states that AI is integral to nursing practice, and Provision 4.2 clarifies that nurses retain final accountability over AI-influenced care decisions. The update is significant because it moves AI governance from an informatics specialty issue to a professional obligation for every practicing nurse.

The ACCME[7] has issued guidance on the responsible use of AI in accredited continuing education, establishing disclosure requirements and data handling standards for CME programs that use AI-generated content or AI-assisted learning tools (Accreditation Council for Continuing Medical Education 2025). For AMC CME offices that have adopted AI for content development, this is an immediate operational compliance issue, not a future planning item.

The practical implication is that workforce development for AI is no longer a discretionary investment. Programs that do not have documented, accreditor-compliant AI policies and competency-based training in place are at accreditation risk. The AMC that treats this as an IT initiative rather than an academic and clinical leadership priority will discover the error during its next site visit.

[5] https://www.acgme.org
[6] https://lcme.org
[7] https://www.accme.org

15.7 Building a Living Curriculum

The half-life of specific AI technical knowledge is short. A training module written around GPT-4's capabilities in early 2023 was already outdated within six months. A curriculum built around the tools of 2025 will require revision by 2026. This is not a reason to avoid building curricula — it is a design constraint that distinguishes AI literacy training from other required training programs.

The institutions that have built durable AI literacy programs share a structural characteristic: they design for modularity and update cadence from the beginning, rather than treating each module as a finished product. Stanford Medicine's approach treats AI literacy content as living documentation — updated when model capabilities change materially, not on an annual review cycle. Mayo Clinic's Harper Family Foundation AI Education Program uses micro-credentialing to recognize competency at specific points in time, acknowledging that the credential represents "AI literacy as of this date" rather than a permanent certification.

The modular structure that makes curricula updatable also makes them role-specific without requiring separate parallel programs. Consumer-tier content covers the concepts and failure modes that every staff member needs. Translator-tier content goes deeper on model evaluation and governance processes. Developer and governor tiers go deeper still on technical architecture and risk management. When the underlying technology changes, the consumer module updates; the translator module updates to a different depth; the developer module updates to reflect new architectural patterns. The institution does not rebuild four programs — it updates the four tiers of one program.

The CME and CNE infrastructure required to sustain a living curriculum is not trivial to build, but it is not novel either. The ACCME's guidance on AI in continuing education provides a framework for awarding credit for AI literacy content that meets the disclosure and data handling requirements for accredited CME. An AMC that routes its AI literacy modules through its existing CME infrastructure — with appropriate documentation of AI-assisted content development where that applies — can offer credit for the training that the accreditation mandates now require, without building a separate credentialing system.

15.8 The Faculty Development Gap

The most significant systemic bottleneck in AMC AI workforce development is the faculty development gap: the clinicians and educators responsible for training the next generation cannot teach AI literacy they do not have themselves. Faculty who were trained before large language models existed, and who have not had protected time or institutional support for AI education, are not positioned to integrate AI literacy into medical education curricula.

Addressing this gap requires institutional investment in faculty development specifically for AI, not as a one-day workshop but as a sustained program. The most effective models combine

peer learning — faculty who have developed AI competence mentoring colleagues — with micro-credentialing that provides formal recognition of AI literacy achievement and creates a career incentive for the investment. The AAMC and AMIA have both published frameworks for faculty AI development that AMC education offices can adapt.

15.9 Where to Start

15.9.1 Starter Project 1: Role-Based AI Literacy Module Deployment

What it is: Mandatory role-based AI literacy training for all clinical and administrative staff, tiered by role (consumer, translator, governor), with completion required within 90 days of the launch and annually thereafter.

Why now: Colorado SB 24-205 requires that staff using high-risk AI in healthcare understand the tool's purpose and limitations. More broadly, no AMC AI governance program can operate effectively if the staff expected to apply governance controls do not understand why they exist.

How to execute: The AAMC, AMIA, and AMA STEPS Forward have all published training content that can be adapted without building from scratch. Consumer-tier training should cover: what is an LLM, what are its failure modes, what is shadow AI and why does it matter, how to document AI use. Translator-tier training should add model evaluation, bias audit interpretation, and regulatory basics. Governor-tier training should cover risk governance, vendor evaluation, and board-level AI accountability. Track completion in the LMS and include AI literacy completion as a metric in the annual AI governance report.

Buy vs. build: Primarily adapt and configure existing content. Build additional modules for institution-specific tools and institutional policies.

15.9.2 Starter Project 2: AI Champions Program

What it is: A formal program that identifies, trains, and supports clinician AI champions across departments — translators who can bridge between the informatics team and bedside clinical practice, help colleagues adopt new AI tools, and surface governance concerns from the frontline.

Why now: Adoption of clinical AI tools without champions in the departments is slower, and the governance feedback loop from frontline use back to the AISC is weaker. Champions are the mechanism by which governance reaches the point of care.

How to execute: Identify one champion per department or service line. Provide advanced translator-tier training plus explicit instruction in governance processes and escalation paths. Give champions protected time (one to two hours per week) for the AI champion role. Connect

champions to each other through a community of practice that shares experience across departments. Route frontline governance feedback — anomalous AI outputs, patient concerns, workflow problems — from champions to the AI Steering Committee on a regular cadence.

Buy vs. build: Program design and staff time. No technology purchase required.

16 Ethics, Equity, and Institutional Accountability

The ethical challenges of AI in the AMC are not primarily about individual decisions by individual clinicians. They are structural. A predictive model that systematically underestimates the health needs of Black patients does not fail because the clinician using it is biased; it fails because it was trained on data that encodes decades of inequitable access to care, and deployed without monitoring that would detect the systematic underperformance (Obermeyer et al. 2019). An ambient documentation system that malfunctions differently for patients with non-standard accents does not fail because the clinician was careless; it fails because the model was trained on data that overrepresented certain speech patterns and deployed without demographic performance stratification. The pattern is consistent: the ethical failures that have actually occurred in deployed healthcare AI are structural failures, predictable in advance, and correctable through governance — if the governance exists.

This chapter argues that AMC AI ethics requires a structural turn. The question is not "does this AI tool respect individual patient autonomy?" but "does the process by which this institution deploys, monitors, and governs AI tools systematically protect patient equity and institutional accountability?" Individual ethical review is necessary but not sufficient. Structural governance is the mechanism through which individual ethical commitments become institutional practice.

16.1 Algorithmic Bias as a Structural Problem

The most cited demonstration of algorithmic bias in healthcare involves a commercial risk stratification algorithm used by major health systems to identify high-risk patients who would benefit from care management programs (Obermeyer et al. 2019). The algorithm used healthcare costs as a proxy for health need — a reasonable proxy if access to care were uniformly distributed, which it is not. Black patients with the same actual health burden as white patients had systematically lower costs, because they had systematically less access to care. The algorithm therefore scored them as lower risk, directing care management resources away from patients who needed them more.

The finding was not that the algorithm was malicious. It was that the algorithm was trained on data that encoded an existing inequity, using a proxy variable that faithfully reproduced that inequity at scale. The authors estimated the bias caused the algorithm to miss 43% of high-risk

Black patients compared to a race-neutral approach (Obermeyer et al. 2019). Subsequent work has documented similar patterns in algorithms for kidney disease (using race as a correction factor), cardiac risk assessment, and dermatology imaging models (Zack et al. 2024).

The structural governance response requires three elements that individual ethical review cannot provide alone. First, demographic stratification of performance metrics as a standard part of AI validation — not an optional audit, but a required table in every model evaluation. Second, monitoring that continues after deployment, because bias can emerge over time as the population served changes or as the model is applied to use cases outside the validation set. Third, a reporting structure that routes performance stratification findings to clinical leadership and the governance committee, not just to the informatics team.

16.2 Health Equity as a Performance Metric

The framework proposed by Badal and colleagues, introduced in Chapter 6, includes the alleviation of health disparities as the first principle of responsible clinical AI. In practice, this requires operationalizing "equity" as a quantitative performance dimension alongside accuracy. A model that achieves 85% accuracy overall but 70% accuracy for the subpopulation with the highest disease burden is not a high-performing model — it is a model that performs best for the patients who need it least.

The practical implementation of equity monitoring requires demographic data in the validation and monitoring datasets. This is straightforward in principle and difficult in practice, because demographic data in EHRs is often missing, inconsistent, or coded in ways that do not capture the granularity needed for subgroup analysis. AMCs that serve diverse populations should treat demographic data quality as an AI readiness issue: the return on investment for improving race, ethnicity, and language data completeness comes precisely when the institution needs to assess whether its AI tools are performing equitably.

The FUTURE-AI[1] principles for trustworthy AI in medical imaging establish fairness as one of six foundational requirements, alongside universality, traceability, usability, robustness, and explainability (Lekadir et al. 2022). These principles map to operational AMC practices: fairness requires subgroup performance evaluation; traceability requires audit logging of model outputs; explainability requires that clinicians can access the reasoning behind a model recommendation. The NIST AI RMF (National Institute of Standards and Technology 2023) maps these same principles to the Govern/Map/Measure/Manage operational structure that AMC governance programs can implement directly.

[1] https://future-ai.eu/

16.3 Informed Consent in the Continuous AI Era

The consent framework governing clinical AI does not fit the technology. The traditional consent model is episodic and discrete: a patient is informed about a specific intervention at a specific moment and chooses whether to accept it. Clinical AI tools are continuous and ambient: a predictive model runs on every patient in the database, continuously, updating as new data arrives. The patient is not present when the model generates a risk score; there is no discrete moment at which they can meaningfully consent or decline.

Ambient documentation creates a more visible consent challenge — the patient is in the room when the AI is active — but the design solution is tractable. The verbal consent model described in Chapter 13 and Chapter 12 provides explicit, encounter- specific authorization. The harder case is the background AI: the readmission risk model, the sepsis predictor, the no-show prediction algorithm. These tools affect every patient encounter without any patient-facing disclosure.

The emerging consensus from bioethics and patient advocacy is that "invisible" background AI requires institutional-level disclosure rather than individual patient consent: the institution publishes a clear statement of which AI tools are used in patient care, how they affect clinical decisions, and what patients can do to learn more. This does not resolve the ethical tension between efficient deployment and individual autonomy, but it addresses the transparency gap. The WHO ethics guidance on AI for health specifically recommends meaningful transparency about automated decision-making as a prerequisite for the ethical deployment of clinical AI (World Health Organization 2024).

16.4 Intellectual Property and the AI Authorship Gap

The U.S. Copyright Office and the U.S. Patent and Trademark Office have both taken consistent positions that AI cannot be an author or an inventor; copyright and patent protection requires human creative contribution. For AMCs, this creates several practical IP implications.

Research outputs that are substantially AI-generated — whether grant applications, academic papers, or clinical protocol drafts — may not be protectable under copyright if the human contribution is insufficient. The ICMJE standards for medical journal authorship state clearly that AI systems cannot be listed as authors and that authors are responsible for the integrity of AI-assisted content (International Committee of Medical Journal Editors 2023). This places the accountability for AI-generated research content squarely on the human authors, including accountability for errors, hallucinations, and fabricated citations introduced by AI tools.

For clinical documentation, the liability implication is parallel: the clinician who attests to an AI-generated note takes professional responsibility for its accuracy. The attestation is not a rubber stamp; it is a claim that the clinician has reviewed and accepts the content as their professional documentation of the encounter. Governance policies and training programs should

be explicit about this responsibility, because the ease and speed of AI-assisted documentation can inadvertently erode the clinician's sense of authorship and accountability.

16.5 The Regulatory Turn: HHS Section 1557 and the Duty to Mitigate

For most of the 2010s, AI equity was a governance aspiration — a principle that showed up in AMC values statements and academic papers but carried no specific legal obligation. The 2024 HHS Section 1557 final rule changed that (U.S. Department of Health and Human Services, Office for Civil Rights 2024a). Under 45 C.F.R. § 92.210, covered entities are required to take reasonable steps to identify and mitigate discrimination in patient care decision support tools that are used to make, recommend, or facilitate clinical decisions. The rule explicitly covers algorithmic and AI-assisted tools. The compliance deadline for covered entities was May 2025.

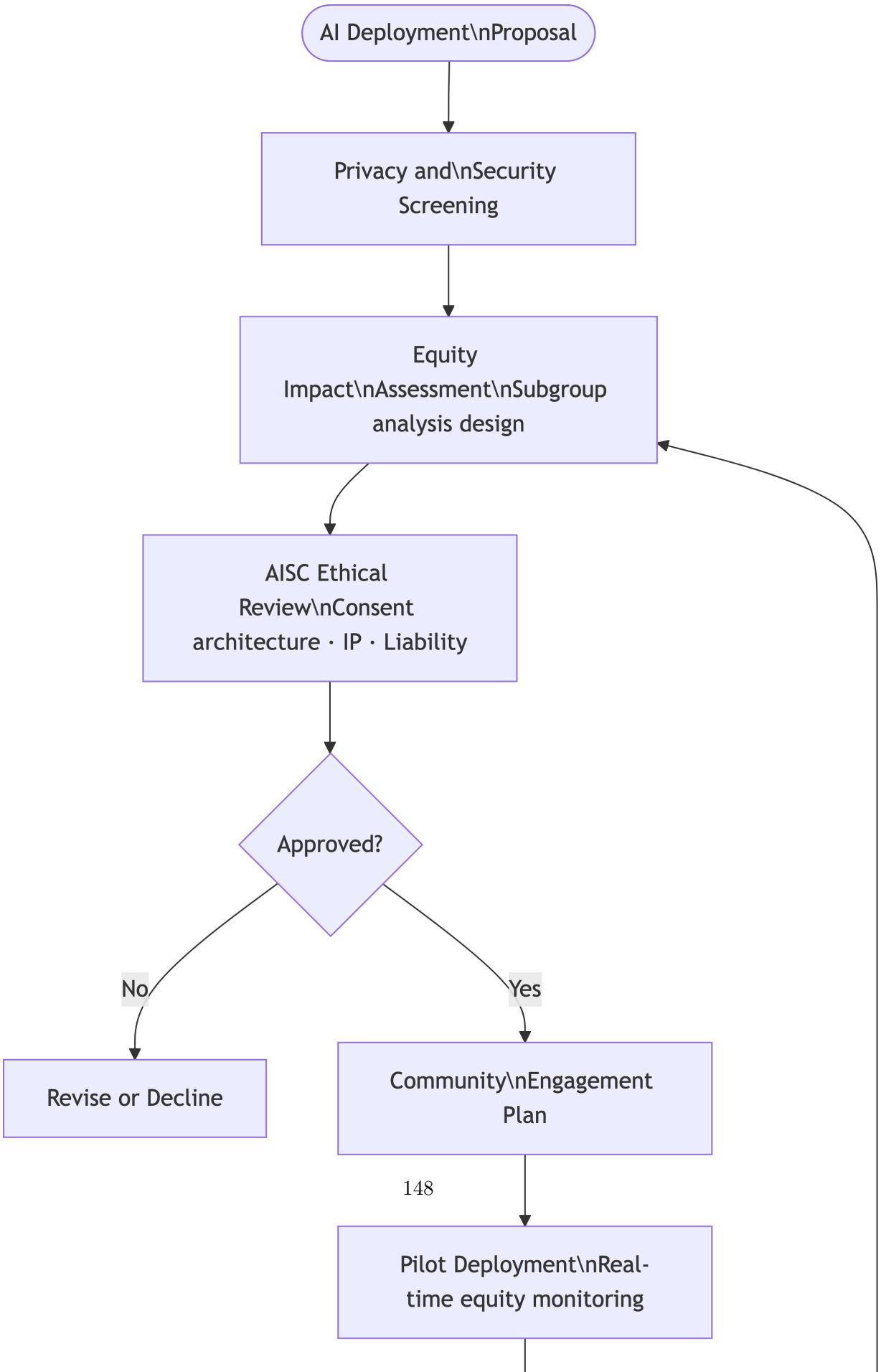
What "reasonable steps" means in practice is not fully defined by the rule, but the regulatory record is instructive. HHS explicitly cited the Obermeyer 2019 risk stratification algorithm as the paradigmatic case the rule is designed to address. The implication is that an institution deploying a care management algorithm — a readmission risk model, a care gap identification tool, a utilization management system — without having assessed its performance across demographic groups cannot demonstrate compliance. The assessment does not need to be a clinical trial; it needs to be documented evidence that someone looked. The equity audit process in Section 16.11 is what that documented evidence looks like.

The risk in Section 1557 is not just regulatory penalty. It is reputational and evidentiary. If a patient files a discrimination complaint and the institution cannot produce documentation that it evaluated whether its AI tools affected that patient's demographic group differently, the absence of documentation is itself evidence of unreasonable practice. Building the equity audit function is not a compliance checkbox — it is the institutional record that will matter when a complaint arrives.

16.6 Beyond Obermeyer: Recent Cases of Algorithmic Bias

The Obermeyer 2019 finding — that a commercial risk stratification algorithm used a cost proxy that systematically underestimated the health needs of Black patients — is the most cited demonstration of algorithmic bias in healthcare, and it risks becoming a comfortable historical example that lets institutions off the hook for examining what their own deployed tools are doing right now.

The 2022 to 2025 literature documents the pattern continuing. Daneshjou and colleagues demonstrated that dermatology imaging AI performs substantially worse on images of patients


AI Deployment\nProposal
Privacy and\nSecurity Screening
Equity Impact\nAssessment\nSubgroup analysis design
AISC Ethical Review\nConsent architecture · IP · Liability
Approved?
No
Yes
Revise or Decline
Community\nEngagement Plan
Pilot Deployment\nReal-time equity monitoring

with darker skin tones, a predictable consequence of training datasets that overrepresented lighter skin phenotypes. Ambient documentation tools have been found to have higher error rates for patients with non-standard accents — specifically omitting nuances in social history that require parsing speech patterns underrepresented in the training data. A 2024 Senate investigation documented that AI-assisted care denial systems used by Medicare Advantage insurers produced denial rates up to 16 times higher than human review in post-acute care — with the automation of denial decisions structured to make human review functionally impossible at the volumes the AI generated (U.S. Senate Committee on Homeland Security and Governmental Affairs 2024).

The ProPublica[2] investigation into Cigna's PxDx system described physicians reviewing AI-generated denial recommendations at a rate of 1.2 seconds per claim (Kirchner and Waldman 2023). That is not human- in-the-loop review. It is human-in-the-loop theater. For an AMC that uses AI-assisted prior authorization or utilization management tools, the governance question is not whether those tools carry demographic bias — they almost certainly do, to some degree — but whether the human review process is substantive enough to catch and override it when it manifests.

16.7 State Privacy Laws and the Post-HIPAA Landscape

HIPAA remains the dominant privacy framework for clinical AI, but it is no longer sufficient as a complete governance guide. A patchwork of state laws has emerged in the 2022 to 2025 period that creates obligations for AI use at AMCs that operate in, or serve patients from, specific states.

Washington's My Health MY Data Act, effective 2023, regulates consumer health data that falls outside HIPAA's scope — data collected by apps, wellness tools, and AI systems that are not covered entities (Washington State Legislature 2023). It requires a separate opt-in consent for collection and sharing, imposes restrictions on data retention, and bans geofencing around healthcare facilities, which has implications for location-based AI tools and mobile health applications. Because the Act's definition of "consumer health data" is broad enough to capture AI-generated health inferences, an AMC deploying patient-facing AI tools that touch Washington residents needs to analyze the Act's requirements specifically.

Colorado HB 24-1139, signed in 2024, prohibits health insurers from using AI as the sole basis for an adverse medical determination, requiring that all AI-generated denial recommendations receive substantive review by a qualified clinician (Colorado General Assembly 2024). The bill explicitly addresses the pattern documented in the Senate Medicare Advantage investigation. For AMCs that operate health plans or manage care programs with AI-assisted utilization management, this is a direct compliance obligation in Colorado.

[2] https://www.propublica.org

Illinois BIPA's healthcare exemption — clarified in a 2023 Illinois Supreme Court ruling — exempts biometric data used for healthcare treatment, payment, or operations from the Act's consent and notice requirements. This is relevant for AMCs using ambient audio, retinal scans, or other biometric identification in clinical workflows. The exemption is narrower than it appears; biometric data used for security access, time and attendance, or administrative identification may not fall within the healthcare exemption.

The broader pattern is that HIPAA compliance is a floor, not a ceiling. Each state where an AMC operates patients, employs staff, or deploys patient-facing digital tools may impose additional requirements on AI-related data handling, consent, and human review. The institutional legal review process for AI deployments needs to include state law analysis, not just HIPAA review.

16.8 The Workforce and Labor Dimension

An ethics chapter about AI in the AMC that does not address what happens to the people whose work AI changes is incomplete. The institutional ethics question here is not whether to deploy AI tools that make some existing roles redundant — that is already happening — but how the institution manages the human consequences of that displacement.

The roles most directly affected by AI automation in the current wave are not clinical roles requiring complex judgment. They are roles involving high-volume, structured, repetitive cognitive work: medical coders whose work is partially automated by AI-assisted coding tools; prior authorization specialists whose decisions are increasingly pre-populated or reviewed by AI; transcriptionists who have seen their role transformed or eliminated by ambient documentation; certain radiology reading functions where AI handles high-volume, lower-complexity cases.

The institution that deploys AI tools that reduce the need for these roles without an explicit workforce transition program — retraining, reassignment, severance, outplacement — is making an ethical choice, whether or not it acknowledges it as one. AMCs that have invested in the relationships with their frontline staff that make clinical quality possible should not treat AI-driven workforce changes as a pure efficiency calculation. The social compact that allows an AMC to function as a clinical and community institution is relevant to how it manages the people affected by AI-driven change, not just to how it treats patients.

16.9 Community Trust and the Social License to Deploy

Healthcare AI operates not just within a regulatory framework but within a social one. Patients have expectations about how their health data is used, how AI figures in their care, and what control they retain over algorithmic decisions that affect them. Those expectations are not uniformly positive, and they are not uniformly distributed.

Survey data from 2024 to 2025 shows that patient trust in AI-assisted healthcare varies significantly by demographic group, with Black and Hispanic patients expressing more skepticism about clinical AI than white patients in multiple studies. This asymmetry is not irrational — it reflects historical experience with healthcare systems that produced the very biases that AI tools now replicate at scale. An institution that deploys AI tools with demonstrated demographic performance disparities, in service of a patient population that has historically been underserved, and then frames the resulting errors as "algorithmic" rather than institutional, is trading on trust it may not have fully earned.

The response to this is not to delay AI deployment until trust is perfect. Trust is built through transparency and accountability in practice, not in advance of it. The institutional mechanisms that build social license are the same mechanisms that the governance chapters elsewhere in this book require: meaningful disclosure about which AI tools are used and how they affect care decisions, community engagement in AI governance processes, an equity audit process that reports findings publicly, and a willingness to suspend tools that produce harm even when the RO business case for them is positive.

The Coalition for Health AI[3]'s 2024 patient trust survey found that 51 percent of patients reported trusting healthcare less due to AI, but 80 percent said their trust would increase if they knew who was accountable for the AI's decisions and that training on those decisions was documented (Coalition for Health AI 2024). The accountability and training documentation that this book's governance chapters describe are not just operational mechanisms. They are the substance of the social license that AMC AI requires to function.

16.10 Liability, the Standard of Care, and the Duty to Use

The liability landscape for clinical AI is developing, not settled, but the direction is clear in both directions. Clinicians and institutions face potential liability both for harms caused by following AI recommendations without adequate oversight, and — as AI tools become validated for specific clinical tasks — for failing to use tools that have become part of the standard of care.

The second direction is counterintuitive but increasingly argued in the literature. As AI tools for specific diagnostic tasks — retinal disease screening, dermatology imaging, sepsis prediction — accumulate evidence of performance at or above specialist-level accuracy, the ethical argument for using them begins to shade into a professional obligation (Zemmar et al. 2023). A radiologist who does not use an FDA-cleared AI tool to detect pneumothorax, when the tool has demonstrated sensitivity superior to unassisted reading, may eventually face liability for the missed diagnosis.

The prudent governance posture is to document, for each deployed clinical AI tool, the institutional reasoning about when and how it should be used — not just the existence of

[3] https://www.coalitionforhealthai.org

the tool, but the clinical judgment about its appropriate role. When a clinician overrides an AI recommendation, that decision should be documentable. When a clinician relies on an AI recommendation, that reliance should be documentable. The medical record is the primary liability defense; it should reflect the clinician's engagement with AI tools, not hide it.

16.11 Where to Start

16.11.1 Starter Project 1: Equity Audit of Deployed Clinical AI

What it is: A structured retrospective audit of performance stratification for the two or three highest-impact clinical AI tools currently deployed, assessing whether performance metrics vary significantly by race, ethnicity, age, insurance status, and language.

Why now: HHS Section 1557 requires that covered entities not deploy discriminatory patient care decision-support tools. The section 1557 final rule is in effect. An institution that has not assessed its clinical AI tools for demographic performance variation cannot certify compliance, and more importantly, cannot know whether its tools are harming the patients most at risk.

How to execute: Work with the clinical informatics team to extract retrospective performance data for each tool, stratified by available demographic dimensions. Identify subgroups with statistically significant performance differences. Assess whether the difference is clinically meaningful and whether it reflects a correctable bias in the model or an irreducible clinical population difference. Report findings to clinical leadership and the governance committee. For tools with significant performance disparities, develop a remediation plan.

Buy vs. build: Analytical work using existing institutional data. Commercial bias audit tools (Credo AI, IBM OpenScale) can accelerate the analysis but are not prerequisites.

16.11.2 Starter Project 2: Clinical AI Ethics and Accountability Policy

What it is: A published institutional policy on the ethical deployment of clinical AI that addresses the four structural elements described in this chapter: equity monitoring requirements, consent architecture for background AI, IP and authorship accountability, and documentation requirements for AI-assisted clinical decisions.

Why now: Without a published policy, there is no institutional standard to hold deployments to, no governance anchor for the ethics review pipeline in Figure 16.1, and no document to point to when a patient asks why an AI tool was used in their care.

How to execute: Draft using the NIST AI RMF as the governance scaffold and the FUTURE-AI principles as the technical requirements framework. Review with legal (liability and IP), compliance (Section 1557 and Colorado SB 24-205), clinical leadership (standard of care

implications), and patient representatives (consent and disclosure language). Publish as institutional policy with a defined review cycle aligned with the annual AI governance report.

17 Data Access and Governance

The most common bottleneck in AMC AI deployment is not model selection, compute availability, or vendor relationships. It is data governance. Not in the abstract sense — every institution has HIPAA policies — but in the specific sense that most AMC data governance frameworks were designed for uses that look nothing like LLM training or inference. The HIPAA Safe Harbor de-identification standard was written to support data sharing for research and public health uses. It was not designed to account for a model that can infer a patient's identity from the statistical patterns in clinical notes after every explicitly identifying field has been removed. Understanding why existing frameworks are insufficient, and what a governance structure adequate to the LLM era requires, is the starting point for sound data strategy.

17.1 The AMC Data Mosaic

An academic medical center holds data that falls under at least four distinct regulatory regimes, and AI pipelines frequently cross the lines between them without anyone noticing. Clinical data generated by patient care is governed primarily by HIPAA. Student and trainee evaluations are governed by FERPA, which prohibits disclosure of educational records without consent. Research data involving human subjects is governed by the Common Rule (45 CFR 46) and, for federally funded genomic data, by additional NIH consent and data sharing requirements. Administrative and business data is governed by institutional policy and non-disclosure obligations.

The governance failure happens when a researcher asks an LLM to analyze de-identified clinical notes alongside trainee evaluation records and administrative email threads, passes all of it into a proprietary API with a standard enterprise agreement, and calls the result compliant because no single data type was explicitly prohibited. The pipeline crosses four regulatory regimes, creates potential for cross-dataset re-identification, and sends everything to an external service with whatever data retention policies the standard enterprise agreement allows. This is not a hypothetical; it is a pattern that occurs regularly at institutions without explicit AI data governance policies.

17.2 Data Classification for AI

A data classification framework designed for the LLM era needs at least four tiers that account not just for the sensitivity of the data in isolation but for its sensitivity in combination with the capabilities of the model processing it.

Public data can be used freely, including with public AI services: institutional press releases, public-facing policy documents, de-identified aggregate statistics.

Internal data includes non-PHI business data — operational metrics, general financial information, non-patient-facing communications. This data should be processed only through services with a signed data processing agreement, but does not require a HIPAA BAA. Many institutions use consumer-grade AI tools for internal data tasks without adequate contractual protections; this is a gap the classification framework should close.

Regulated data is the primary risk tier: PHI under HIPAA, student records under FERPA, and individually identifiable research data under the Common Rule. This data requires a BAA or equivalent privacy agreement, zero-data-retention provisions in vendor contracts, and where possible, local processing rather than transmission to external services.

Restricted data is the highest tier: genomic sequences, psychiatric treatment records, HIV status, substance use treatment records, and data subject to specific research consent restrictions. NIH policy has begun restricting the export of AI models trained on certain genomic datasets, recognizing that the model weights themselves can encode identifying information. Restricted data should not leave the institutional network boundary without specific governance approval.

17.3 The Limits of De-identification

The HIPAA Privacy Rule defines two de-identification methods. Safe Harbor removes 18 specific categories of direct identifiers. Expert Determination uses a statistical assessment to verify that the residual re-identification risk is very small (U.S. Department of Health and Human Services, Office for Civil Rights 2012). In 2012, when these standards were codified, the primary risk cases were database linkage attacks using public records. The threat model has changed materially since then.

Gymrek and colleagues demonstrated that genomic data de-identified under Safe Harbor could be re-identified using public genealogical databases and surname inference — using techniques available to any moderately sophisticated analyst (Gymrek et al. 2013). The threat model for clinical notes has evolved similarly. Models trained on large corpora of clinical text can memorize rare clinical strings — unusual diagnoses, distinctive procedure sequences, specific medication combinations — and reproduce them under adversarial prompting. A note that contains no direct identifiers may still contain a clinical signature unique enough to identify a specific patient to someone with access to supplementary information.

The governance implication is direct: Safe Harbor is not an adequate privacy guarantee for LLM training datasets. Expert Determination — a statistical verification that the specific data in the specific modeling context poses negligible re-identification risk — should be the default for AI training on clinical data. For inference, the appropriate safeguard is a BAA with zero-data-retention provisions that prevent the model provider from retaining prompt content after the session concludes.

17.4 The AI-Ready Honest Broker

Most AMCs have an existing honest broker function: a person or team that processes data requests for research use, applies de-identification, and manages sharing agreements. This function was designed for structured data exports. It was not designed for the AI-era pattern: unstructured text, imaging data, multi-modal inputs, and requests for ongoing access to live data streams rather than static exports.

The AI-ready data broker function needs to add several capabilities. It needs the ability to assess re-identification risk for unstructured text, not just structured fields. It needs to evaluate vendor BAAs specifically for AI provisions — standard BAAs written before 2022 frequently do not address model training, prompt retention, or output logging. It needs to maintain an approved tool registry mapping permitted data types to approved services, so that individual researchers do not make case-by-case risk assessments without institutional guidance.

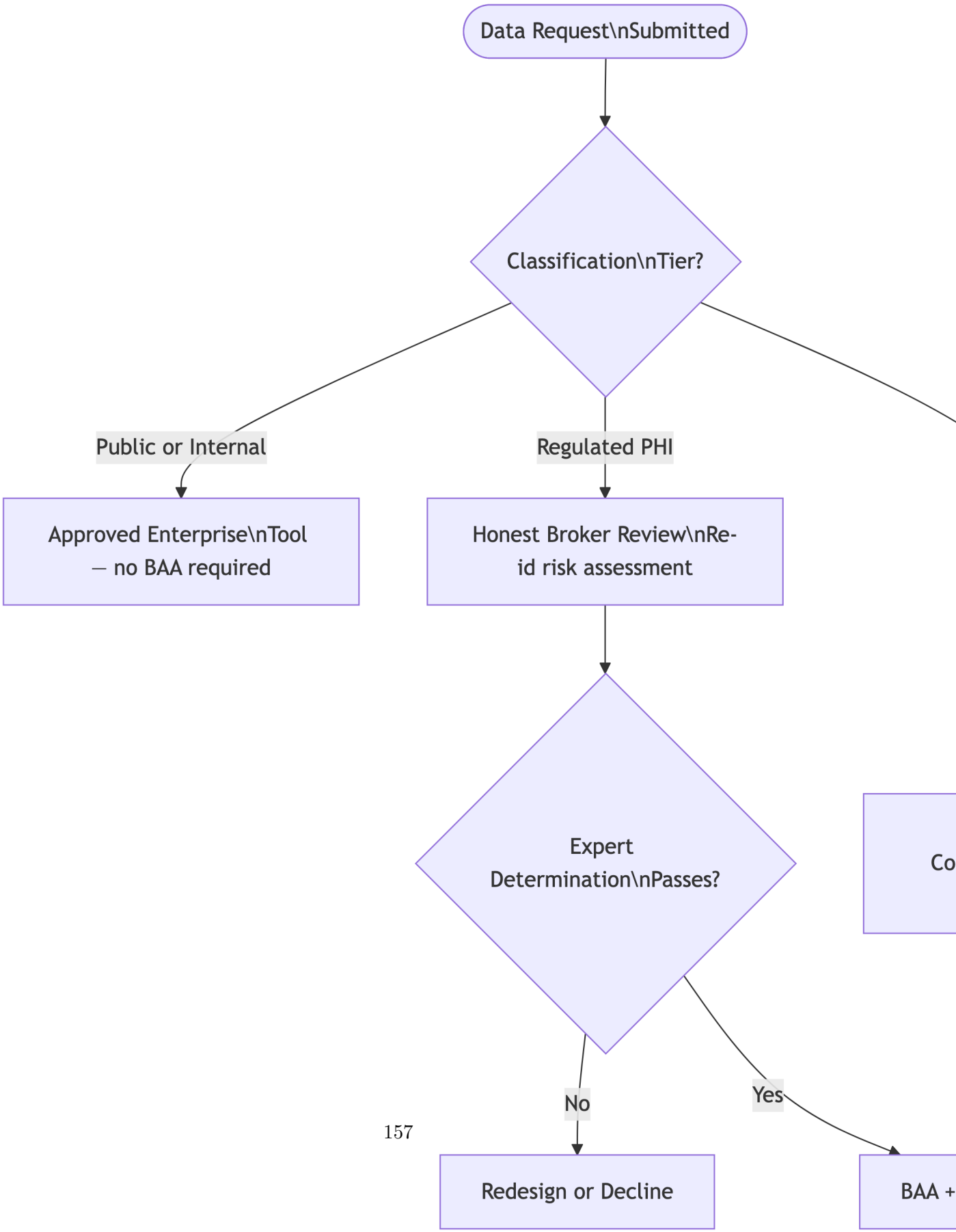
17.5 FHIR and OMOP as AI Substrate

The interoperability standards that AMCs have adopted for data exchange — HL7 FHIR[1] and the OMOP Common Data Model[2] — are not just data formats. They are the substrate on which clinical AI is increasingly built and validated.

FHIR R5 provides standardized resource definitions for clinical data that allow AI models to query, retrieve, and write structured information across different EHR implementations. The SMART on FHIR authorization framework enables granular, patient-specific access scopes that are essential for the least-privilege agentic system design described in Chapter 11. For AMCs procuring AI tools that need EHR interaction, FHIR compatibility is increasingly the minimum requirement — both for richer data access and for the access logging that governance requires.

The OMOP CDM is the dominant standard for large-scale observational research, and it has become the basis for multi-institutional AI validation (Singhal et al. 2023). A model validated on OMOP-formatted data from one institution can be tested against OMOP-formatted data

[1] https://hl7.org/fhir/
[2] https://www.ohdsi.org/data-standardization/the-omop-common-data-model/


Data Request\nSubmitted
Classification\nTier?
Public or Internal
Regulated PHI
Approved Enterprise\nTool — no BAA required
Honest Broker Review\nRe-id risk assessment
Expert Determination\nPasses?
Co
No
Yes
Redesign or Decline
BAA +

from another without custom integration work. For AMCs participating in research consortia — NIH N3C, PCORnet, NIH Bridge2AI — OMOP compatibility is typically required.

17.6 Vendor Contracts: The Non-Negotiables

The BAA for any AI vendor processing PHI must address provisions that standard HIPAA BAAs frequently omit. AMC legal teams should require:

No-training clause: The vendor may not use AMC data — prompts, completions, user interactions — to train or fine-tune models. This is standard in enterprise tiers from major providers but absent in default tiers; it must be explicitly verified.

Zero-data-retention for prompts: Prompt content containing PHI must not be retained after the session concludes. This means no logging for safety monitoring involving human review, and no storage in training pipelines.

Output ownership: Clinical notes and analyses generated from institutional data are institutional property. Vendor agreements should not claim ownership of AI-generated outputs based on institutional inputs.

Algorithmic change notification: The vendor must notify the institution before making significant model changes that could affect clinical outputs. An unannounced model update in a clinical workflow integration is a safety event.

Table 17.1: Non-negotiable BAA provisions for AI vendors processing clinical data. Standard HIPAA BAAs pre-dating 2023 frequently omit the first three rows.

Contract Provision	Risk Addressed	Negotiability
No-training on customer data	PHI memorization in model weights	Mandatory
Zero-data-retention for prompts	PHI leakage through session logging	Mandatory
Output ownership by institution	IP rights to AI-generated clinical content	Mandatory
Right to audit	Verify data handling compliance	Required for high-risk use
Algorithmic change notification	Unexpected behavior change in clinical AI	Required for clinical integration
US-only data residency	Jurisdictional compliance	Required for restricted data

17.7 TEFCA and the Nationwide Exchange Layer

The Trusted Exchange Framework and Common Agreement[3] — TEFCA — established a network of Qualified Health Information Networks[4] (QHINs) for nationwide clinical data exchange when it went live in 2023 (Office of the National Coordinator for Health Information Technology 2023). Epic's Nexus Health Network, Health Gorilla[5], Oracle Health, and a handful of other organizations have received QHIN designation, creating for the first time a governed national pipe through which clinical data can flow across institutional boundaries.

The implications for AI data access are significant, and more complicated than they appear. TEFCA defines a set of permitted Exchange Purposes — Treatment, Payment, Health Care Operations, Public Health, Individual Access, and a handful of others — that determine under what conditions data can be queried across the network. Research is notably absent from the permitted purposes for nationwide exchange in the current framework. An AMC that wants to use TEFCA-accessed data to train or validate an AI model is operating in legally ambiguous territory unless it can characterize the use as Health Care Operations, a framing that has a defined statutory meaning and does not extend to all AI development activities.

The practical implication is that TEFCA is most immediately useful for clinical AI applications that operate at the point of care — retrieving a patient's medication history from an external health system to inform a clinical decision, for example — and least immediately useful for the large-scale data aggregation that model training requires. AMCs participating in AI research consortia will need to route their data sharing through existing research frameworks — IRB oversight, data use agreements, OMOP standardization — rather than through TEFCA's exchange infrastructure, at least until research is added as a permitted exchange purpose.

The QHIN connection that TEFCA requires does, however, create an infrastructure opportunity for AI governance that did not exist before: every query that passes through a QHIN- connected endpoint is logged. For the first time, the institution has a national-scale audit trail for the external data it accesses. That audit trail is an asset for the AI governance program, providing evidence of the provenance of external data inputs to clinical AI systems in a way that was previously unavailable.

17.8 The NIH Data Management and Sharing Policy Tension

NIH's Data Management and Sharing Policy, which took effect in January 2023, requires that all research conducted with NIH funding produce a data management plan and share the resulting scientific data to the extent permitted by law and subject to privacy and ethical constraints (National Institutes of Health 2020). The policy was written to address the reproducibility

[3] https://www.healthit.gov/topic/interoperability/trusted-exchange-framework-and-common-agreement-tefca
[4] https://rce.sequoiaproject.org/qhin/
[5] https://www.healthgorilla.com

crisis in biomedical science — if every investigator shared their data, more findings could be independently verified. In the AI context, it creates a specific governance problem: what counts as "scientific data" when the research involves training or fine-tuning a language model?

NIH clarified in a 2025 notice that AI models trained on controlled-access genomic data are Data Derivatives subject to the same access restrictions as the underlying data (National Institutes of Health 2025). The model weights — the billions of numerical parameters that encode what the model learned — may contain information about the training data that must not be shared openly. This is the "model weights as PHI" problem: a model trained on genomic data from a controlled-access cohort may memorize rare clinical signatures in ways that allow adversarial reconstruction of individual patient data, and releasing that model is equivalent to releasing a derivative of the controlled-access dataset.

The resulting tension is not abstract. An investigator conducting NIH-funded research on a clinical AI model is simultaneously obligated to share scientific data (by the DMS Policy) and obligated not to share data that could enable re-identification (by HIPAA, IRB consent terms, and the 2025 genomic AI notice). The resolution requires distinguishing among what must be shared, what may be shared, and what must not be shared. Model documentation — model cards, performance reports, training data descriptions — can satisfy much of the transparency obligation without releasing the weights themselves. Code and analysis scripts can be shared. The trained model weights require a case-by-case governance determination that should involve the institution's research compliance office, not the individual investigator.

17.9 Synthetic Data as a Governance Instrument

One response to the tension between data access and privacy protection is to replace real clinical data with synthetic data — artificial patient records that preserve the statistical structure of the original data without containing information about any real individual. The technology for generating high-quality synthetic clinical data has matured substantially since 2020. Variational autoencoders, generative adversarial networks, and diffusion models have all been applied to EHR data generation, with recent methods producing synthetic records that are nearly indistinguishable from real records on most clinical research tasks.

The governance value of synthetic data is that it allows the institution to provide data access — for model development, testing, algorithm validation, and education — at a risk level substantially lower than real PHI. A developer testing a new clinical NLP pipeline does not need real patient data to verify that the pipeline runs correctly; synthetic data with the same structural properties serves that purpose without the privacy exposure. A student learning to write SQL queries against clinical data does not need to practice on actual patients.

But synthetic data has real limitations that institutional governance needs to acknowledge. First, synthetic data inherits the biases of the data it was generated from. A synthetic dataset generated from an EHR that underrepresents certain demographic groups will underrepresent

those groups in the synthetic version. Bias auditing of synthetic data requires the same demographic stratification that bias auditing of real data requires. Second, for purposes that require population-level validity — epidemiological analysis, model external validation, clinical trial simulations — the fidelity of synthetic data is an empirical question that requires evaluation, not an assumption. The institution should not claim that a model trained on synthetic data performs equivalently to a model trained on real data without evidence.

Third, the privacy protection that synthetic data provides is not absolute. Membership inference attacks — attempts to determine whether a specific individual's data was in the training set — can be applied to synthetic data generators as well as to trained models. The privacy guarantee of synthetic data depends on the generation method and the characteristics of the source data. For high-dimensional or rare-condition data, where individuals may be uniquely identifiable by their clinical signature, synthetic generation requires formal privacy guarantees — differential privacy, for example — rather than relying on the assumption that no record in the synthetic set corresponds to any real person.

17.10 Federated Learning and the Governance of Distributed Data

Federated learning offers a different architecture for the same problem: rather than sharing data with a central model, participating institutions keep their data local and share only model updates — the numerical gradients that represent what the model learned from a local training round. A central aggregator combines the updates from multiple institutions into an improved global model, which is then distributed back to the nodes for the next training round. The data never leaves the hospital's firewall.

Production federated learning infrastructure exists in healthcare. The NIH Bridge2AI[6] program is developing standardized data and AI-ready infrastructure across four large data generation projects, with federated coordination as a design principle (National Institutes of Health 2023a). The Medical Imaging and Data Resource Center[7] (MIDRC) has used federated learning to validate imaging AI models across multiple institutions without centralizing imaging data (MIDRC Consortium 2024). NVIDIA FLARE[8] is the dominant production framework for healthcare federated learning, with documented deployments at major academic medical centers.

Governance of federated learning participation is meaningfully different from governance of centralized data sharing, but it is not governance-free. Participating in a federated training consortium requires a data use agreement with the consortium that specifies what the model is being trained to do, who controls the aggregated model, and how the model can be used after training. An AMC that contributes gradients to a federated training round for a commercial AI product is contributing to the development of that product — a contribution that has

[6] https://bridge2ai.org
[7] https://www.midrc.org
[8] https://nvflare.ai

intellectual property implications and that may or may not be reflected in the benefit the institution receives from the trained model.

The most important governance question for federated learning is model poisoning risk: a malicious participant can contribute corrupted gradients that degrade the global model or embed adversarial behaviors. For clinical AI models where the output influences patient care, this is not a theoretical concern. Federated learning consortia need technical safeguards — gradient inspection, anomaly detection, trusted node certification — and governance frameworks that define what audit mechanisms participants can rely on to verify the integrity of their contribution and the safety of the models they receive.

17.11 Where to Start

17.11.1 Starter Project 1: AI Data Governance Policy and BAA Audit

What it is: A review of all current vendor agreements for AI tools handling PHI, assessing whether each BAA includes the provisions in Table 17.1, combined with a draft institutional AI data governance policy operationalizing the classification framework and honest broker requirements above.

Why now: Many AMC AI deployments occurred under standard enterprise agreements not specifically negotiated for AI use. A BAA audit surfaces contractual gaps before an adverse event. A data governance policy closes the structural gap that allowed non-compliant tools to be deployed.

How to execute: Legal and compliance review each active vendor agreement. Gaps are prioritized by data sensitivity and tool risk tier. Renegotiation is requested for high-priority gaps; tools with unresolvable gaps are candidates for replacement. The governance policy is drafted by legal and informatics jointly, reviewed by the AI Steering Committee, and published with a defined annual review cycle.

Buy vs. build: Legal and governance work. Commercial data governance platforms can maintain the inventory, but the policy analysis requires legal judgment.

17.11.2 Starter Project 2: Institutional AI Data Enclave

What it is: A technically isolated compute environment in which regulated AMC data can be used for AI development and experimentation without transmitting PHI to external services.

Why now: Demand for AI development access to clinical data is increasing, and case-by-case IRB and data governance review does not scale. A standing enclave with pre- approved data protocols, audit logging, and network isolation provides a faster path for approved researchers while preventing accidental data egress.

How to execute: Build on existing research computing infrastructure with added network controls and an approved tool list. Pre-approve OMOP-formatted clinical datasets at appropriate de-identification levels. Establish a lightweight application process for enclave access that replaces per-project data governance review for compliant use cases.

Buy vs. build: Build infrastructure on HIPAA-eligible cloud or on-premises HPC. The institutional work is in governance design, access controls, and audit processes layered on top of the compute environment.

18 Project Management and AI Portfolio Governance

The most common failure mode in AMC AI programs is not technical. It is organizational. Across health systems that have published candid accounts of AI implementation, the recurring pattern is a graveyard of pilots: technically sound models that achieved strong performance on held-out test sets and then stalled or failed at the point of clinical deployment. The failures trace not to the models but to the institutional machinery around them — absent ownership, underfunded maintenance, clinicians who were not consulted during development, and a go-live event that mistook deployment for done.

Avoiding this pattern requires treating the AI program as a portfolio management function, not a series of one-off projects. The difference is not semantic. A portfolio function has a governing body with the authority and budget to prioritize, fund, and terminate projects. It has a standardized intake process that evaluates every proposal against a consistent set of criteria before committing resources. It has stage gates that require rigorous validation before any tool influences a clinical decision. And it has a post-deployment infrastructure that treats every deployed tool as a living clinical asset requiring continuous monitoring, calibration, and eventual decommissioning. This chapter describes what that machinery looks like at an AMC that has built it.

18.1 The AISC as Portfolio Manager

In most AMCs, the AI Steering Committee begins as a governance and review body: a committee that meets monthly to evaluate proposals, review vendor contracts, and oversee compliance with the policies described in Chapter 10 and Chapter 16. This is a necessary starting point. It is not a sufficient end state.

The AISC that drives sustainable AI adoption has evolved beyond the ethics-review model into an active portfolio manager — an executive body with four specific powers that passive review committees lack. First, it holds and allocates a central AI portfolio budget, distinct from individual department IT budgets, that can fund feasibility work, shadow deployments, and pilot infrastructure without requiring each sponsoring department to independently fund the technical overhead. Second, it maintains an actively managed project registry that tracks every AI tool from initial proposal through deployment and eventual decommission, creating the institutional memory that prevents the same failed vendor from being re-proposed three

years later by a department that was not involved in the original evaluation. Third, it has the authority to terminate: a project that fails its stage-gate review or underperforms in post-deployment monitoring can be decommissioned without requiring the originating department's agreement. Fourth, it produces a quarterly portfolio report to the executive team and an annual report to the board — described in Section 10.6 — that makes the institution's AI risk posture visible at the level of governance where accountability actually resides.

The AISC chair is typically the CMIO, supported by the CIO for infrastructure decisions, the CISO for security review, and General Counsel for regulatory and liability matters. Clinical representation — department chairs or their designees in the service lines with the highest AI deployment density — ensures that portfolio decisions are grounded in operational reality rather than purely in technical or financial criteria. The ethics, workforce, and patient engagement leads described throughout this book should have standing membership or reporting relationships to the AISC, because the decisions made at the portfolio level are the ones that determine whether the governance commitments in individual chapters are real or theoretical.

18.2 The Intake Engine: Triage Before Resource Commitment

The intake process is the first gate in the AI portfolio — the point at which the institution decides whether a proposed tool or project is worth the structured evaluation that follows. Getting it right is operationally significant: at a mid-sized AMC, the volume of AI proposals — from departments evaluating vendor products, from researchers seeking approval for LLM-assisted analysis, from clinical informatics fellows with internal development ideas — can easily exceed the AISC's evaluation capacity if there is no pre-screening step.

A mature intake process has three components. The first is a structured intake form that captures what the AISC needs to assess strategic fit and risk tier without yet committing to a full evaluation. Minimum required fields include: the clinical or operational problem being addressed, the proposed tool or approach (vendor product, internal build, or academic partnership), the data types that will be used, the patient population affected, the anticipated volume of AI-influenced decisions per month, the sponsoring department and named champion, and a preliminary assessment of whether the tool meets the ONC definition of a Decision Support Intervention under HTI-1 (Office of the National Coordinator for Health Information Technology 2024). The DSI classification field matters because it triggers vendor transparency obligations that the institution can enforce at procurement rather than discovering post-deployment.

The second component is a rapid risk-tier assignment — a structured screen that places each proposal in one of three tiers based on patient safety exposure, regulatory classification, and data sensitivity. Tier 1 covers administrative and operational tools that do not directly influence clinical decisions: scheduling optimization, supply chain AI, administrative documentation drafting. Tier 1 proposals can proceed to vendor evaluation or internal development with AISC notification but without full committee review. Tier 2 covers clinical decision support tools that influence but do not automate clinical decisions, and research tools processing de-identified

data. Tier 2 proposals require full AISC evaluation. Tier 3 covers tools that directly influence high-stakes clinical decisions, tools processing restricted data, and any tool that qualifies as Software as a Medical Device under FDA regulations. Tier 3 proposals require a dedicated risk assessment with legal, CISO, and clinical leadership sign-off before advancing to stage-gate.

The third component is a vendor Model Card requirement for any Tier 2 or Tier 3 commercial product. Mitchell and colleagues' model card framework — now widely adopted by major AI vendors — specifies a standardized format for reporting training data sources, performance across demographic subgroups, intended and out-of-scope uses, and known limitations (Mitchell et al. 2019). The Coalition for Health AI[1] (CHAI) has adapted this format for clinical AI, adding performance reporting requirements specific to healthcare regulatory standards. Requiring a model card at intake catches a substantial fraction of vendor governance gaps before the institution has committed to a procurement process.

Table 18.1: AMC AI project intake checklist, tiered by risk. The named champion field is required at all tiers because absence of a named owner predicts deployment failure regardless of technical quality.

Intake Field	Purpose	Required Tier
Problem statement and clinical need	Validates that AI is the right solution	All
Proposed tool and development path	Buy/build/connect assessment	All
Data types and regulatory classification	HIPAA, FERPA, Common Rule scoping	All
Patient population and decision volume	Risk exposure quantification	All
Named champion and department sponsor	Ownership accountability	All
DSI classification assessment	ONC HTI-1 compliance trigger	2–3
Model card or equivalent	Vendor transparency verification	2–3
Equity impact pre-assessment	Section 1557 compliance baseline	2–3
FDA SaMD classification	Regulatory pathway determination	3
Preliminary IRB assessment	Research use determination	3

[1] https://www.coalitionforhealthai.org

18.3 Stage-Gate Discipline: From Ideation to Scale

The clinical trial phase model is the correct analogy for AI deployment in a clinical institution. A compound that passes safety screens in preclinical work is not approved for patient use; it proceeds through Phase I, II, and III trials with pre-registered hypotheses, independent monitoring, and pre-defined stopping rules. An AI tool that performs well on a held-out test set has cleared the equivalent of a preclinical screen. Deploying it directly to clinical use without a supervised piloting phase is the equivalent of moving from animal testing to widespread patient administration.

The DECIDE-AI[2] reporting guidelines for early-stage clinical evaluation of AI decision support systems — developed by a multinational consensus group and published in Nature Medicine — define the specific requirements for pilot evaluation that parallel Phase I and Phase II trial standards: prospective design, pre-registered primary endpoints, independent oversight, and monitoring for unexpected harms (Vasey et al. 2022). The stage-gate model operationalizes these requirements within the AMC portfolio management process.

Figure 18.1: AMC AI project stage-gate model. Each gate requires documented evidence before the project advances. A failed gate triggers remediation or termination, not automatic recycling. The monitoring loop reflects the continuous nature of post-deployment governance.

Shadow deployment — sometimes called dry-run or silent-mode — is the stage that most AMC AI programs skip and that most AMC AI failures trace to. In a shadow deployment, the tool runs in parallel with existing clinical workflows, generating outputs that are logged and reviewed but not presented to clinicians or integrated into clinical decisions. The shadow period generates evidence that no test set evaluation can provide: performance in the actual clinical environment (not the curated dataset), the distribution of cases on which the tool is

[2] https://www.decide-ai.org/

invoked (which may differ significantly from the training distribution), and early signals of demographic performance disparities.

The Duke Health Sepsis Watch program, one of the most extensively documented clinical AI implementations in the peer-reviewed literature, ran a multi-year shadow deployment before the tool influenced clinical decisions — and the shadow period identified operational patterns that required significant model recalibration before live use was safe (Sendak et al. 2020). The lesson is not that every deployment requires two years of shadow testing; it is that the shadow duration should be calibrated to the tool's risk tier, the stability of the clinical environment, and what the shadow data reveal, rather than compressed by deployment pressure from executive sponsors.

Gate 3 — the transition from shadow deployment to clinical pilot — is the highest-stakes gate in the model. It requires a documented safety review that includes the equity audit described in Section 16.11, a simulation exercise in which clinicians walk through the tool's failure modes with the clinical informatics team, and a formal AISC vote. The pre-registered primary endpoints for the pilot — the metrics that will determine whether the tool advances to enterprise deployment or returns to remediation — must be locked before Gate 3. Post-hoc success criteria are a governance failure.

18.4 The Integration Tax and Pilot Design

A clinical AI pilot whose primary outcome is model accuracy is measuring the wrong thing. By Gate 3, model accuracy should already be established — that is the purpose of the shadow deployment and feasibility review. The clinical pilot's primary purpose is to measure impact: on clinical workflow, on clinician cognitive load, on the specific patient outcomes the tool was designed to improve, and on the demographic equity of those outcomes across the patient population served.

The integration tax concept captures a specific and pervasive failure mode: a tool that is technically accurate but operationally burdensome enough that clinicians develop workarounds, ignore alerts, or route around the tool entirely. The literature on alert fatigue in clinical decision support systems documents this pattern in detail — tools that generate high volumes of low-specificity alerts are overridden at rates exceeding ninety percent, creating an alert environment in which genuine high-priority warnings are indistinguishable from background noise (Parasuraman and Manzey 2010). An AI tool that adds net cognitive load without proportional clinical value carries a negative integration tax even when its model metrics are strong.

Pilot design for integration tax measurement requires instrumentation at the workflow level, not just the model output level. The relevant metrics are: time-to-decision for clinical tasks the tool is designed to support, alert override rates and override documentation quality, workflow steps added versus removed by the tool, and clinician-reported experience using validated instruments.

The pilot should also measure non-use: what fraction of target patient encounters result in the tool being invoked, and what fraction of invocations are dismissed without substantive review. A tool with a forty-percent non-use rate at pilot is not ready for enterprise deployment regardless of its performance on the cases where it was used.

The integration tax calculation does not end at go-live. Tools that appear workflow-neutral during a short pilot can accumulate friction over time as edge cases multiply, as the tool's outputs diverge from evolving clinical practice, and as initial novelty wears off and interaction behavior reverts to pre-tool patterns. The monitoring infrastructure described in the next section should include workflow metrics alongside performance metrics to detect the gradual accumulation of integration tax that often precedes tool abandonment.

18.5 The Total Product Lifecycle

The go-live event is not the end of the AI project management process. It is the transition from pre-deployment governance to post-deployment governance — a shift that requires dedicated infrastructure and defined responsibilities that are distinct from the team that built or procured the tool.

The Total Product Lifecycle concept holds that every deployed tool requires ongoing monitoring for three categories of change that can degrade its performance without any modification to the tool itself. The first is dataset shift: changes in the clinical environment that alter the statistical distribution of inputs the model receives. Finlayson and colleagues documented how protocol changes and coding practice shifts during the COVID-19 pandemic caused multiple deployed models to systematically underperform, with no change to the models themselves — the clinical context had shifted in ways the models' training data did not anticipate (Finlayson et al. 2021). The second is population shift: demographic changes in the patient population served that move it away from the population on which the model was validated. The third is protocol shift: changes in clinical guidelines or institutional protocols that alter the clinical context within which the tool's outputs are interpreted.

Post-deployment monitoring requires three institutional commitments. First, a monitoring cadence: scheduled performance reviews at defined intervals — monthly for Tier 3 tools, quarterly for Tier 2 — that compare current performance metrics against the validation benchmark established at Gate 3. Second, a drift detection mechanism: an automated alert when model performance drops below a pre-defined threshold, triggering escalation to the AISC rather than waiting for the next scheduled review. Third, a model update protocol: a defined process for requesting and evaluating vendor model updates, aligned with the PCCP provisions described in Chapter 10, that treats an unannounced model update as a safety event requiring the same Gate 3 review process as a new tool deployment.

For internally developed tools, total product lifecycle management also requires a decommissioning protocol: a defined process for retiring a tool when monitoring finds that it no longer

meets safety or effectiveness standards, or when a superior tool is available. Decommissioning is systematically neglected in AMC AI programs, resulting in registries that grow without pruning and an operational environment in which clinicians are exposed to tools of varying and undocumented performance quality (Mitchell et al. 2019). The stage-gate model's decommission branch — Gate 4's failure path — is only actionable if someone is assigned to act on it.

18.6 New Human Infrastructure: Architects and Champions

The institutional roles that make AI portfolio governance operational are not adequately described by existing job classifications in most AMCs. Two roles represent genuinely new professional functions rather than extensions of existing IT or clinical informatics positions.

The AI Solution Architect is the technical-organizational bridge between the platform infrastructure described in Chapter 14 and the clinical departments deploying AI tools. The role requires fluency in both directions: deep enough technical knowledge to evaluate model performance, design RAG pipelines, configure API gateway access controls, and read FHIR resource specifications, and deep enough clinical knowledge to translate between algorithm behavior and clinical workflow implications. This combination is rare; most AMCs will need to develop it through structured training of existing clinical informatics staff rather than recruiting for it from the general technology market.

The Solution Architect's responsibilities include: technical review of vendor model cards at intake, shadow deployment instrumentation and monitoring, performance dashboard maintenance for the deployed portfolio, and first-line response to drift alerts that do not require AISC escalation. One architect can realistically support ten to fifteen deployed tools; an AMC with a large deployment portfolio may need multiple architects organized by clinical domain. The role belongs formally within clinical informatics, not within the general IT department, because the primary work — assessing clinical plausibility, understanding workflow context, communicating with clinical leadership — requires clinical domain knowledge that a general infrastructure team does not provide.

The Clinician AI Champion, introduced in Chapter 15, has a specific portfolio management function that complements the Solution Architect's technical role. Champions are the primary mechanism through which the AISC receives frontline intelligence about tool performance, workflow integration problems, and safety concerns that would not be visible in dashboard metrics alone. A tool that is technically performing within normal parameters but is being systematically misused because clinicians misunderstand its output format is a safety risk that dashboard monitoring will not detect — but a well-connected champion in the department will. The champion's dual reporting relationship — to the department for operational matters and to the AISC for governance matters — creates the bidirectional communication channel that connects governance to point-of-care reality.

Table 18.2: RACI matrix for AMC AI portfolio governance decisions. R = Responsible (does the work), A = Accountable (owns the outcome), C = Consulted (provides input), I = Informed (receives notification). The AISC Chair holds Accountable status on the highest-stakes decisions; the Clinical Lead holds Responsible status for the clinical safety gate.

Decision	AISC Chair	CISO	Legal	Clinical Lead	Solution Architect
Approve vendor contract	A	C	R	C	C
Advance project past Gate 3	A	C	C	R	C
Declare AI safety incident	A	R	C	C	I
Authorize shadow deployment	R	C	I	C	A
Retire deployed tool	A	I	C	C	R
Approve model update	R	C	C	C	A

18.7 Valuing AI: The Return on Health Framework

Traditional IT ROI calculation — cost of implementation divided by projected efficiency gains — maps poorly onto clinical AI investments, because the most significant value created by clinical AI tools is frequently not captured in direct cost reduction. An ambient documentation system that saves a physician forty-five minutes per day does not generate direct revenue; it generates reclaimed time that may be applied to additional patient care, research, or recovery from the burnout that drives costly physician attrition. A sepsis prediction model that identifies cases fourteen hours earlier than clinical suspicion does not save costs that appear in a line item; it prevents mortality and complications whose downstream costs are measured in readmissions, litigation exposure, and regulatory scrutiny.

The Return on Health framework, developed through collaborative work between the AMA[3] and digital health stakeholders, defines six value streams that together constitute a more complete accounting of clinical AI investment value. The clinical value stream captures direct

[3] https://www.ama-assn.org

improvement in patient outcomes — prevented adverse events, reduced diagnostic error rates, earlier treatment initiation. The patient value stream captures patient experience, engagement, and access. The provider value stream captures clinician time, cognitive load, and burnout — a driver of physician workforce retention that AMC CFOs increasingly recognize as a balance-sheet issue with consequences that dwarf the implementation cost of most AI tools. The operational value stream captures workflow efficiency, throughput, and administrative cost reduction. The financial value stream captures revenue generation and cost avoidance that appear in standard accounting. The equity value stream captures reduction in care disparities — a dimension that the HHS Section 1557 regulatory landscape makes increasingly difficult to exclude from institutional accountability frameworks.

The Return on Health framework does not resolve the measurement challenge; quantifying equity value and burnout prevention in common units with financial return requires methodological choices that reasonable analysts will make differently. What it does is force the institution to account for the dimensions of value it cares about before committing to the investment, rather than discovering post-deployment that the business case was built on metrics that do not capture the outcomes the institution actually needs to move. For AMCs with an academic mission, a seventh value stream — educational and research value — deserves explicit representation: the learning generated from rigorous shadow deployments, from DECIDE-AI-compliant pilots, and from post-deployment monitoring constitutes a scholarly asset that contributes to the institution's research portfolio and its capacity to train the next generation of clinical AI practitioners.

18.8 Where to Start

18.8.1 Starter Project 1: Intake Process and Stage-Gate Framework

What it is: A formalized AI project intake process — a standardized submission form, a risk-tiering protocol, and documented stage-gate criteria — implemented as the mandatory path for all new AI tool evaluations, whether vendor procurements or internal builds.

Why now: An AMC without a formal intake process is making AI investment decisions based on informal advocacy and executive attention rather than structured evidence. The intake process is the prerequisite for portfolio management, because you cannot manage a portfolio you cannot see. The regulatory pressure from Colorado SB 24-205's annual impact assessment requirement and HHS Section 1557's equity audit obligation creates a compliance incentive: institutions that begin documenting their evaluation process now will be in a substantially better position for the first annual reporting cycle than those that start from scratch when the deadline arrives.

How to execute: Draft the intake form using the fields in Table 18.1. Implement it as a

REDCap[4] survey or equivalent institutional form routed to the AISC administrative coordinator. Define the three risk tiers and the gate criteria in a governance document approved by the AISC. Communicate the new process to department chairs and informatics leadership with explicit framing that the process is designed to accelerate compliant projects, not to create bureaucratic delay. Track the fraction of new AI tool introductions that pass through the intake process as a governance metric in the annual report.

Buy vs. build: Governance design and process work. A REDCap instance or equivalent institutional survey tool handles the form; no dedicated software purchase is required. Commercial AI governance platforms (Credo AI, OneTrust AI Governance) offer intake workflow features that can reduce configuration work for institutions with large portfolios, but the policy analysis underlying the tier criteria requires internal governance judgment that no platform automates.

18.8.2 Starter Project 2: Deployed AI Portfolio Dashboard

What it is: A centralized monitoring dashboard that tracks performance metrics, usage patterns, and drift signals for all Tier 2 and Tier 3 tools currently deployed — the operational implementation of the total product lifecycle monitoring described above.

Why now: An institution that cannot answer "which AI tools are currently deployed, and are they performing within acceptable parameters?" is not meeting its governance obligations under the NIST AI RMF's Manage function, the Colorado SB 24-205 impact assessment requirement, or the basic standard of care accountability the AMA has articulated for AI-assisted clinical practice (National Institute of Standards and Technology 2023; American Medical Association 2024a). The dashboard is also the operational foundation for the annual board report described in Section 10.6: without systematically collected performance data, the report is a narrative rather than an evidence document.

How to execute: Build on the AI tool inventory created in the clinical workstream (Section 6.4). For each Tier 2 and Tier 3 tool, define two to four performance metrics that can be automatically extracted from existing institutional data — EHR data, audit logs, API gateway logs. Configure automated alerts at pre-defined drift thresholds. Build a quarterly summary view for AISC review that compares current metrics against the baseline established at Gate 3. The dashboard does not require a dedicated analytics platform; a structured connection between the API gateway audit logs, the EHR reporting infrastructure, and a business intelligence tool the institution already uses is a sufficient foundation for most deployed portfolios.

Buy vs. build: The analytics infrastructure is primarily a configuration project layered on existing institutional systems. Commercial MLOps platforms (Azure AI, Arize AI) offer purpose-built model monitoring capabilities that can accelerate the build for institutions with active internal AI development programs. For vendor-managed tools, performance monitoring

[4] https://projectredcap.org/

obligations should be included in vendor contracts at procurement — with data export rights specified — so that vendor-provided metrics can be integrated into the central dashboard rather than reviewed only in vendor portals.

19 Evaluation and Monitoring

Deploying an AI tool without a monitoring plan is not a deployment. It is an experiment, one in which patients, students, researchers, and administrative staff are the unwitting subjects. The experiment ends when something goes wrong visibly enough that someone shuts it off. What the institution rarely knows at that point is how long the model had been quietly performing below the threshold it was validated against, or how many people were affected before the degradation became obvious.

Evaluation and monitoring are not bureaucratic overhead bolted onto AI deployment to satisfy governance committees. They are the mechanism by which an institution finds out whether a tool is actually doing what it claimed to do, in the specific population and context where it was deployed, after the vendor demonstration is a memory. Without them, the institution is flying on instruments it has not calibrated.

The principles underlying good evaluation and monitoring are largely domain-agnostic. Local validation before deployment, shadow-mode testing before live use, drift detection after go-live, structured feedback loops from the people using the tool: these apply whether you are deploying a sepsis prediction model in the ICU or an AI-assisted grant writing tool in the research office. What differs substantially by domain is what you measure, who you ask, what a bad result means, and what governance body you escalate to when you find one. A clinical AI tool that degrades in performance is a patient safety event. The same degradation in a business operations tool is a financial exposure. The same degradation in a student assessment tool is an equity problem. Same principle, completely different response.

19.1 Why Benchmark Performance Does Not Transfer

Vendor-reported performance numbers tell you very little about how a model will perform at your institution.

Every prediction model is a function of the data it was trained on. That data reflects a specific case mix, specific documentation practices, and specific clinical workflows. When any of those change, model performance changes too. Sometimes the change is modest. Sometimes it is not.

The Epic Sepsis Model is the most-cited illustration. The vendor-reported AUC was approximately 0.76 to 0.83. When researchers at Michigan Medicine conducted an independent

external validation using actual patient outcomes rather than the training-population data, the AUC dropped to 0.63, indistinguishable from the performance of a much simpler severity score already in clinical use (Wong et al. 2021). The model had been implemented at hundreds of health systems before that external validation was published. The question is not whether external validation should have been required. It obviously should have been. The question is what institutional infrastructure makes local validation the default rather than the exception.

Performance degradation does not only happen at implementation. It also happens over time, in production, to models that performed adequately at go-live. Finlayson and colleagues documented how the COVID-19 pandemic induced rapid dataset shift across a wide range of clinical models, not because the models were flawed, but because the patient population, treatment patterns, and documentation practices in 2020 were structurally different from the training data those models had learned from (Finlayson et al. 2021). This form of drift, where the underlying relationship between inputs and outputs changes, is particularly insidious because the model continues to generate predictions with the same apparent confidence. The output does not flag that the world has changed.

This is why every AI deployment requires two distinct activities: pre-deployment evaluation to confirm adequate local performance before the tool influences any decision, and post-deployment monitoring to confirm that it continues to do so over time.

19.2 Pre-Deployment Evaluation

Pre-deployment evaluation has two components: local validation and shadow deployment.

Local validation asks whether the model performs adequately on a retrospective dataset drawn from your own institution's population, documentation practices, and time period. This is not the same as accepting the vendor's validation dataset, even if that dataset includes data from institutions that resemble yours. A model validated exclusively on academic medical center data from the northeast may not perform adequately at a safety-net hospital serving a different demographic mix. The validation data needs to be yours.

Shadow deployment, sometimes called silent mode, runs the model in parallel with real clinical or operational workflow, generating outputs that are not shown to the users making decisions. This phase answers questions that retrospective validation cannot. What is the alert burden this model generates in real operations? What is the false positive rate under live documentation patterns? Are there systematic differences in performance across demographic subgroups that did not appear in the retrospective dataset? The answers sometimes differ substantially from what the validation study suggested, and finding out during silent mode rather than during live rollout is the point.

For clinical AI tools, the DECIDE-AI[1] reporting standard provides a structured framework for this early evaluation phase, covering the evidence required before moving from silent mode to supervised deployment and from supervised deployment to unsupervised live use (Vasey et al. 2022). TRIPOD+AI[2] provides the reporting standard for the pre-deployment validation study itself: a 27-item checklist specifying what must be documented about model development, calibration, and performance before a prediction model goes live in any setting (Collins et al. 2024).

For large language models, a third pre-deployment step applies that does not exist for traditional prediction models. LLMs can fail in ways prediction models do not: hallucinating plausible-sounding clinical information, exposing protected patient data in responses to queries that appear innocuous, producing outputs with demographic biases that standard performance metrics do not capture. A pre-deployment red-teaming exercise, structured adversarial testing by domain experts attempting to elicit harmful outputs, is the minimum bar for any LLM in a clinical or educational context.

Model cards, standardized by CHAI[3] as a "nutrition label" for AI tools, document the intended population, known performance limitations, demographic subgroup performance, and deployment constraints. Requiring vendors to provide a CHAI-compliant model card before evaluation begins shifts accountability appropriately. If the vendor cannot characterize subgroup performance, the institution has learned something important about what the vendor actually knows about its own model (Coalition for Health AI 2024).

19.3 Post-Deployment Monitoring: Catching Drift

Once a model is live, the monitoring question changes from "does this model perform well enough to deploy?" to "is it still performing well enough to remain deployed?"

Drift comes in three forms. Covariate drift occurs when the statistical distribution of the input data changes: the patient population ages, a new referral pattern alters the case mix, a workflow change affects how lab values are captured. The model was not trained on this new distribution and may perform worse on it without anyone noticing, because it is still generating outputs and users have adapted their interpretation of those outputs based on the tool's historical behavior. Concept drift is more fundamental: the underlying relationship between inputs and the outcome the model predicts changes. The COVID pandemic induced concept drift across virtually every clinical model trained on pre-2020 data. Concept drift cannot be fixed by retraining on more data from the same period; it requires recognizing that the world has changed and deciding whether the model can be updated to reflect that change. Data pipeline failures, upstream changes to EHR configuration or coding practices, can cause models to receive inputs structurally different from their training data without any change in

[1] https://www.decide-ai.org/
[2] https://www.tripod-statement.org/
[3] https://www.coalitionforhealthai.org/

the patient population. This is the most operationally tractable form of drift because it has an identifiable technical cause, but it is also easily missed because it does not look like a model failure from the outside.

The pharmacovigilance analogy is useful here. A drug approved by the FDA carries a post-market surveillance obligation: the manufacturer must monitor for adverse events that did not appear in the clinical trials. AI/ML-based Software as a Medical Device is now subject to analogous requirements. As of early 2025, the FDA's Total Product Lifecycle guidance for AI/ML-enabled device software functions remains a draft document and has not been finalized into a binding rule, but the agency's direction is clear: ongoing performance monitoring and active surveillance for drift and bias are expected elements of any clinical AI program (U.S. Food and Drug Administration 2024). Flag this as an area to check; a final rule may be in effect by the time you read this.

19.4 KPI Architecture: Three Tiers

The KPIs that matter for AI monitoring fall into three tiers that are frequently conflated, with real consequences for what gets measured and what gets missed.

Technical KPIs measure model performance: calibration, discrimination (AUC-ROC, F1), sensitivity and specificity at the operating threshold, subgroup performance across demographic categories. These are what data scientists measure. They are necessary but not sufficient, because a model that performs adequately on technical metrics can still fail to improve the outcomes it was deployed to affect, or actively harm them.

Operational KPIs measure behavior in the workflow: utilization rate (what fraction of eligible cases are receiving the model's output?), override rate (how often do clinicians or administrators reject the recommendation?), alert burden, and time effects. The override rate is particularly informative. A model with a high override rate either has poor accuracy that frontline users have recognized and are compensating for, or has accurate outputs that the workflow integration is presenting poorly. Both are problems requiring different responses, and neither is visible on a calibration curve.

Outcome KPIs measure whether the tool is actually improving what it was deployed to improve. Sepsis mortality rates, medication error rates, student board exam pass rates, revenue cycle denial rates: whatever the tool was meant to move. These are the hardest to measure because they require attributing outcomes to the tool in a context where many other things are changing simultaneously. They are also the only KPIs that answer the question the institution actually cares about.

Table 19.1: KPI taxonomy for AMC AI monitoring. Technical KPIs catch model degradation. Operational KPIs catch workflow and adoption problems. Outcome KPIs answer whether the tool is achieving its purpose. All three tiers are required; monitoring only technical KPIs is the most common failure mode in practice.

KPI Category	Example Metric	What It Detects	Who Interprets It	Monitoring Cadence
Technical	AUC-ROC by demographic subgroup	Performance degradation, bias	Data science / MLOps	Monthly
Technical	Calibration curve	Probability over/underprediction	Data science	Quarterly
Operational	Override / rejection rate	Workflow misfit, user distrust	Department head, AI champion	Weekly / monthly
Operational	Alert burden per user-hour	Automation bias risk	Clinical informatics	Weekly
Operational	Time-to-task delta	Efficiency impact	Operations / QI	Quarterly
Outcome	Downstream outcome rate	Whether the tool is working	QI officer, domain lead	Quarterly / annually
Outcome	Subgroup outcome equity	Disparity amplification	Equity officer, ethics	Quarterly

19.5 Stakeholder Feedback as a Monitoring Method

Quantitative dashboards miss failure modes that only show up in how people experience a tool.

A clinician who has learned to distrust a sepsis model's outputs because the false positive rate is too high, and who now processes the alert as background noise, is a monitoring signal that no AUC metric captures. A student who submits AI-generated text not to cheat but because the assessment format rewards performance over learning is a monitoring signal about assessment design. An administrative staff member who has developed informal workarounds for an AI-assisted billing tool because it fails on certain claim types is a monitoring signal about the model's coverage gaps.

Structured mechanisms for collecting this kind of signal are part of a monitoring program, not an optional supplement to it. For clinical AI, this means building feedback pathways for frontline clinicians and, where the tool directly affects patient care, for patients. Patient advisory boards and community advisory bodies have been used by some health systems

as governance mechanisms for equity-sensitive tools, providing ongoing input from affected communities rather than one-time consent at enrollment. The literature on participatory design in health AI is still developing; replicable methods exist but most programs remain institutional pilots.

For educational AI, faculty are both users and evaluators. Their sense of whether students are developing genuine clinical reasoning or outsourcing it to AI tools is a monitoring input that no automated system can provide. For research AI, the IRB and the research integrity office can surface concerns from researchers and trainees about data provenance, analysis reproducibility, and authorship questions that would otherwise not reach institutional visibility. For business operations AI, frontline administrative staff are the first to notice when an AI-assisted authorization tool starts generating unusual denial patterns, or when an automated scheduling system creates downstream inefficiencies invisible to management dashboards.

19.6 Domain-Specific Dimensions

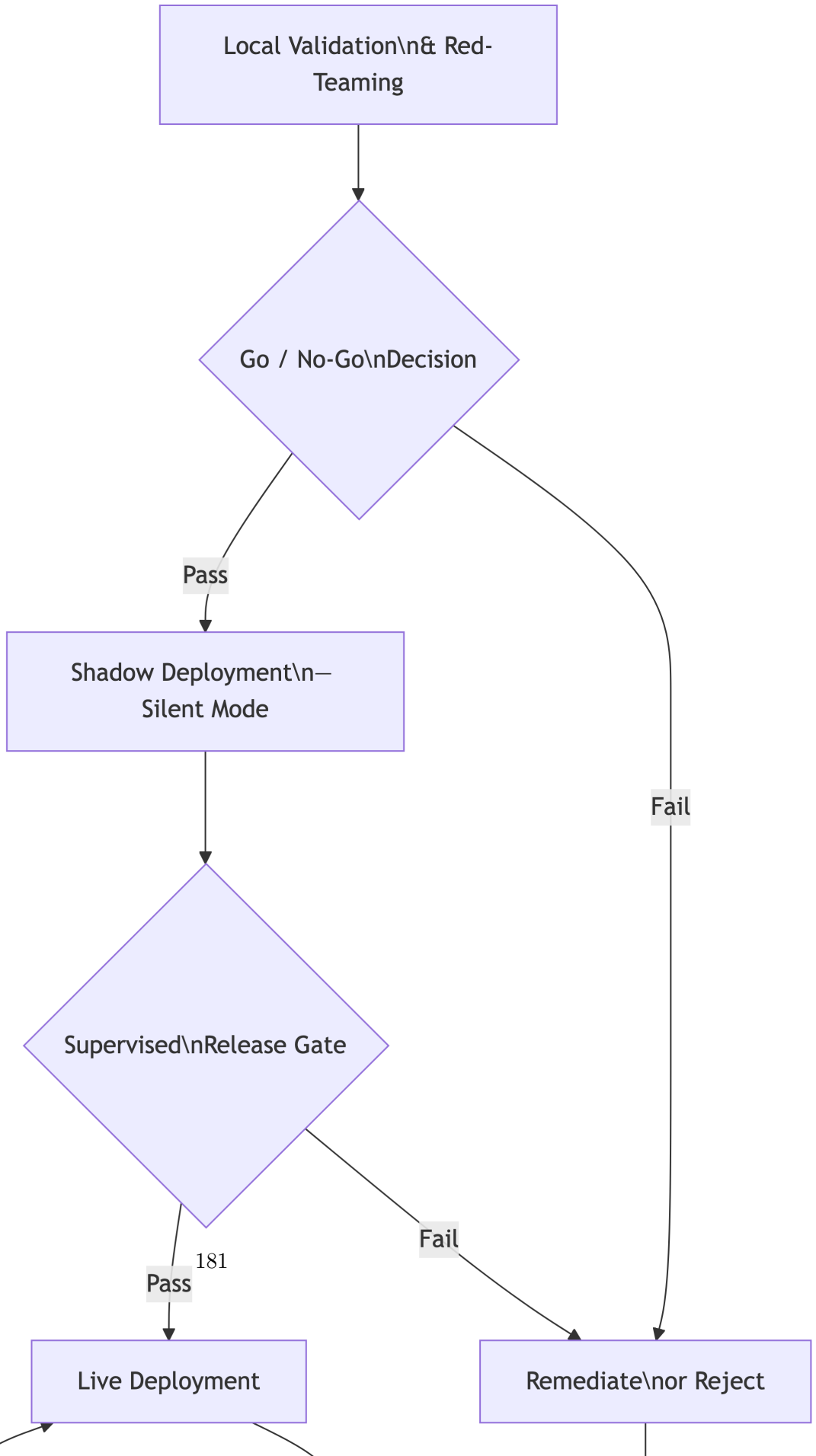
The general principles above apply across all four domains. What differs are the KPIs that matter most, who the primary stakeholders are, what the main ethical obligation is, and what governance body is responsible for escalation.

19.6.1 Clinical

Clinical AI monitoring operates under the most developed regulatory framework and carries the highest stakes for patient harm. Patient safety KPIs — near-miss events, adverse outcomes where the model's output was in the clinical record — are the core of outcome monitoring. Clinician override rates serve as a leading indicator; high false positive rates tend to show up as alert fatigue well before they appear as measurable patient harm events, and that pattern is visible through operational monitoring.

Equity monitoring is a legal requirement under the HHS Section 1557 final rule, which creates an affirmative obligation for health systems to identify and mitigate discrimination in their use of clinical decision support (U.S. Department of Health and Human Services, Office for Civil Rights 2024b). Demographic subgroup performance is not an optional component of technical monitoring. It is a compliance obligation.

The FDA PCCP framework requires that any predetermined changes to an AI/ML-based Software as a Medical Device — retraining triggers, performance floor thresholds, the training data update protocol — be specified and approved before they are implemented (U.S. Food and Drug Administration 2024). This gives the monitoring program a regulatory anchor: the decommissioning criteria and retraining triggers are documented commitments made at deployment time, not decisions improvised when a problem has already become visible.


Local Validation\n& Red-Teaming
Go / No-Go\nDecision
Pass
Fail
Shadow Deployment\n— Silent Mode
Supervised\nRelease Gate
Pass
Fail
Live Deployment
Remediate\nor Reject

Patient engagement in clinical AI monitoring extends beyond consent forms. Community advisory bodies with ongoing governance roles provide a channel for affected communities to surface concerns before they reach the level of documented harm. The literature on what this looks like in practice is still thin; the most replicable models involve structured advisory processes with defined input pathways into the AI governance committee, rather than ad hoc consultation that happens when a problem is already public.

19.6.2 Research

Research AI monitoring focuses on a different set of failure modes. The primary risks are not immediate patient harm but the slower, harder-to-reverse harms of corrupted science: results that are not reproducible, published findings not attributable to real effects, and training data used in violation of its collection consent.

Data provenance monitoring — tracking whether the data used for AI-assisted analysis was collected, consented, and used under IRB protocols that cover that use — is the research equivalent of clinical safety monitoring. An AI tool used in ways that fall outside its approved IRB protocol should trigger the same reporting pathway as any other protocol deviation. The IRB is the governance body for escalation in this domain.

Reproducibility KPIs include whether the code and configuration needed to reproduce an analysis are archived alongside the data, whether model outputs are stable across re-runs, and whether AI-assisted analysis steps are disclosed in publications per ICMJE standards (International Committee of Medical Journal Editors 2023). The reproducibility crisis in biomedical research predates AI: about 70% of researchers in one large survey reported being unable to reproduce another scientist's results, and about 50% had failed to reproduce their own (Baker 2016). AI-assisted analysis adds new dimensions to this problem, particularly when models are used to identify patterns in data without explicit hypothesis specification. Faster analysis is not more rigorous analysis, and the monitoring program needs to include mechanisms that reflect that distinction.

19.6.3 Education

The monitoring challenge in education is distinct because the failure mode is often invisible from outside. A student who produces AI-generated work that satisfies the stated learning objectives on paper, while the cognitive development those objectives were meant to build does not occur, is not detectable by any automated system.

The most important education monitoring KPI is assessment validity: does the assessment format still distinguish students who have developed genuine clinical reasoning from students who have not? This is a question for faculty, not for a dashboard. ACGME has introduced requirements for programs to demonstrate that trainees can engage competently with AI tools and that faculty are equipped to evaluate that competency (Accreditation Council for

Graduate Medical Education 2025). Those external requirements create an institutional monitoring obligation that exists independent of what any individual program director chooses to prioritize.

The case against AI detection tools as a primary integrity monitoring mechanism rests on two documented problems. They produce false positives for non-native English speakers at rates high enough to constitute a discriminatory monitoring practice (Liang et al. 2023). And they are a defensive posture that treats the symptom, AI-generated text submitted for assessment, rather than the cause, which is assessment design that does not require authentic engagement. Moving monitoring from policing toward assessment design, asking whether the assessments themselves require reasoning that AI cannot substitute for, is a more durable response. For health professions education specifically, peer-reviewed literature on AI monitoring in clinical training contexts remains limited as of 2025; most available evidence comes from general higher education contexts and should be applied with awareness that the competency stakes in clinical training are higher.

19.6.4 Business Operations

Business operations AI monitoring combines operational efficiency KPIs with equity obligations that are often underweighted in standard ROI analyses.

The efficiency KPIs are tractable: throughput rates, processing time, error rates, revenue cycle metrics. For ambient documentation tools, the relevant operational KPI is the actual reduction in documentation burden measured against the time required for AI output review and correction, not against the theoretical maximum (Tierney et al. 2024). Tools that reduce total documentation time are different from tools that redistribute documentation burden from typing to reviewing.

Algorithmic hiring bias monitoring is a legal requirement in jurisdictions that have enacted AI employment law. New York City's Local Law 144 requires employers using automated employment decision tools to conduct annual bias audits and disclose the results publicly (Local Law 144 2021). AMC human resources teams using AI in any part of the hiring or promotion process should know whether they fall under this or analogous state-level requirements, and build the audit infrastructure to comply before an enforcement inquiry requires it.

Frontline administrative staff are both the primary operational monitors and the primary workforce equity stakeholders. When AI automation displaces tasks that staff were previously performing, monitoring should include tracking of role changes, retraining uptake, and whether the efficiency gains from automation are distributed equitably across staff levels. The AMC that deploys an AI billing tool to reduce administrative headcount without monitoring the workforce equity dimensions of that reduction is carrying risk — regulatory, reputational, and labor relations — that standard operational dashboards will not surface.

Table 19.2: Domain comparison for AI monitoring. Each domain shares the general monitoring lifecycle but differs in what KPIs matter most, which stakeholders are primary, and what governance body has escalation authority.

Domain	Primary KPI Type	Key Stakeholders	Main Ethical Obligation	Governance Body	Illustrative Metric
Clinical	Safety and equity	Clinicians, patients, patient advisory board	Non-maleficence; HHS 1557 compliance	AISC, CMIO, patient safety	Override rate; subgroup AUC-ROC
Research	Reproducibility and provenance	Researchers, trainees, IRB	Research integrity; consent compliance	IRB, research integrity office	Protocol compliance rate; analysis reproducibility
Education	Assessment validity	Faculty, students, accreditors	Genuine competency development	Curriculum committee, ACGME	Assessment authenticity; faculty evaluation
Business Ops	Efficiency and equity	Admin staff, HR, finance	Fairness; workforce equity	CFO, CHRO, labor relations	Denial rate delta; hiring bias audit results

19.7 Decommissioning

Every deployed AI tool needs pre-defined criteria for when it should be taken offline. Setting those criteria at deployment time, rather than when a problem has become large enough to be undeniable, is one of the clearest markers of a monitoring program that is designed to work.

Decommissioning criteria should specify performance floors below which the tool goes offline regardless of workflow integration costs, subgroup equity thresholds that trigger mandatory review, data pipeline failure modes that trigger automatic suspension, and a review cadence at which continued deployment requires affirmative re-authorization rather than passive continuance. The specifics will differ by domain and by the risk profile of the tool. What matters is that they exist in writing before go-live.

Wiens and colleagues argued that the harm from deploying a poorly performing AI tool in a high-stakes clinical setting is not a theoretical risk that might materialize. It is a predictable consequence of a decision made without adequate evaluation infrastructure (Wiens et al. 2019).

Decommissioning criteria are the mechanism by which that infrastructure acknowledges its own limits.

19.8 Minimum Responsible Bar

It is worth being direct about what the minimum responsible bar actually is, because "comprehensive" monitoring programs are sometimes invoked as an argument for indefinite delay rather than a target for implementation.

For clinical AI, the minimum bar is: local validation on institutional data before go-live, shadow deployment with monitoring before clinician-facing rollout, pre-defined decommissioning criteria in writing, and at least quarterly review of technical and operational KPIs with a named reviewer. Below this bar, the institution is not governing its AI program. It is hoping the vendor got it right.

The minimum bar for research, education, and business operations AI is different in some dimensions and lower in others, but the structure is identical: every deployment should have a named monitoring owner, defined KPIs, a defined review cadence, and criteria that would result in the tool being taken offline. These are not aspirational targets for mature programs. They are the preconditions under which responsible deployment is possible at all.

A AI Principles Across Governance Frameworks

The WHO, the FDA, NIST, the AMA, and the EU have each produced AI governance frameworks from different starting points and for different audiences — and they converge on the same small set of concerns: safety, transparency, fairness, human oversight, and accountability. That convergence is meaningful. It suggests these are not arbitrary categories but genuine pressure points where AI systems consistently generate governance problems, regardless of who is doing the governing.

This appendix describes the major frameworks relevant to AMC AI governance, notes what is distinctive about each, and closes with a comparison table for governance teams designing programs that need to work across multiple standards simultaneously.

A.1 The Frameworks

WHO Ethics and Governance of Artificial Intelligence for Health (2021, updated 2024) (World Health Organization 2024) is the most comprehensive international health-sector AI ethics framework. Its six principles — human autonomy, human well-being and safety, transparency and explainability, responsibility and accountability, inclusiveness and equity, and responsiveness and sustainability — are organized around the patient and community as the primary stakeholders, not the health system. The sustainability principle, which asks whether AI deployment is globally equitable and environmentally responsible, is distinctive among health-sector frameworks and is often the least operationalized in AMC governance programs.

FUTURE-AI[1] is an international multi-stakeholder consortium that developed a framework structured around six properties whose initials spell the name: Fairness, Universality, Traceability, Usability, Robustness, and Explainability (Lekadir et al. 2022). The Universality principle — that AI tools should perform equitably across diverse populations and be validated on representative global datasets — goes further than most frameworks in demanding that equity be built into the development process, not added in post-deployment monitoring.

Good Machine Learning Practice for Medical Device Development: Guiding Principles[2] is a joint statement from the FDA, Health Canada, and the UK's MHRA

[1] https://future-ai.eu/

[2] https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles

(U.S. Food and Drug Administration et al. 2021). It is the most operationally specific of the frameworks listed here, specifying practices rather than values: representative datasets, independent training and test sets, thorough clinical testing, and explicit post-deployment monitoring requirements. Of the listed frameworks, it comes closest to a checklist that a vendor validation team can work through. AMC procurement teams evaluating clinical AI vendors should use this framework as a minimum due diligence standard.

AMIA's Artificial Intelligence Principles (American Medical Informatics Association 2024) map traditional bioethics principles — autonomy, beneficence, non-maleficence, justice — onto an AI governance context and add informatics-specific properties: explainability, interpretability, fairness, dependability, auditability, and knowledge management. The knowledge management principle, which asks how AI systems contribute to or degrade the institution's knowledge infrastructure, is distinctive and underrepresented in other frameworks.

NIST AI Risk Management Framework (AI RMF 1.0) (National Institute of Standards and Technology 2023) is a voluntary US federal framework organized around four functions — Govern, Map, Measure, Manage — with nine trustworthiness characteristics: accountable, explainable and interpretable, fair with bias managed, privacy-enhanced, reliable, safe, secure and resilient, transparent, and valid and accurate. The RMF is distinctive in its process orientation: unlike the principle frameworks, it specifies organizational functions and roles, making it useful as an implementation scaffold rather than a values statement. The AISC governance structure described in Chapter 3 maps naturally onto the RMF's Govern function.

AMA Principles for Augmented Intelligence (American Medical Association 2024a) are addressed to clinicians and health systems and place particular weight on physician agency: AI tools should support rather than supplant clinical judgment, and physicians must retain meaningful control over AI-assisted decisions. The AMA principles also address the obligation to inform patients when AI plays a role in their care, connecting to the notice and explanation principle in the OSTP Blueprint (Appendix B).

A.2 Where the Frameworks Agree

Table A.1: Convergence of principle themes across major AI governance frameworks relevant to AMC settings. A check mark indicates the framework addresses this theme explicitly; absence indicates the theme is not a primary focus, not that the framework opposes it.

Principle Theme	WHO	FUTURE-AI	FDA/MHRA	AMIA	NIST AI RMF	AMA
Safety and non-maleficence						
Transparency and explainability						
Fairness and equity						
Human oversight and autonomy		—				
Accountability		—				
Post-deployment monitoring						—
Privacy	—	—	—	—		

Principle Theme	WHO	FUTURE-AI	FDA/MHRA	AMIA	NIST AI RMF	AMA
Global/demographic representativeness						—

Safety, transparency, fairness, and post-deployment monitoring appear in every framework. That is the minimum common denominator — the set of themes any credible AMC AI governance program has to address if it wants to be legible to regulators, accreditors, and institutional partners drawing from any of these sources. The divergences are instructive too: privacy, sustainability, knowledge management, and global representativeness each appear in only one or two frameworks. These are not areas of disagreement so much as areas where different communities have different starting assumptions. Governance teams should treat them as prompts for explicit institutional choices rather than gaps to paper over.

B OSTP Blueprint for an AI Bill of Rights

The White House Office of Science and Technology Policy published the Blueprint for an AI Bill of Rights in October 2022 as a nonbinding framework for the design, use, and deployment of automated systems affecting the American public. It is not a law, and it creates no enforceable obligations. It matters for AMC AI governance in a specific way: it is the clearest statement of the federal government's normative expectations for AI systems before those expectations were codified into binding regulation, and several of its five principles have since been incorporated, in varying forms, into enforceable rules — including the HHS Section 1557 algorithmic nondiscrimination requirements, the ONC HTI-1 transparency mandates, and the FTC's enforcement posture on AI capability claims.

Reading the Blueprint alongside the regulatory chapters of this book shows which principles have made the transition from aspiration to obligation and which remain in the aspirational category. That distinction is useful for governance planning: AMC AI programs that have operationalized the Blueprint's principles are better positioned for the regulatory requirements that have followed from them.

B.1 The Five Principles

Safe and Effective Systems. Automated systems should be tested for safety and effectiveness before deployment, and should not be used where they pose unacceptable risks of harm to individuals. For clinical AI, this principle maps directly onto the validation requirements in Chapter 18 and the staged deployment framework in Chapter 4. The ONC HTI-1 rule's requirement for performance documentation on validated populations is the regulatory operationalization of this principle.

Algorithmic Discrimination Protections. Automated systems should not discriminate on the basis of protected characteristics, and institutions should proactively ensure equitable design and deployment. This principle is the direct precursor to the HHS Section 1557 duty-to-mitigate requirement. The equity audit process in Section 16.11 and the demographic performance stratification requirement that runs throughout the clinical and ethics chapters are the institutional operationalization.

Data Privacy. People should be protected from abusive data practices, with built-in privacy protections and meaningful agency over how data about them is used. The data governance framework in Chapter 17 — including the honest broker function, the BAA non-negotiables,

the expert determination standard for AI training data — implements this principle. The state laws that have followed (Washington My Health MY Data Act, California AB 3030) give it additional specificity in their respective jurisdictions.

Notice and Explanation. People should know when an automated system is being used and understand how and why it contributes to outcomes affecting them. This principle maps to the consent architecture in Chapter 16, the ONC HTI-1 source attribute requirements, and the California AB 3030 disclosure mandate for AI-generated patient communications. It is the principle with the largest gap between aspiration and current institutional practice: most patients whose care involves AI-assisted documentation, prediction, or decision support receive no meaningful disclosure.

Human Alternatives, Consideration, and Fallback. People should be able to opt out of automated systems where appropriate and have access to a human who can remedy problems. For clinical AI, this principle requires that every AI-assisted clinical process have a defined human override path — that a clinician can decline an AI recommendation without penalty, that a patient can request human-only care documentation, and that the governance program has a mechanism for escalating and investigating AI-related errors. The CMS requirement that AI systems cannot substitute for human clinical review in Medicare coverage decisions gives this principle binding force in one specific regulatory context.

B.2 Current Status

The Blueprint was published in October 2022 under the Biden administration and was connected to a broader federal AI governance agenda that included Executive Order 14110 on the Safe, Secure, and Trustworthy Development of Artificial Intelligence (October 2023). Executive Order 14110 was revoked by Executive Order 14179 in January 2025 (Executive Office of the President 2025). The Blueprint itself — as an OSTP white paper rather than an executive order — was not formally revoked and remains publicly available, but its status as an expression of federal normative expectations has changed. Governance teams citing the Blueprint as evidence of federal alignment should note this distinction: the Blueprint's five principles remain a useful values framework, but they no longer carry the same signal about the direction of federal regulatory development that they did in 2022–2024.

B.3 The Blueprint's Limitations

Two aspects of the Blueprint are worth naming honestly. First, it was written primarily with consumer-facing AI in mind — recommendation algorithms, hiring screens, benefits determinations — and its application to complex clinical AI requires interpretation. The "notice and explanation" principle is tractable for a patient receiving a denial letter. It is considerably harder to operationalize for a sepsis prediction model running continuously in the

background of an ICU. Governance teams should treat the Blueprint as a values framework that requires domain-specific translation, not a procedure manual.

Second, the Blueprint is nonbinding. Institutions that have aligned their AI governance programs to its principles have done so voluntarily. The regulatory obligations that have followed from the Blueprint — Section 1557, HTI-1, the FTC's AI enforcement actions — are binding, but they do not cover the full scope of the Blueprint's five principles. An AMC that interprets "compliance with the Blueprint" as "compliance with the regulations it inspired" will have operationalized some of its principles and left others as aspirations. That is not necessarily wrong — governance programs have to prioritize — but it should be an explicit institutional choice rather than an implicit one.

B.4 Primary Sources

- Blueprint for an AI Bill of Rights (OSTP, October 2022): https://www.whitehouse.gov/ostp/ai-bill-of-rights/
- Applying the Blueprint — sector-specific guidance: https://www.whitehouse.gov/ostp/ai-bill-of-rights/applying-the-blueprint-for-an-ai-bill-of-rights/
- Executive Order 14179, Removing Barriers to American Leadership in Artificial Intelligence (January 2025) (Executive Office of the President 2025)

C ONC HTI-1: Algorithm Transparency and Interoperability

The ONC Health Data, Technology, and Interoperability (HTI-1) final rule, published in January 2024 and effective later that year, is the federal regulation that AMC AI governance programs encounter most frequently in the clinical domain (Office of the National Coordinator for Health Information Technology 2024). It amends the Health IT Certification Program established by the 21st Century Cures Act and introduces specific transparency requirements for AI and predictive models embedded in certified Electronic Health Records. This appendix summarizes the rule's AI-relevant provisions for governance teams who need a working understanding of what it requires without reading the full Federal Register text.

C.1 The Decision Support Intervention Framework

The HTI-1 rule restructures how the certification program treats clinical decision support by creating a new regulatory category: the "Predictive Decision Support Intervention" (Predictive DSI). A Predictive DSI is any clinical decision support function that uses AI or machine learning to generate patient-specific output — a risk score, a recommendation, an alert — that a clinician is expected to act on. The rule distinguishes Predictive DSIs from evidence-based DSIs (which cite explicit guideline references) and from workflow tools (which do not generate patient-specific clinical recommendations).

The distinction matters because Predictive DSIs are subject to substantially more rigorous transparency requirements than other DSI types. EHR vendors must make 31 structured source attributes available for every Predictive DSI in a certified system. Those attributes include training data sources and date ranges, the populations on which the model was validated, performance characteristics on those validation populations, known limitations and failure modes, update history and versioning, and instructions for appropriate interpretation. The attributes must be accessible at the point of care — meaning a clinician using a sepsis prediction model should be able to see the model's source attributes without leaving the clinical workflow.

C.2 What This Means for AMC Governance

For AMC AI governance programs, HTI-1 creates obligations in two directions.

For clinically deployed AI tools procured from vendors: the institution can and should require that every AI tool used in a certified EHR context provide the full set of HTI-1 source attributes. Vendors who cannot or will not provide this documentation are not in compliance with federal certification requirements, and that non-compliance transfers risk to the institution that deploys the tool. The intake checklist in Chapter 18 includes HTI-1 source attribute documentation as a required intake item for clinical AI tools.

For internally developed clinical AI tools deployed in or alongside certified EHR systems: the institution is effectively in the position of a developer and bears the documentation obligations itself. An internally developed readmission risk model that generates patient-specific recommendations in a certified EHR workflow is a Predictive DSI for purposes of the rule, and the institution must be able to provide the 31 source attributes for it. This requirement has driven several major AMC AI governance programs to adopt the model card standard as the format for their internal AI documentation — a format that maps naturally onto the HTI-1 attribute list.

C.3 USCDI v3 and AI Data Access

HTI-1 also establishes the United States Core Data for Interoperability version 3 (USCDI v3) as the baseline data standard for certified health IT. USCDI v3 expands the required data elements that EHR systems must be able to capture and exchange, including patient demographic and social determinants fields that are directly relevant to AI equity monitoring. The addition of more granular race, ethnicity, sexual orientation, and gender identity fields to the required data set means that institutions updating to USCDI v3 compliance will have better demographic data available for bias auditing than they had under earlier versions. Treating the USCDI v3 compliance effort as an AI readiness investment — specifically, as infrastructure for the demographic stratification that equity monitoring requires — is a practical way to align two compliance obligations that might otherwise proceed independently.

C.4 Information Blocking and AI Data Access

The HTI-1 rule includes updates to the information blocking regulations under 45 C.F.R. Part 171. Information blocking — practices by EHR developers, health information networks, and health care providers that interfere with the access, exchange, or use of electronic health information — is prohibited with exceptions. The updated exceptions framework is relevant to AI governance in two ways. First, institutions that restrict clinical AI tool access to certain data elements need to ensure those restrictions fall within a defined exception, or they may constitute information blocking. Second, institutions using FHIR APIs to connect AI tools to EHR data must ensure their API access policies comply with the information blocking rules,

which require that certified APIs be made available to any application a patient authorizes, not just to institution-selected tools.

The full text of the rule is available at the Federal Register: https://www.federalregister.gov/documents/2024/01/09/2023-28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency-and

D How This Book Was Written: A Multi-Model AI Authorship Workflow

This appendix documents the workflow used to research and write the chapters of this book. We include it not as a novelty item but because it is itself an instance of the kind of human-AI collaboration the book describes — and because the specific tool choices, failure modes, and tradeoffs we encountered are illustrative of the decisions any organization faces when it tries to use AI for substantive knowledge work.

The workflow involved four distinct agents with different roles: a human author who owned the intellectual direction, Claude Opus 4.7 as the planning and architectural intelligence, Claude Sonnet 4.6 as the orchestrating author, and Gemini as the research engine. Each played to its comparative advantage, and the division of labor was not arbitrary — it emerged from thinking carefully about what each tool does well and where each fails. The human author was present and directing at every phase: writing the prompts, evaluating the outputs, redirecting scope when the AI's framing was wrong, and deciding what to accept, revise, or discard. No phase ran autonomously to completion without human review.

D.1 The Workflow in Brief

The workflow proceeded in five phases. Table D.1 summarizes them.

Table D.1: The six-phase multi-model authorship workflow. The human author directed and evaluated every phase; no phase ran to completion without human review. The AI agents contributed throughput and breadth; the human author contributed judgment, direction, and accountability for the final text.

Phase	AI agent	What the AI did	Human author's role	Output
1. Scaffold	—	—	Built the Quarto book structure, wrote initial chapter drafts, established the domain/workstream framework	Existing `.qmd` files, `_quarto.yml`, `references.bib`

Phase	AI agent	What the AI did	Human author's role	Output
2. Analysis and planning	Claude Opus 4.7	Reviewed the repo, identified gaps, drafted a research plan covering structure, voice, citation standards, and figure requirements	Directed the analysis with specific prompts; reviewed the master prompt and revised scope before accepting it	`_research/MASTER_PROMPT` (~500 words)
3. Brief writing	Claude Sonnet 4.6	Wrote per-chapter research briefs specifying argument arc, source targets, figure sketches, and open questions	Reviewed each brief; redirected scope where framing was wrong (e.g., ensuring all four AMC domains were covered, not just clinical)	17 briefs in `_research/briefs/` (~16,500 words total)
4. Deep research	Gemini (via CLI)	Executed web-search-backed research for each brief, producing annotated source lists, argument spines, section outlines, and figure sketches	Reviewed each dossier for scope gaps; evaluated source quality; accepted, rejected, or re-ran research sessions as needed	17 dossiers in `_research/dossiers/` (~30,900 words total)
5. Authorship	Claude Sonnet 4.6	Wrote chapter prose from dossiers, verified citations, created diagrams and tables, added BibTeX entries	Reviewed each chapter; evaluated voice, argument, and factual accuracy; directed revisions and approved final text	Chapter `.qmd` files, updated `references.bib`
6. Review pass	Gemini + Claude Sonnet 4.6	Audited completed chapters for unsupported factual claims and unlinked entities; flagged gaps for resolution	Evaluated each flagged item; approved or rejected proposed citations; made final editorial calls on what to add, revise, or leave as synthesis	Updated `.qmd` files, new BibTeX entries, `_research/REVIEW_PROMPT` `_research/review_groups/`

D.2 Why These Tools in These Roles

The tool assignments were deliberate.

Claude Opus 4.7 for planning. The planning phase required understanding an existing codebase, identifying structural gaps, and writing a specification detailed enough that a separate

agent could execute against it without constant supervision. This is closer to architectural reasoning than text generation, and it benefited from Opus's stronger judgment on ambiguity and its ability to produce a coherent multi-thousand-word specification in a single pass. We used Opus only for the planning phase and switched to Sonnet for execution — the planning phase is the most cognitively expensive, but it is also the one that generates the fewest tokens, so the cost differential is acceptable.

Gemini for research. The research phase requires a different capability profile than authorship: breadth over depth, web search over reasoning, high recall over precise citation. Gemini's deep research mode — which runs multiple web searches, synthesizes across sources, and produces structured outputs — is well suited to this. It is also meaningfully faster and cheaper than having a frontier model perform iterative web searches itself. The limitation we encountered was that Gemini, when given complex multi-part research prompts, tends to produce planning text ("I will now search for…") without always completing the structured output the prompt requested. We addressed this by re-running failed chapters sequentially rather than in parallel to avoid rate-limiting, and by specifying the output structure in the master prompt precisely.

Gemini in autonomous (YOLO) mode also took actions we did not explicitly request — creating chapter stub files and editing `_quarto.yml` to add them to the book structure. This is a useful illustration of the principal-agent problem in AI deployment: an agent given broad autonomy in a codebase will take actions consistent with its understanding of the task, which may or may not match the operator's intent. In this case, the actions were sensible and we kept them. In a production deployment, review before accepting is the right default.

Claude Sonnet for orchestration and authorship. Sonnet handles the execution loop: read the dossier, verify suspicious citations before citing them, write prose, create diagrams, edit `references.bib`, check that the Quarto build passes. This is a high-volume task that benefits from fast output and low cost per token more than it benefits from the deeper reasoning Opus provides. The important thing is that the human reviewed each chapter after completion, not that the model was the most capable available.

D.3 The Citation Problem

The most consistent failure mode in the research phase was citation hallucination. Gemini's dossiers reliably produced regulatory primary sources (NIST, NIH, FDA, EEOC) with correct URLs, and reliably produced real peer-reviewed papers from high-profile journals with correct DOIs. Where it failed was with newer or more obscure literature: it generated plausible-sounding papers with plausible-sounding DOIs that do not exist. The pattern is recognizable in retrospect — sequential-looking identifiers (`10.1109/COMPUTER.2025.1234567`, `10.5281/zenodo.1234567`), papers from journals that would be unlikely venues for the specific claim, and 2025 publication dates for papers that were too specific to be plausible given the knowledge cutoff.

The structural fix we implemented after discovering fabricated DOIs in early dossiers was a two-tier citation discipline codified in `_research/MASTER_PROMPT.md`. Tier 1 covers sources Gemini can retrieve and verify directly during a session — federal documents, regulatory filings, institutional publications, vendor documentation — and requires a live URL confirmed during the research run. Tier 2 covers peer-reviewed biomedical and informatics papers, and requires only a placeholder in the format `[PUBMED-PENDING: Author YYYY - descriptor - journal]`, with an explicit prohibition on writing any DOI unless the publisher landing page was retrieved during that session. The placeholder system shifted verification responsibility to the authorship phase, where PubMed MCP tools and direct DOI resolution can confirm or reject each citation before it enters `references.bib`. This is not a perfect solution — it requires more authorship-phase work — but it eliminated the most damaging failure mode: a confident, wrong DOI cited in finished prose.

The authorship phase treats every Tier 2 placeholder as unverified until confirmed by direct lookup. The BibTeX entries in `references.bib` contain only sources that were either in the original project bibliography, confirmed real by direct URL or PMID retrieval, or drawn from prior knowledge with high confidence. When no verified citation is available, the prose makes the claim without a citation rather than citing a hallucinated paper.

Table D.2 summarizes the citation quality we observed across the completed dossiers.

Table D.2: Observed citation accuracy rates by source type across Gemini research dossiers. Rates are estimates based on spot-checking; not every citation was independently verified. The two-tier PUBMED-PENDING discipline was implemented to prevent fabricated Tier 2 citations from reaching `references.bib`.

Source type	Accuracy rate (observed)	Notes
Federal regulatory documents	~95%	Correct URLs, correct document numbers
Major journal policies (Nature, NEJM, JAMA)	~90%	Correct, with occasional date errors
High-profile peer-reviewed papers	~80%	Well-known papers reliably correct
Newer or niche peer-reviewed papers	~40%	Frequent fabrication of DOIs and author lists

Source type	Accuracy rate (observed)	Notes
Vendor documentation	~70%	URLs often stale or approximate
News/trade press	~60%	Article may exist but headline or date wrong

The practical implication is that Gemini research output should never be committed to a bibliography without verification. The workflow's structure — Gemini produces a dossier with verified Tier 1 sources and placeholders for Tier 2, a human or second model resolves the placeholders — is the right one for any research-intensive writing task.

D.4 Costs and Volume

Table D.3 summarizes the artifacts produced and their approximate scale. Token counts for the AI interactions were not systematically recorded; the estimates below are based on output sizes and typical input-to-output ratios.

Table D.3: Artifacts produced by the multi-model workflow across all authorship sessions. Word counts are approximate; individual chapter and dossier files can be examined in the repository for exact lengths.

Artifact	Files	Words	Notes
Research master prompt (`_research/MASTER_PROMPT.md`)	1	~500	Written by Claude Sonnet 4.6; two-tier citation discipline codified here
Chapter briefs (`_research/briefs/`)	17	~16,500	Written by Claude Sonnet 4.6
Automation script (`scripts/gemini-research.sh`)	1	76 lines	Written by Claude Sonnet 4.6
Gemini research dossiers (`_research/dossiers/`)	17 complete	~30,900	Produced by Gemini CLI (sequential runs)
Chapter prose (`.qmd` files)	17+	~22,300	Written by Claude Sonnet 4.6

Artifact	Files	Words	Notes
New BibTeX entries (`references.bib`)	30+ new entries	—	Added during authorship phase; Tier 1 verified, Tier 2 resolved via PubMed
Quarto stub files (Gemini-initiated)	several	~200 each	Autonomous action by Gemini; accepted and incorporated

The total research and planning effort — phases 2 through 4 — produced approximately 47,400 words of structured inputs for the authorship phase. The authorship phase produces finished chapters of approximately 2,500–4,500 words each, depending on the domain. The ratio of research material to finished prose reflects the verification and synthesis work the authorship phase requires: not all research material makes it into the chapter, the prose is more compressed than the dossiers, and citation verification eliminates a meaningful fraction of the sourced claims.

D.5 The Review Pass

After the authorship phase produced a complete draft, a structured review pass identified two categories of gap that the writing phase does not reliably close on its own: unsupported factual claims that carry meaningful epistemic weight, and significant entities — organizations, tools, frameworks, standards bodies — mentioned in the text without a hyperlink.

The review was structured by having Gemini perform a first-pass audit of completed chapters, using instructions codified in `_research/REVIEW_PROMPT.md`. Six review group files in `_research/review_groups/` specified which chapters each Gemini run should examine; runs were executed sequentially rather than in parallel to avoid rate-limiting. The prompt asked Gemini to identify, per chapter:

- **Unsupported claims**: Factual assertions that carry substantive weight and lack a citation or reference, prioritized by how much the claim's validity bears on the chapter's argument.
- **Unlinked entities**: Organizations, tools, frameworks, and regulatory bodies mentioned by name without a hyperlink that would allow a reader to verify or explore further.

The prompt imposed strict rules: Gemini was not to suggest fabricated URLs or invent citation keys. For unsupported claims, it could flag the claim and name the type of source that would

address it, but not supply a DOI it had not verified during that session. For unlinked entities, it could suggest a URL only if it had retrieved that URL during the current run.

Claude Sonnet reviewed Gemini's outputs and resolved items selectively. A meaningful fraction of flagged claims turned out to already be supported by citations elsewhere in the same paragraph — the audit prompt lacked full bibliography context and could not detect these. Those were closed as non-issues without adding redundant citations. Only high-confidence gaps were resolved: claims traceable to a verifiable primary source, and organizations with stable institutional URLs.

The net output: several new BibTeX entries for sources that had been mentioned without citation, and approximately forty inline links added to first occurrences of significant organizations and tools across the chapter set. The process also confirmed that many flagged claims were already adequately supported — which is useful negative space: it documents that the review was done rather than skipped.

The review pass is intentionally conservative. It closes verifiable gaps rather than filling word count. A chapter that makes a specific numerical claim without a citation, where the claim is traceable to a primary source, is a closure candidate. A chapter that makes a framing claim without a citation — one that represents synthesis rather than a specific verifiable fact — is not. The response to "this could use a citation" is different from "this is documented at a specific primary source we failed to cite." Conflating them inflates the bibliography without improving the book.

D.6 What This Means for AI-Assisted Knowledge Work

This workflow illustrates several things about AI-assisted knowledge work that are not obvious from reading about AI in the abstract.

AI at its most valuable is a research collaborator, not a ghostwriter. The chapters in this book are not AI-generated prose edited by a human. They are human-structured arguments, grounded in AI-gathered sources, written by a model that was given specific direction about what to argue and how to argue it, and reviewed by a human who can tell when the argument is wrong. The value of the AI is in the research throughput — ~30,900 words of sourced, structured research material produced in hours rather than days — not in generating text that the human then accepts.

The division of labor matters. Running Gemini research sessions in parallel consistently caused rate-limiting failures that produced empty or truncated outputs; running them sequentially, one at a time, succeeded. Using Opus for planning and Sonnet for execution reduced cost without sacrificing quality, because the task profiles genuinely differ. These are engineering decisions, not philosophical ones.

Autonomy requires clear boundaries. Gemini in YOLO mode created files and edited configuration — actions that were correct in this case but that could easily have been wrong in a different context. Any workflow that gives an AI agent write access to a production system needs explicit boundaries around what the agent is and is not authorized to change.

The bottleneck is authorship, not research. The research phase produced ~30,900 words of structured material in approximately four hours of elapsed time (including re-runs for failed dossiers). The authorship phase produces approximately 4,300 words per chapter at higher quality but significantly more elapsed time. At scale, the constraint is the human review step and the authorship pass, not the research. Tools that improve the authorship pass — better citation verification, tighter prose controls — will have a larger impact on throughput than further improvements to the research phase.

The workflow is reproducible. The `scripts/gemini-research.sh` wrapper, the `_research/briefs/` input structure, and the `_research/MASTER_PROMPT.md` template are all checked into the repository. Any contributor can run the same workflow for a new chapter or an update pass on an existing chapter. The intellectual bottleneck is writing a good brief; once the brief is right, the research and authorship passes follow a repeatable pattern.

References

AB 3030: Health Care: Artificial Intelligence (2024). https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB3030.

Accreditation Council for Continuing Medical Education. 2025. *Guidance on the Responsible Use of AI in Accredited Continuing Education*. Https://www.accme.org/news-releases/guidance-responsible-use-ai-accredited-ce[1]. https://www.accme.org/news-releases/guidance-responsible-use-ai-accredited-ce.

Accreditation Council for Graduate Medical Education. 2025. *Common Program Requirements: July 2025 Updates*. Https://www.acgme.org/what-we-do/accreditation/common-program-requirements/[2]. https://www.acgme.org/what-we-do/accreditation/common-program-requirements/.

American Hospital Association. 2025. *Health Care Workforce System Under Pressure, Poised for Reinvention*. AHA Center for Health Innovation Market Scan. https://www.aha.org/aha-center-health-innovation-market-scan/2025-12-09-health-care-workforce-system-under-pressure-poised-reinvention.

American Medical Association. 2023. *Physicians Optimistic about AI in Health Care with Some Concerns*. https://www.ama-assn.org/practice-management/digital/physicians-optimistic-about-ai-health-care-some-concerns.

American Medical Association. 2024a. *AMA Principles for Augmented Intelligence Development, Deployment, and Use*. https://www.ama-assn.org/practice-management/digital/ama-principles-augmented-intelligence-development-deployment-and-use.

American Medical Association. 2024b. *Doctors Work Fewer Hours, but EHR Still Follows Them Home*. AMA. https://www.ama-assn.org/practice-management/physician-health/doctors-work-fewer-hours-ehr-still-follows-them-home.

American Medical Informatics Association. 2024. *AI Competencies for Health Professionals*. https://amia.org/education-events/ai-competencies.

[1] https://www.accme.org/news-releases/guidance-responsible-use-ai-accredited-ce
[2] https://www.acgme.org/what-we-do/accreditation/common-program-requirements/

Association of American Medical Colleges. 2024. *Principles for the Responsible Use of Artificial Intelligence in and for Medical Education*. https://www.aamc.org/about-us/mission-areas/medical-education/principles-responsible-use-artificial-intelligence-and-medical-education.

Association of American Medical Colleges. 2025. *AI Competencies Across the Learning Continuum*. Https://www.aamc.org/about-us/mission-areas/medical-education/ai-competencies[3]. https://www.aamc.org/about-us/mission-areas/medical-education/ai-competencies.

Badal, Kimberly, Carmen M Lee, and Laura J Esserman. 2023. "Guiding Principles for the Responsible Development of Artificial Intelligence Tools for Healthcare." *Communication & Medicine* 3 (1): 47. https://doi.org/10.1038/s43856-023-00279-9.

Baker, Monya. 2016. "1,500 Scientists Lift the Lid on Reproducibility." *Nature* 533: 452–54. https://doi.org/10.1038/533452a.

Bedoya, Armando D, Nicoleta J Economou-Zavlanos, Benjamin A Goldstein, et al. 2022. "A Framework for the Oversight and Local Deployment of Safe and High-Quality Prediction Models." *Journal of the American Medical Informatics Association* 29 (9): 1631–36. https://doi.org/10.1093/jamia/ocac078.

Black Book Market Research. 2025. *AI Integration and Shadow IT in Health Systems*. Https://blackbookmarketresearch.com/healthcare-technology-reports[4]. https://blackbookmarketresearch.com/healthcare-technology-reports.

Centers for Medicare and Medicaid Services. 2024a. "Medicare and Medicaid Programs; Patient Protection and Affordable Care Act; Advancing Interoperability and Improving Prior Authorization Processes for Medicare Advantage Organizations, Medicaid Managed Care Plans, State Medicaid Agencies, CHIP Fee-for-Service Programs, CHIP Managed Care Entities, and Qualified Health Plan Issuers in the Federally-Facilitated Exchanges (Final Rule, CMS-0057-F)." In *Federal Register*, No. 25, vol. 89. https://www.federalregister.gov/documents/2024/02/08/2024-00895/medicare-and-medicaid-programs-patient-protection-and-affordable-care-act-advancing-interoperability.

Centers for Medicare and Medicaid Services. 2024b. "Medicare Program; Contract Year 2025 Policy and Technical Changes to the Medicare Advantage Program, Medicare Prescription Drug Benefit Program, Medicare Cost Plan Program, and Programs of All-Inclusive Care for the Elderly (Final Rule)." In *Federal Register*, No. 79, vol. 89. https://www.federalregister.gov/documents/2024/04/23/2024-07105/medicare-program-contract-year-2025-policy-and-technical-changes.

[3] https://www.aamc.org/about-us/mission-areas/medical-education/ai-competencies
[4] https://blackbookmarketresearch.com/healthcare-technology-reports

Chelli, Mikaël, Julien Descamps, Vincent Lavoue, et al. 2024. "Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Study." *Journal of Medical Internet Research* 26: e53745. https://doi.org/10.2196/53745.

Cleveland Clinic. 2025. *Cleveland Clinic Announces the Expanded Rollout of Bayesian Health's AI Platform for Sepsis Detection*. Cleveland Clinic Newsroom. https://newsroom.clevelandclinic.org/2025/09/23/cleveland-clinic-announces-the-expanded-rollout-of-bayesian-healths-ai-platform-for-sepsis-detection.

Coalition for Health AI. 2024. *Patient Trust and AI Accountability Survey*. Https://www.coalitionforhealthai.org/publications/[5]. https://www.coalitionforhealthai.org/publications/.

College of Healthcare Information Management Executives, and Censinet. 2025. *2025 CHIME/Censinet Healthcare Cybersecurity and AI Governance Survey*. Https://chimecentral.org/chime-censinet/[6]. https://chimecentral.org/chime-censinet/.

Collins, Gary S, Karel G M Moons, Paula Dhiman, et al. 2024. "TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods." *BMJ* 385: e078378. https://doi.org/10.1136/bmj-2023-078378.

Colorado General Assembly. 2024. *HB 24-1139: Concerning Human Review of Health Insurance Denials*. Https://leg.colorado.gov/bills/hb24-1139[7]. https://leg.colorado.gov/bills/hb24-1139.

Dhakal, Prithvi et al. 2024. "Evaluation of GPT-4 on USMLE Step 2 Clinical Knowledge Questions." *JMIR Medical Education*.

Executive Office of the President. 2025. "Removing Barriers to American Leadership in Artificial Intelligence." In *Federal Register*, No. 8741, vol. 90. https://www.federalregister.gov/documents/2025/01/23/2025-01953/removing-barriers-to-american-leadership-in-artificial-intelligence.

Federal Trade Commission. 2024. *Operation AI Comply: Continuing Actions Against AI-Related Deceptive Practices*. https://www.ftc.gov/news-events/press-releases/2024/09/ftc-announces-operation-ai-comply.

Fierce Healthcare. 2026. *75% of US Healthcare Systems Use or Plan to Use AI Platform in 2026*. Fierce Healthcare. https://www.fiercehealthcare.com/ai-and-machine-learning/75-us-healthcare-systems-use-plan-use-ai-platform-2026.

[5] https://www.coalitionforhealthai.org/publications/
[6] https://chimecentral.org/chime-censinet/
[7] https://leg.colorado.gov/bills/hb24-1139

Finlayson, Samuel G, Adarsh Subbaswamy, Karandeep Singh, et al. 2021. "The Clinician and Dataset Shift in Artificial Intelligence." *New England Journal of Medicine* 385: 283–86. https://doi.org/10.1056/NEJMc2104626.

Gao, Catherine A, Frederick M Howard, Nikolay S Markov, et al. 2023. "Comparing Scientific Abstracts Generated by ChatGPT to Real Abstracts with Detectors and Blinded Human Reviewers." *Npj Digital Medicine* 6: 75. https://doi.org/10.1038/s41746-023-00772-w.

Garabet, Liana, Alexei Kasparov, Marc Minciullo, John Tran, and Medhat Sarhan. 2024. "Performance of GPT-4 on USMLE Step 1–Style Questions." *Medical Science Educator* 34: 89–97. https://doi.org/10.1007/s40670-023-01968-3.

Greshake, Kai, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. "Not What You've Signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection." *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. https://doi.org/10.48550/arXiv.2302.12173.

Gymrek, Melissa, Amy L McGuire, David Golan, Eran Halperin, and Yaniv Erlich. 2013. "Identifying Personal Genomes by Surname Inference." *Science* 339 (6117): 321–24. https://doi.org/10.1126/science.1229566.

Hippel, Ted von, and Courtney von Hippel. 2015. "The Time Allocation Effects of Uncertainty: Grant Writing and Scientific Productivity." *PLOS ONE* 10 (5): e0127948. https://doi.org/10.1371/journal.pone.0127948.

HL7 International. 2024. *SMART App Launch Framework V2.2.0*. Https://hl7.org/fhir/smart-app-launch/[8]. https://hl7.org/fhir/smart-app-launch/.

International Committee of Medical Journal Editors. 2023. *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals*. https://www.icmje.org/recommendations/.

ISO/IEC 42001: Information Technology — Artificial Intelligence — Management System (2023). https://www.iso.org/standard/81230.html.

JAMA Network. 2023. *Instructions for Authors: Use of Artificial Intelligence, Language Models, Machine Learning, or Similar Technologies*. https://jamanetwork.com/journals/jama/pages/instructions-for-authors.

James, Casey A, Robert M Wachter, and James O Woolliscroft. 2022. "Preparing Clinicians for a Clinical World Influenced by Artificial Intelligence." *JAMA* 327 (14): 1333–34.

[8] https://hl7.org/fhir/smart-app-launch/

https://doi.org/10.1001/jama.2022.3580.

Jones, Clare, Janet Thornton, and Jeremy C Wyatt. 2023. "Artificial Intelligence and Clinical Decision Support: Clinicians' Perspectives on Trust, Trustworthiness, and Liability." *Medical Law Review* 31 (3): 501–20. https://doi.org/10.1093/medlaw/fwad013.

Kaufman Hall. 2026. *Hospitals Face 2026 New Normal: Rising Expenses and Shifts in Revenue Mix*. Kaufman Hall National Hospital Flash Report. https://www.kaufmanhall.com/news/hospitals-face-2026-new-normal-rising-expenses-and-shifts-revenue-mix.

Kirchner, Lauren, and Annie Waldman. 2023. *How Cigna Saves Millions by Having Its Doctors Reject Claims Without Reading Them*. Https://www.propublica.org/article/cigna-pxdx-medical-health-insurance-rejection-claims-without-reading[9]. https://www.propublica.org/article/cigna-pxdx-medical-health-insurance-rejection-claims-without-reading.

Kung, Tiffany H, Morgan Cheatham, Arielle Medenilla, et al. 2023. "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models." *PLOS Digital Health* 2 (2): e0000198. https://doi.org/10.1371/journal.pdig.0000198.

Lekadir, Karim, Aasa Feragen, Abdul Joseph Faris, et al. 2022. "FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging." *Insights into Imaging* 13: 169. https://doi.org/10.1186/s13244-022-01307-w.

Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. "GPT Detectors Are Biased Against Non-Native English Writers." *Patterns* 4 (7): 100779. https://doi.org/10.1016/j.patter.2023.100779.

Local Law 144: Automated Employment Decision Tools (2021). https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page.

McCoy, Liam G, Sulaiman Nagaraj, Farouk Morgado, Vinyas Harish, Sunit Das, and Leo Anthony Celi. 2020. "What Do Medical Students Actually Need to Know about Artificial Intelligence?" *Npj Digital Medicine* 3 (1): 86. https://doi.org/10.1038/s41746-020-0294-7.

MIDRC Consortium. 2024. *Medical Imaging and Data Resource Center (MIDRC)*. Https://www.midrc.org/[10]. https://www.midrc.org/.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, et al. 2019. "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287596.

[9] https://www.propublica.org/article/cigna-pxdx-medical-health-insurance-rejection-claims-without-reading
[10] https://www.midrc.org/

Mollick, Ethan R, and Lilach Mollick. 2023. *Assigning AI: Seven Approaches for Students, with Prompts*. SSRN. https://ssrn.com/abstract=4475995.

National Institute of Standards and Technology. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1.

National Institute of Standards and Technology. 2024. *Artificial Intelligence 600-1: Generative Artificial Intelligence Profile*. NIST AI 600-1. U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1.

National Institutes of Health. 2020. *Final NIH Policy for Data Management and Sharing (NOT-OD-21-013)*. Https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html[11]. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.

National Institutes of Health. 2023a. *Bridge2AI: Generating AI-Ready Data for All*. Https://commonfund.nih.gov/bridge2ai[12]. https://commonfund.nih.gov/bridge2ai.

National Institutes of Health. 2023b. *The Use of Generative Artificial Intelligence Technologies Is Not Permitted for NIH Peer Review Activities*. NOT-OD-23-149. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html.

National Institutes of Health. 2025. *Guidance on Generative AI and Controlled-Access Human Genomic Data (NOT-OD-25-081)*. Https://grants.nih.gov/grants/guide/notice-files/NOT-OD-25-081.html[13]. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-25-081.html.

National Library of Medicine. 2024. *PubMed Statistics*. https://pubmed.ncbi.nlm.nih.gov/stats/.

National Science Foundation. 2024. *Notice on the Use of Generative Artificial Intelligence Technology in the NSF Merit Review Process*. https://new.nsf.gov/news/notice-on-the-use-of-generative-artificial-intelligence.

Ng, Felix Y C, Arun James Thirunavukarasu, Helen Cheng, et al. 2023. "Artificial Intelligence Education: An Evidence-Based Medicine Approach for Consumers, Translators, and Developers." *Cell Reports Medicine* 4 (10): 101230. https://doi.org/10.1016/j.xcrm.2023.101230.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366

[11] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html
[12] https://commonfund.nih.gov/bridge2ai
[13] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-25-081.html

(6464): 447–53. https://doi.org/10.1126/science.aax2342.

Office of Management and Budget. 2024. *M-24-10: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence*. https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10.pdf.

Office of the National Coordinator for Health Information Technology. 2023. *Trusted Exchange Framework and Common Agreement (TEFCA), Common Agreement Version 1.1*. Https://www.healthit.gov/tefca[14]. https://www.healthit.gov/tefca.

Office of the National Coordinator for Health Information Technology. 2024. "Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing (HTI-1)." In *Federal Register*, No. 8, vol. 89. https://www.federalregister.gov/documents/2024/01/09/2023-28824/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency-and.

Parasuraman, Raja, and Dietrich H Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Updated Understanding." *Human Factors* 52 (3): 381–410. https://doi.org/10.1177/0018720810376055.

Pew Research Center. 2023. *Americans' Views of Artificial Intelligence Use in Health Care*. Https://www.pewresearch.org/science/2023/02/22/60-of-americans-would-be-uncomfortable-with-provider-relying-on-ai-in-their-own-health-care/[15]. https://www.pewresearch.org/science/2023/02/22/60-of-americans-would-be-uncomfortable-with-provider-relying-on-ai-in-their-own-health-care/.

Polubriaginof, Fernanda C G et al. 2024. "Implementation of a Generative Artificial Intelligence Solution for Patient Message Drafting in an Academic Medical Center." *JAMIA Open*, ahead of print. https://doi.org/10.1093/jamiaopen/ooae024.

Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (Artificial Intelligence Act) (2024). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689.

SB 24-205: Consumer Protections for Artificial Intelligence (2024). https://leg.colorado.gov/bills/sb24-205.

Sendak, Mark P, William Ratliff, Dina Sarro, et al. 2020. "Real-World Integration of a Sepsis Deep Learning Technology into Routine Clinical Care: Implementation Study." *JMIR*

[14] https://www.healthit.gov/tefca

[15] https://www.pewresearch.org/science/2023/02/22/60-of-americans-would-be-uncomfortable-with-provider-relying-on-ai-in-their-own-health-care/

Medical Informatics 8 (7): e15182. https://doi.org/10.2196/15182.

Singhal, Karan, Shekoofeh Azizi, Tao Tu, et al. 2023. "Large Language Models Encode Clinical Knowledge." *Nature* 620: 172–80. https://doi.org/10.1038/s41586-023-06291-2.

Thirunavukarasu, Arun James, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. "Large Language Models in Medicine." *Nature Medicine* 29: 1930–40. https://doi.org/10.1038/s41591-023-02448-8.

Tierney, Aaron A, Shreya Bhatt, Aakash Houndie, et al. 2024. "Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation." *NEJM Catalyst Innovations in Care Delivery* 5 (1): CAT.23.0404. https://doi.org/10.1056/CAT.23.0404.

U.S. Department of Health and Human Services, Office for Civil Rights. 2012. *Guidance Regarding Methods for de-Identification of Protected Health Information in Accordance with the HIPAA Privacy Rule*. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

U.S. Department of Health and Human Services, Office for Civil Rights. 2024a. *Nondiscrimination in Health Programs and Activities (45 C.F.R. § 92.210)*. 89 FR 37522. https://www.federalregister.gov/d/2024-08967. https://www.federalregister.gov/d/2024-08967.

U.S. Department of Health and Human Services, Office for Civil Rights. 2024b. *Section 1557: Nondiscrimination in Health Programs and Activities (Final Rule)*. https://www.hhs.gov/civil-rights/for-individuals/section-1557/index.html.

U.S. Department of Health and Human Services, Office of the Surgeon General. 2022. *Addressing Health Worker Burnout: The U.S. Surgeon General's Advisory on Building a Thriving Health Workforce*. Https://www.hhs.gov/surgeongeneral/reports-and-publications/health-worker-burnout/index.html[16]. https://www.hhs.gov/surgeongeneral/reports-and-publications/health-worker-burnout/index.html.

U.S. Equal Employment Opportunity Commission. 2024. *Strategic Enforcement Plan Fiscal Years 2024-2028*. https://www.eeoc.gov/strategic-enforcement-plan-fiscal-years-2024-2028.

U.S. Food and Drug Administration. 2024. *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning-Enabled Device Software Functions*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-aiml-enabled-device-software.

[16] https://www.hhs.gov/surgeongeneral/reports-and-publications/health-worker-burnout/index.html

U.S. Food and Drug Administration, Health Canada, and UK Medicines and Healthcare products Regulatory Agency. 2021. *Good Machine Learning Practice for Medical Device Development: Guiding Principles*. Https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles[17]. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles.

U.S. Senate Committee on Homeland Security and Governmental Affairs. 2024. *AI in Medicare Advantage: A Review of Care Denials*. Https://www.hsgac.senate.gov/wp-content/uploads/2024-Senate-MA-AI-Report.pdf[18]. https://www.hsgac.senate.gov/wp-content/uploads/2024-Senate-MA-AI-Report.pdf.

Vasey, Baptiste, Myura Nagendran, Bruce Campbell, et al. 2022. "Reporting Guideline for the Early-Stage Clinical Evaluation of Decision Support Systems Driven by Artificial Intelligence: DECIDE-AI." *Nature Medicine* 28: 924–33. https://doi.org/10.1038/s41591-022-01772-9.

Washington State Legislature. 2023. *My Health MY Data Act*. RCW 19.373. https://app.leg.wa.gov/RCW/default.aspx?cite=19.373. https://app.leg.wa.gov/RCW/default.aspx?cite=19.373.

Wiens, Jenna, Suchi Saria, Mark Sendak, et al. 2019. "Do No Harm: A Roadmap for Responsible Machine Learning for Health Care." *Nature Medicine* 25: 1337–40. https://doi.org/10.1038/s41591-019-0548-6.

Wolters Kluwer. 2025. *Survey: Generative AI in Healthcare — Clinician Use and Risks*. Https://www.wolterskluwer.com/en/news/survey-reveals-doctors-using-ai[19]. https://www.wolterskluwer.com/en/news/survey-reveals-doctors-using-ai.

Wong, Andrew, Erkin Otles, John P Donnelly, et al. 2021. "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients." *JAMA Internal Medicine* 181 (8): 1065–70. https://doi.org/10.1001/jamainternmed.2021.2626.

World Health Organization. 2024. *Ethics and Governance of Artificial Intelligence for Health*. https://www.who.int/publications/i/item/9789240029200.

Yaneva, Victoria et al. 2024. "Evaluating GPT-4 on USMLE Step 2 CK: Implications for Clinical Reasoning Assessment." *Academic Medicine*.

Zack, Travis, Eric Lehman, Mirac Suzgun, et al. 2024. "Assessing the Potential of GPT-4 to

[17] https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles

[18] https://www.hsgac.senate.gov/wp-content/uploads/2024-Senate-MA-AI-Report.pdf

[19] https://www.wolterskluwer.com/en/news/survey-reveals-doctors-using-ai

Perpetuate Racial and Gender Biases in Health Care: A Model Evaluation Study." *The Lancet Digital Health* 6 (1): e12–22. https://doi.org/10.1016/S2589-7500(23)00225-X.

Zemmar, Adjmal, Andres M Lozano, and Bradley J Nelson. 2023. "The Ethical Imperative for the Use of Artificial Intelligence in Surgery." *Nature Machine Intelligence* 5: 1184–90. https://doi.org/10.1038/s42256-023-00742-6.