Ethical Implications of AI in Healthcare Management

How do we harness AI's power while protecting what matters most?

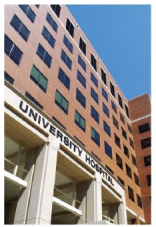
- **Betsy Hopson, PhD, MSHA**

  University of Alabama at Birmingham

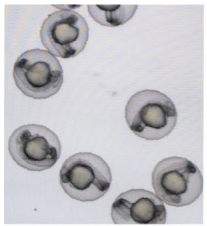
SAU 2026 Annual Meeting at the AUA | May 14, 2026

UAB MEDICINE



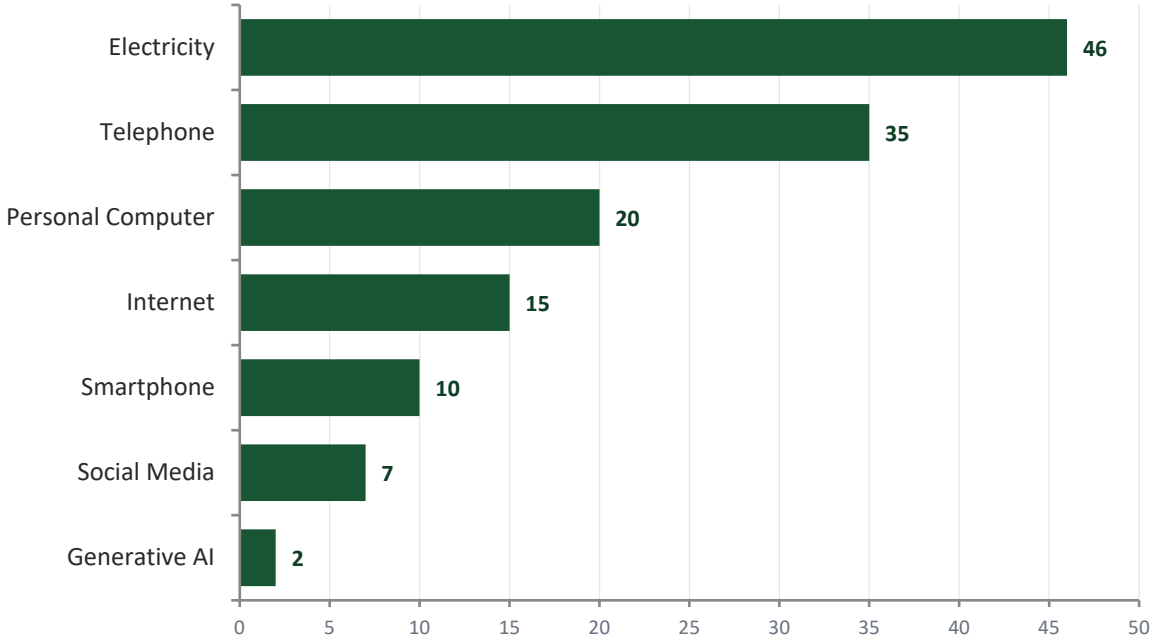







The Fastest Technology Adoption in History

ChatGPT reached 100 million users in 2 months. Compare to every prior technology.



Years required to reach 100 million users

Gen AI adoption

doubled in ONE YEAR

33% → 71%

2023 → 2024 · McKinsey / Stanford HAI

UAB MEDICINE

Why Ethics Must Lead — Not Follow — the AI Revolution

✓ AI's Promise in Medicine

- Earlier, more accurate diagnoses
- Reduced physician burnout via automation
- Personalized treatment plans
- Optimized workflows & admin tasks
- AlphaFold: solved 50-year protein folding problem
- AI scribe tools cutting documentation burden

⚠ Unresolved Ethical Risks

- AI ethics guidelines developed WITHOUT clinicians, patients, or developers
- 'Black box' decisions — AI cannot explain its reasoning
- Biased training data perpetuating health disparities
- Data privacy risks under HIPAA & GDPR
- Erosion of human empathy and the physician-patient relationship

The Black Box Problem: When AI Can't Explain Itself

> *"Can clinicians make vital decisions without grasping the fundamental reasoning behind them?"*
> — The core clinical AI dilemma

Trust

Clinicians and patients need to understand the rationale behind AI recommendations — not just accept outputs at face value

Consistency

AI models trained in one hospital setting may fail when deployed with different patient populations or clinical practices

Accountability

When AI causes harm, who bears responsibility — the algorithm, the developer, the hospital, or the clinician?

Explainability is not optional — it is an ethical imperative in clinical AI

UAB MEDICINE

Explainability in Clinical AI: Who Is Requiring It

Six authoritative sources calling it a prerequisite, not a preference

FDA

AI/ML action plan & SaMD guidance

Explainability is a prerequisite for clinical deployment, not optional.

WHO

Ethics & Governance of AI for Health, 2021

One of six core ethical principles that must be present before deployment.

American Medical Association

AMA Policy, 2023

AI in clinical decision-making must be explainable to the clinician AND patient.

OWASP

LLM Top 10 AI Security Framework, 2025

Lack of explainability listed as a top-tier security and safety risk.

The Joint Commission

AI governance in accreditation standards

Explainability named as a component of safe AI deployment.

Peer-reviewed evidence

Tang, Li & Fantus (2023) — 36 studies

Explainability is the strongest predictor of clinician trust and AI adoption.

The bottom line: Demanding explainability is not a fringe academic position — it is the consensus of FDA, WHO, AMA, OWASP, The Joint Commission, and peer-reviewed evidence.

Bias in Medical AI: Who Gets Left Behind?

Race Bias in Risk Algorithms

A widely used U.S. algorithm relied on healthcare spending as a proxy for illness — resulting in >50% fewer Black patients being identified as needing extra care, despite equal illness burden to White patients.
Obermeyer et al., Science, 2019

Cardiovascular Disease in Women

Most heart disease prediction models are trained predominantly on male data. Cardiovascular disease manifests differently in women — AI trained this way routinely misses diagnoses in female patients.
Mihan & Van Spall, npj Cardiovascular Health, 2024

Dermatology & Skin Tone

Skin lesion AI trained primarily on lighter skin shows significantly lower accuracy for patients with darker skin — particularly troubling given higher melanoma mortality in these populations.
Daneshjou et al., Science Advances, 2022

If AI learns from historically biased data, it doesn't fix disparities — it encodes and scales them.

UAB MEDICINE

AI Won't Say No.

And that is one of its most dangerous qualities.

Unlike a colleague or consultant, AI is designed to be helpful ALWAYS. It has no hesitation, no professional doubt, no discomfort that makes a human say 'I'm not sure.' Ask it a question, it will answer. Ask it the same question 10 ways, it answers all 10. Confidently. Even when it is wrong.

No built-in refusal

Ask AI to recommend a dosing regimen or draft a denial letter — it will. Even with insufficient context.

Confident hallucination

AI presents fabricated information with the same certainty as accurate information. No 'I made that up' signal.

Sycophancy effect

Research shows AI changes a correct answer when users push back. It tends to agree with you — even when you are wrong.

You own the consequences

AI can surface options and flag patterns. But the clinician bears the legal, ethical, and human weight of the decision.

The physician's judgment and accountability cannot be delegated to an algorithm.

What Even the Developers Don't Fully Understand

Three phenomena that reveal how much remains unknown about how AI actually works

2x

The Repeated Prompt Effect — *Google Research, 2025*

Simply copying and pasting your exact prompt twice — saying it twice in a row — improves AI accuracy by up to 76% across GPT-4, Gemini, Claude, and DeepSeek. No rephrasing. Just repetition. No one can fully explain why, even the engineers who built these systems.

?!

Hallucinations: Confident & Wrong — *FDA / NIH / OpenAI, 2024-25*

AI generates false medical information with complete confidence. GPT-3.5 hallucinated 39.6% of medical references. OpenAI's own testing found its newest models hallucinated 30-50% of the time. A therapy chatbot told a patient in addiction recovery to use methamphetamine.

!>

The Overthinking Problem — *2025 Research*

Asking AI to 'think harder' or 'reason more deeply' can make answers WORSE. Researchers found consistent performance decline after extended reasoning. More compute does not equal more accuracy — a failure mode with no parallel in human cognition.

Deploying technology whose inner workings even its creators don't fully understand demands humility, oversight, and governance — not blind trust.

Hidden Threats: Prompt Injection & Invisible Text Attacks

What is it?

Attackers embed malicious instructions invisible to the human eye — white text on white background, hidden in metadata — but fully readable by AI. When the AI processes the document, it silently obeys the hidden commands, with no visible sign anything went wrong.

EchoLeak — Microsoft 365 Copilot (2025) · CVE-2025-32711

A hidden prompt in a routine email hijacked Copilot with zero clicks. When the AI summarized the email, it silently executed attacker commands and forwarded sensitive inbox data. Healthcare risk: same attack works on AI tools processing patient referral letters.

PDF Document Poisoning (2025) — OWASP #1 AI Threat

Researchers created PDFs with invisible hidden text. When a clinician uploaded the document for AI summarization, concealed prompts activated and instructed the AI to exfiltrate sensitive data.

Medical Peer Review Manipulation (2025)

Researchers embedded invisible prompts in journal submissions instructing AI reviewers to give positive evaluations regardless of quality. Confirmed on Claude, GPT-4, and Gemini. The medical research pipeline is an active attack surface.

Any AI tool summarizing clinical notes, referrals, insurance docs, or research papers is a potential attack surface — and patient PHI is at risk.

Protecting Against Prompt Injection: What You Can Do

THE SOBERING TRUTH: OpenAI and the UK National Cyber Security Centre agree — prompt injection may NEVER be fully solved. It stems from the fundamental architecture of how AI works: there is no clean line between data and instructions inside the model.

For Clinicians: Behavioral Defenses

1. Never let AI take final action on a document you did not personally review — treat AI outputs as drafts, not decisions
2. Be specific with AI instructions. Vague prompts like 'handle this referral' give attackers more room to exploit
3. If AI output on a document feels unexpected or off, treat it as suspicious — your clinical instinct is a valid security signal
4. Do not upload unverified documents (unsolicited PDFs, referrals from unknown sources) to AI tools with system access

For Health Systems: Technical Defenses

- **A** Least privilege: AI tools should only access the specific data needed for one task — never full EHR access
- **B** Mandatory audit logging: every AI action touching PHI must be timestamped and logged — this is where HIPAA enforcement is heading
- **C** Vendor security assessment: before deploying any AI tool, demand proof of prompt injection testing and their incident response plan
- **D** Assume breach posture: design AI workflows so that if an injection succeeds, the damage is contained and immediately detectable

"Stop treating AI like an oracle. Start treating it like a high-privilege employee with known and dangerous flaws." — UK NCSC, Dec 2025

Patient Data: Privacy, Consent & the HIPAA/AI Collision

HIPAA & AI

- HIPAA §164.508 requires explicit informed consent before PHI is used beyond direct treatment
- AI algorithms routinely process vast patient data for predictive analytics
- 'Minimum necessary' standard (§164.502) must govern all AI data access
- Patients have rights to access, amend, and restrict use of their health data
- GDPR fines totaled $1.3 billion in 2024 — healthcare AI is squarely in regulators' sights

The Real Tensions

- Most patients don't know their EHR data trains AI models
- 'Notice and consent' models fail when patients lack awareness of future data uses
- AI can infer sensitive data from non-sensitive data (GDPR Article 22 challenge)
- Informed consent for AI-assisted care is an emerging and unresolved legal frontier
- Who owns a patient's data when it is used to train a commercial AI product?

UAB MEDICINE

Protecting What AI Cannot Replace: The Human Connection

The #1 ethical concern shared by **BOTH** clinicians AND patients across 36 empirical studies:
AI will depersonalize medicine and eliminate the human relationships that make care work.

— Tang, Li & Fantus (2023). Systematic Review of 36 Empirical Studies on Medical AI Ethics.

AI Cannot Feel Empathy

AI chatbots in mental health (e.g., Woebot) left users feeling isolated and disconnected — responses felt 'robotic' even when evidence-based. AI detects patterns; it cannot genuinely understand suffering.

Autonomy Must Be Preserved

Both patient autonomy (informed consent for AI care) and clinician autonomy (professional judgment) are frequently cited in research as threatened by AI-driven decision-making.

AI Augments — Never Replaces

Final decisions about diagnosis, treatment, and ethics require human judgment, clinical context, and compassion that no algorithm can replicate. The clinician remains irreplaceable.

OpenClaw: When AI Stops Answering and Starts Acting

OpenClaw is a free, open-source AI agent — not a chatbot. It connects AI models to your computer, your files, your email, and your EHR, and executes multi-step tasks autonomously. Launched late 2025, it reached 160,000 GitHub stars in weeks (more than the Linux Kernel). Its creator was acquired by OpenAI in early 2026. Gartner recommended enterprises block it on launch day.

Traditional AI Chatbot	OpenClaw AI Agent
Responds to one question	**Chains multi-step tasks autonomously**
Output = text only	**Reads files, sends emails, executes code**
You act on its suggestions	**It acts on your data — without asking**
No memory between sessions	**Persistent memory; wakes itself on schedule**
Cannot access external systems	**Connects to EHRs, calendars, billing APIs**

⚠ **Key risks:** no HIPAA compliance out of the box · PHI exposure · shadow IT (staff running it on work laptops with EHR access) · vulnerable to prompt injection hijacking

UAB MEDICINE

A Framework for Ethical AI in Healthcare: SERBS

Sensitivity

Recognize & mitigate bias across race, gender, age, and socioeconomic status. Monitor AI continuously for unfair outcomes.

Evaluation

Rigorous testing for fairness, safety, and societal impact — not just predictive accuracy. Assess before deployment.

Responsibility

Transparency, explainability, and accountability. AI decisions must be auditable and justifiable to patients and regulators.

Beneficence

AI must promote long-term well-being of individuals and society. Efficiency alone is not sufficient justification.

Security

Protect data integrity, patient privacy, and prevent unauthorized access. Assume breach; design for containment.

Source: Nasir, Khan & Bai (2024). Ethical Framework for Harnessing the Power of AI in Healthcare. IEEE Access.

You as the AI-Driven Leader

Maintain thought leadership — use AI as your thought partner, not your authority

The Thought Partner Prompt

"I want to maintain being the thought leader. I want you to be my thought partner. Help me think about this more deeply — ask me questions."

Why it matters:

Keeps YOU in control. AI challenges your thinking rather than replacing it.

The Ethics Check Prompt

"Before I implement this AI tool in my department, what are the ethical risks I may not have considered? Push back on my assumptions."

Why it matters:

Forces anticipatory ethics — surfaces blind spots before they become patient harm.

The Equity Lens Prompt

"Review this AI recommendation. Which patient populations might be systematically underserved or misclassified? What data would I need to verify this?"

Why it matters:

Embeds fairness into everyday clinical AI decision-making.

The question is not whether AI will transform medicine.

It already has.

The question is whether WE will shape it ethically — or let it shape us by default.

- Demand explainability from every AI tool in your institution
- Protect patient consent — they have the right to know when AI informs their care
- Interrogate the training data — who is represented, and who is not
- Stay the thought leader — use AI as a partner, never as the authority

Betsy Hopson, PhD, MSHA | University of Alabama at Birmingham | Department of Urology

Questions & Discussion