Collaborative AI in Healthcare:
Preserving Human Autonomy in Clinical Decision-Making

A White Paper on Ethical Design Constraints for Clinical Decision Support Systems

Authors: Jaydee Robles, DACM & FIDM+

January 2026 | Version 1.0

Executive Summary

Clinical artificial intelligence is increasingly deployed across triage, documentation, decision support, patient communication, and care operations. Yet ethical failure modes persist because many systems are designed as if clinical decisions are primarily computational rather than moral, contextual, and accountable. This paper advances a simple boundary condition: in healthcare, AI must support decisions, not make them. Ethical clinical AI should augment clinician judgment, preserve patient autonomy, and maintain unambiguous accountability for care decisions.

We synthesize the most common risk categories observed when AI is placed inside real clinical workflows: privacy and confidentiality failures, unsafe outputs, bias and inequity, autonomy erosion, and accountability diffusion. We then propose an evaluation lens rooted in widely used bioethical principles (autonomy, beneficence, non-maleficence, and justice) and operationalized through governance requirements such as auditability, incident response, documentation boundaries, and vendor accountability.

FIDM+ is presented as a conceptual ethical pattern rather than a technical specification. It illustrates an approach in which patient narrative is captured in a structured way, patient preferences and boundaries are explicitly collected, safety gates are embedded, and the clinician remains the accountable author of the care plan. The objective is not to claim that any single framework is sufficient, but to clarify what ethical structure looks like when the goal is collaborative, human-led care.

Contents

1. Introduction

Healthcare is a domain where decisions are expected to be explainable, accountable, and responsive to the values of the person receiving care. Clinical reasoning integrates evidence, patient history, exam findings, uncertainty management, and the lived context of the patient. Artificial intelligence can strengthen this process by reducing cognitive burden, improving access to relevant knowledge, and supporting documentation quality. However, when AI is used as a decision-maker, its outputs can be interpreted as authority rather than assistance, and the clinical relationship becomes vulnerable to ethical breakdown.

The clinical encounter is not a mere optimization problem. It is a trust relationship. Patients assume that someone is responsible for decisions, that privacy is protected, that recommendations are aligned with their goals and values, and that harms will be addressed through clear accountability. A central premise of patient-centered care is that clinical decisions should be guided by patient preferences, needs, and values. [1]

For clinical AI, the most important ethical design decision is role definition: whether the system is built to advise a clinician and patient, or whether it is built to decide. This paper argues that ethical clinical AI must be built to advise. The clinician remains accountable for the plan; the patient retains autonomy to accept, decline, or negotiate; and AI remains a tool whose outputs must be interpreted in context.

2. Thesis and Position: Augmentation Over Automation

Position statement: Clinical AI should function as decision support, not decision making. Ethical clinical AI must enhance clinician judgment, preserve patient autonomy, and maintain clear accountability for care decisions.

This boundary condition can be understood as a structural safeguard. In many domains, automation is justified when errors are tolerable and accountability can be distributed. In healthcare, errors can be catastrophic, trust is fragile, and accountability must remain legible. Even when an AI model is statistically accurate, the clinical question is never only "what is likely"; it is also "what is safe," "what aligns with this patient's goals," "what is feasible," and "who explains and owns the decision."

A collaborative design posture treats AI as a knowledgeable assistant that can surface options, summarize evidence, identify inconsistencies, and prompt clinicians to consider alternative hypotheses. It does not authorize the system to select diagnoses or prescribe without human verification. This approach protects both patients and clinicians by minimizing false certainty, limiting scope creep, and preventing the diffusion of responsibility.

3. Ethical Problems AI Creates in Real Clinics

Ethical risks emerge when AI is inserted into clinical workflows without healthcare-grade constraints. The following risk categories repeatedly appear in real clinical contexts, regardless of the underlying model architecture.

3.1 Core Risk Categories

- Privacy and confidentiality failures: protected health information (PHI) leakage, secondary data use, vendor exposure, and unauthorized model training on patient data.
- Clinical safety failures: hallucinated outputs, false certainty, inappropriate triage, and unsafe recommendations presented with undue authority.
- Bias and inequity: uneven performance across populations due to training data limitations and "standard patient" assumptions.
- Autonomy erosion: coercive framing, opaque reasoning, and recommendations that patients and clinicians cannot meaningfully challenge.
- Accountability diffusion: unclear responsibility for harm between the clinician, the software, and the vendor.

3.2 Why Generic AI Fails in Clinical Settings

Generic AI systems are typically optimized for broad usability, not healthcare ethics. They may lack robust consent capture, data minimization practices, predictable retention policies, and clear vendor roles. They may also lack clinically appropriate safety gating and escalation logic. In practice, these gaps produce two predictable outcomes: clinicians over-trust outputs because they appear fluent and confident, and patients misunderstand outputs as authoritative medical advice.

The ethical failure is not limited to "hallucinations." Even correct outputs can be unethical if they are delivered without appropriate consent, without respecting patient preferences, or without clarifying who is accountable. Ethical clinical AI therefore requires governance, documentation boundaries, and deliberate scope design, not merely improved accuracy.

4. Ethical Framework for Collaborative Clinical AI

Ethical clinical AI can be evaluated using established biomedical ethics principles: respect for autonomy, beneficence, non-maleficence, and justice. These principles have been widely influential in medical ethics and remain useful as a practical framework for assessing new technologies. [2][3]

4.1 Autonomy: Consent, Choice, and Contestability

Autonomy requires that patients retain meaningful control over their care decisions. In AI-assisted care, autonomy is protected when systems: (a) capture patient preferences explicitly, (b) make recommendations contestable, (c) allow patients to decline modules or questions without penalty, and (d) support informed consent rather than passive acceptance.

Autonomy is undermined when AI outputs are presented as inevitabilities ("you must"), when uncertainty is hidden, or when the clinician cannot explain why an output was generated. In collaborative care, explanation is part of consent: patients should understand the intent, limits, and rationale of recommendations at a level appropriate to their needs.

4.2 Beneficence: Demonstrable Clinical Value

Beneficence requires that AI materially improves the care experience or outcomes. In practice, this can include improvements in access, continuity, patient understanding, and clinical efficiency. However, beneficence is not established by novelty. It is established through evidence that the tool supports better clinical reasoning, safer triage, clearer documentation, or more effective shared decision-making.

4.3 Non-maleficence: Safety Architecture Over Helpful Language

Non-maleficence is the duty to avoid harm. In clinical AI, harm often occurs through predictable pathways: false reassurance, unsafe recommendations, delayed escalation, and misinterpretation of output authority. Ethical systems therefore require explicit safety architecture: red-flag detection, escalation rules, conservative defaults, and clear scope boundaries.

Non-maleficence also includes preventing harms caused by over-collection of data, unnecessary exposure of sensitive information, and the introduction of surveillance-like dynamics into care. A system that gathers excessive personal data without clear clinical utility increases risk without proportional benefit.

4.4 Justice: Equity, Access, and Bias Management

Justice requires equitable care. AI systems can amplify inequities if they perform unevenly across populations, if they encode biased assumptions, or if they are only accessible to privileged groups. Ethical clinical AI should include bias monitoring, inclusive intake language, and mechanisms to evaluate differential performance and impact across patient populations.

4.5 Fidelity and Trust: Legible Accountability

Fidelity is the obligation to be faithful to the clinical relationship. Trust depends on legible accountability: the patient should know who is responsible for decisions, how their data is used, and what recourse exists if something goes wrong. Trust is also supported when AI outputs are presented as suggestions with uncertainty, not as final decisions.

5. Privacy, Data Use, and Documentation Boundaries

Privacy is not merely a legal obligation; it is a clinical and ethical condition of care. Patients disclose sensitive information because they trust that it will be protected and used appropriately. When AI systems process patient data, privacy risks increase through vendor exposure, unclear retention policies, and secondary use incentives.

5.1 Baseline Expectations and Over-Compliance

At minimum, AI used in healthcare should align with applicable privacy and security requirements such as access controls, encryption where supported, audit logging, and appropriate contractual protections for vendors. But ethical systems should aim beyond baseline compliance by making data boundaries explicit: no sale of patient information, no model training on patient

data without separate authorization, and no secondary use that is not directly tied to care operations.

5.2 Documentation Ethics: Separating Story, Synthesis, and Tool Output

A common source of ethical confusion is documentation. Ethical AI-assisted care separates: (1) patient-reported information, (2) clinician interpretation and assessment, and (3) tool-generated suggestions. This separation protects patients from being mischaracterized and protects clinicians from inadvertently adopting tool language as clinical conclusions.

Controlled disclosure is also essential. When records are shared outside the immediate care relationship, the default should be to share clinically necessary summaries and clinician-authored notes. Proprietary tool outputs or internal workflows should not be disclosed unless specifically authorized or legally compelled. This protects intellectual property while keeping patient rights intact.

6. Governance and Evaluation for Ethical Claims

Ethical claims must be demonstrated, not asserted. Governance defines the policies, accountability structures, and monitoring processes that translate ethical intent into reliable practice. Evaluation defines how a system proves that it is safe, equitable, and autonomy-preserving.

6.1 Risk Management as an Ethical Requirement

Formal risk management frameworks provide a practical bridge between ethics and engineering. The NIST AI Risk Management Framework (AI RMF 1.0) frames trustworthy AI as a risk management discipline and emphasizes governance, mapping, measurement, and management across the AI lifecycle. [4][5]

In clinical contexts, risk management should include: (a) clearly defined intended use, (b) human oversight requirements, (c) monitoring for performance degradation, (d) privacy and security controls, (e) incident reporting and response, and (f) regular review of equity impact.

6.2 Lifecycle Monitoring and Clinical Context

AI performance can drift as clinical populations change, as data sources shift, and as practice standards evolve. For AI/ML-enabled medical devices, regulators have articulated Good Machine Learning Practice (GMLP) guiding principles that emphasize safety, effectiveness, quality management, and lifecycle controls. [6] Even when a tool is not regulated as a device, the underlying principle applies: clinical AI should be monitored and governed as a lifecycle system, not a one-time deployment.

6.3 Metrics That Reflect Ethics

Ethical evaluation requires metrics that reflect ethics rather than convenience. Examples include:

- Autonomy and understanding: patient-reported clarity of recommendations, perceived control over decisions, and satisfaction with shared decision-making.
- Safety: red-flag detection accuracy, escalation appropriateness, adverse event tracking, and rates of unsafe recommendations intercepted by clinician oversight.
- Equity: differential performance across demographic groups, language accessibility, and barrier analysis for access.
- Privacy: audit results, vendor compliance tracking, incident response time, and evidence of data minimization.
- Accountability: documentation clarity on who authored decisions, traceability of recommendations, and clinician ability to explain the rationale.

7. FIDM+ as a Conceptual Ethical Pattern

FIDM+ is positioned here as an example of ethical structure rather than a disclosure of proprietary technical mechanisms. The purpose is to illustrate what collaborative clinical AI looks like when it is designed to preserve human autonomy and clinician accountability.

7.1 Human-in-the-Loop as Primary Safety Mechanism

A central ethical constraint is that the clinician remains the accountable author of the plan. AI may summarize, organize, surface hypotheses, and suggest options, but the clinician decides. This role separation reduces false certainty and preserves a clear line of responsibility, consistent with the ethical requirement that clinical accountability remain legible.

7.2 Structured Narrative Capture Without Paternalism

A second ethical constraint is that care begins with the patient story and values, not with a fixed protocol. Structured narrative capture can help clinicians avoid omissions while still preserving the patient's voice. Importantly, structure should not become coercion. Patients should be able to skip questions, decline modules, and choose the degree of personal disclosure.

In integrative settings, this matters even more because patients may wish to incorporate multiple medical paradigms. FIDM+ is designed as an open-framework approach that can incorporate conventional biomedicine alongside traditional systems such as Traditional Chinese Medicine (TCM) and herbal medicine, where appropriate and aligned with patient preferences. This pluralistic posture does not require AI to decide what is true; it requires AI to help the clinician organize data and options while the clinician applies judgment and safety standards.

7.3 Safety Gates and Escalation

Safety gating is an ethical requirement when AI touches triage, symptom interpretation, or patient messaging. A collaborative system should identify urgent symptoms, prompt escalation to emergency care when indicated, and avoid false reassurance. These safeguards protect patients from delayed care and protect clinicians from unsafe scope drift.

7.4 Documentation Boundaries and Controlled Disclosure

FIDM+ treats documentation as an ethics mechanism. Separating patient-reported history from clinician assessment reduces the risk that tool-generated language is misread as a clinical conclusion. Controlled disclosure policies also protect patient privacy and clinical intellectual property by limiting what is shared externally to what is clinically necessary and authorized.

8. Conclusion

Ethical clinical AI is defined less by sophistication and more by restraint. The central ethical demand is not that AI be powerful, but that it be positioned correctly: as a collaborative decision support tool that enhances clinician judgment, preserves patient autonomy, and maintains clear accountability for care decisions.

As clinical AI expands, ethics must be engineered into the structure of systems: consent-forward data capture, privacy-first policies, safety gating, bias monitoring, and governance that supports lifecycle monitoring. The future of clinical AI will be shaped not only by model performance but by whether health systems can keep responsibility legible and preserve the human relationship at the center of care.

Author Statement

Jaydee Robles, DACM is the clinician-author and accountable decision-maker for clinical care decisions referenced in this framework. FIDM+ is presented as a structured, human-in-the-loop clinical support framework intended to strengthen documentation quality, ethical boundaries, and collaborative decision-making without substituting for clinician judgment or patient agency.

References

11. 1. Agency for Healthcare Research and Quality (AHRQ). Six Domains of Healthcare Quality: Patient-centered care definition. https://www.ahrq.gov/talkingquality/measures/six-domains.html
12. 2. Beauchamp TL, Childress JF. Principles of Biomedical Ethics. (Framework summarized in biomedical ethics literature). https://archive.org/details/principlesofbiom0000beau_k8c1
13. 3. Holm S. The four principles of Beauchamp and Childress. (Open-access discussion). https://pmc.ncbi.nlm.nih.gov/articles/PMC3528420/
14. 4. National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework
15. 5. NIST. AI RMF 1.0 publication (NIST.AI.100-1). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
16. 6. U.S. Food and Drug Administration (FDA). 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