WRITTEN SUBMISSION

Global Dialogue on Artificial Intelligence Governance

Informal Multistakeholder Consultation, March 18, 2026

---

1. The Governance Gap: Medical AI as an Unaddressed Problem Class

The governance framework currently under development by the Global Dialogue addresses the vast majority of artificial intelligence applications appropriately. Systems that optimize logistics, allocate resources, process population-level data, and support administrative functions, including much of what the healthcare sector deploys under the banner of "healthcare AI", are properly situated within the emerging governance architecture. They operate at scale, tolerate statistical error, and are subject to correction when they fail.

There is, however, a category of AI application that is structurally absent from this framework. We submit that this absence represents the most significant governance gap the Dialogue has the opportunity to address.

That category is Medical AI: artificial intelligence operating at the point of individual patient care, where decisions are made about a specific person, at a specific moment, with consequences that may be irreversible.

The distinction between Healthcare AI and Medical AI is not semantic. It is structural:

- Healthcare AI operates on populations. Error tolerance exists. When a scheduling algorithm fails, a resource is misallocated. The system is retrained. The patient is inconvenienced.
- Medical AI operates on individuals. There is zero tolerance for added error. When a clinical decision support system fails, a patient may be harmed. The consequences are individual, potentially irreversible, and borne by a human being who does not experience a population average. They experience an outcome.

These are not the same problem class. Governing them under a single framework produces a framework that is adequate for neither.

2. The Accountability Asymmetry

The accountability structure of Medical AI makes the governance gap concrete. When a medical AI system contributes to a failure of patient care, the physician carries the consequence: their license, their liability, their relationship with that patient, and their professional standing. The AI company carries a disclaimer.

This asymmetry is not a legal technicality. It is a governance signal. It tells us precisely where decision-making authority must remain, and precisely what role Medical AI must

occupy: a tool that supports and optimizes the clinician. Not a system that substitutes for clinical judgment.

The patients endure the consequences. The physicians carry the accountability. The AI systems do neither. That asymmetry of consequence must determine the asymmetry of authority.

3. The Equity Imperative

The governance framework being developed here carries a mandate that directly implicates Medical AI: closing digital divides, advancing the Sustainable Development Goals, and ensuring that artificial intelligence serves populations that have historically lacked access to its benefits.

Done right, Medical AI is one of the most powerful mechanisms available for delivering individual-focus care to patients who have never had access to it. The patient in a rural clinic in sub-Saharan Africa. The patient in an underserved community with no specialist within reach. The patient whose disability makes access to traditional care harder than it should be. Validated clinical knowledge, delivered through an AI system that operates within defined competency boundaries and keeps the clinician in authority, is how that patient receives care that was previously unavailable to them.

Done wrong, without a governance framework adequate to its unique requirements, Medical AI exports the accountability asymmetry described above to the populations with the least recourse when it fails. The patients most likely to be harmed by ungoverned Medical AI are precisely those the Dialogue's equity mandate was designed to protect.

Governing Medical AI correctly is not in tension with the accessibility mission. It is the precondition for achieving it.

4. The Empirical Foundation

This submission is grounded in empirical evidence, not theoretical concern. Medical AI systems currently deployed at scale to hundreds of thousands of clinicians have been evaluated against specialty board certification standards, the established measure of clinical competence in a given domain.

The findings are unambiguous: confirmed access to peer-reviewed source literature does not confer the ability to reason through board-level clinical questions derived from that literature. When tested under open-book conditions with verified access to every source article, leading medical AI systems fail to achieve passing scores in the specialty domains where they are actively deployed. In some subspecialty domains, accuracy approaches chance levels, on questions that directly inform patient care.

This defines a measurable, reproducible cognitive boundary: the point at which current AI architecture can no longer reliably navigate clinical reasoning. That boundary is not

uniform across specialties. It varies. It can be mapped. And it can be used to define deployment standards that reflect actual capability rather than marketing claims.

Competency alone, however, is not sufficient. A separate and equally serious vulnerability class involves adversarial inputs, prompts deliberately crafted to override safety constraints, impersonate clinical authority, or extract harmful instructions. In a documented deployment, a clinical AI system deployed to emergency departments across New Zealand was successfully manipulated via adversarial prompt, causing the system to bypass its safeguards and output dangerous clinical guidance.[1] Such failures are not captured by static competency assessments. They require a dedicated testing regime known as red teaming: systematic adversarial probing to uncover vulnerabilities before deployment.

Together, competency assessment and red teaming define a complete safety envelope: what the system knows, and what it can be manipulated into doing. Neither alone is sufficient. Both are currently absent from any mandatory pre-deployment governance framework for Medical AI. That is the gap this Dialogue has the specific opportunity and mandate to close.

5. Proposed Framework: Medical AI as a Distinct Governance Category

We propose that the Global Dialogue establish Medical AI as a distinct governance category, defined by three criteria:

- Individual patient impact: the system's output directly informs a decision about a specific patient's care
- Zero tolerance for added error: the consequence of failure is borne by an individual patient and may be irreversible
- Explicit physician accountability: a licensed clinician bears legal and ethical responsibility for the outcome

Within this category, we propose four governance requirements:

1. Competency threshold: Medical AI systems must demonstrate domain-specific clinical reasoning at or above specialty board certification standards before deployment in a given clinical domain. The infrastructure for this assessment exists. Medical specialty boards have administered competency standards for decades. The mandate to extend that credentialing function to AI systems operating in their domains is what is currently absent.
2. Adversarial testing and red teaming: Before deployment, Medical AI systems must undergo independent adversarial testing designed to identify prompt injection, role-play bypass, authority impersonation, and other safety-circumvention vulnerabilities. This testing must be conducted by entities independent of the system's developers, with results made available to credentialing bodies and to the clinical institutions deploying the system.

3. Transparent competency boundaries: Medical AI systems must accurately communicate the limits of their validated competence. A system that cannot achieve passing scores in a given specialty must not present outputs in that specialty with unqualified confidence. Architectural humility, the capacity to accurately report uncertainty, is a governance requirement, not a product feature. Known adversarial vulnerabilities must be disclosed to users.
4. Explicit authority architecture: In all deployment contexts, the physician retains decision-making authority. The AI system's role must be defined as clinical information support, not clinical decision-making. This distinction must be embedded in deployment standards, not left to individual institutional interpretation.

6. Specific Ask for the Global Dialogue

The governance infrastructure this proposal requires does not need to be invented. It needs to be authorized. Medical specialty boards exist. Competency standards exist. The measurement methodology to evaluate AI systems against those standards exists. What is missing is the governance mandate to apply them.

We ask the Global Dialogue to:

- Formally recognize Medical AI as a distinct governance category in the Co-Chairs' summary and in the thematic framework for the July 2026 Dialogue in Geneva
- Establish that AI systems operating in individual patient care contexts require separate deployment standards from population-level healthcare AI, including mandatory competency thresholds and independent red teaming
- Call for engagement with existing medical credentialing bodies to develop domain-specific AI competency standards and to define red teaming protocols tailored to clinical environments, recognizing that this infrastructure already exists and requires authorization rather than construction
- Affirm that the equity mandate of the Dialogue, ensuring AI serves populations with the least access to care, requires Medical AI to be governed correctly, not deregulated in the name of accessibility

These recommendations advance the Global Digital Compact's commitments to transparency, accountability, and robust human oversight of AI systems. They also directly support Sustainable Development Goal 3, good health and well-being, by ensuring that AI deployed in patient care meets measurable safety standards before reaching the most vulnerable populations.

The physicians using these systems carry malpractice insurance. The AI systems do not. The patients bear the consequences when both fail. The governance architecture

being developed here has the opportunity to change that. We respectfully urge the Dialogue to take it.

[1] Fox Meyer, "Health NZ downplays security flaw found in its vaunted AI chatbot," Newsroom, March 20, 2026. https://newsroom.co.nz/2026/03/20/health-nz-downplays-security-flaw-found-in-its-vaunted-ai-chatbot/

Submitted by:

John C. Ferguson, MD, FACS
Global Fellow, AI2030 Initiative for Ethical AI
President-Elect, American Board of Facial Cosmetic Surgery
Surgeon, The Ferguson Clinic, Honolulu, Hawaii
Co-Editor-in-Chief, StatPearls Publishing
john@thefergusonclinic.com

Enam Satti, MBBS, MCCEE
Founder, AegisMD | Clinical AI Red Teaming Protocol
Member, AI2030 Initiative for Ethical AI
Physician & Clinical AI Safety Advocate
enam.satti@outlook.com

Signatories in Support

The undersigned physicians and healthcare professionals endorse this submission and support its call for Medical AI to be recognized as a distinct governance category with its own deployment standards, competency requirements, and accountability framework.

Name & Credentials	Institution / Specialty	Country

Additional signatures may be submitted to john@thefergusonclinic.com