Comment

ILLUSTRATION NICO BRAUSCH

Tackling the blurring lines between physicians and AI

Kyle Lam, Mindy Nunez Duffourc, Jiankai Sun, Eric Topol & Jianing Qiu

A staging system based on how big a role AI has in patient care can help to identify who is responsible when treatment fails.

Artificial intelligence is fast becoming embedded in hospitals and healthcare clinics. Yet, with the benefits of AI tools come fresh patient-safety risks – and questions about who is responsible when things go wrong.

Conventionally, responsibilities in health care are clearly delineated. Clinicians must provide a set standard of treatment. Institutions must organize safe treatment processes. Manufacturers must deliver non-defective medical devices. Regulators and bodies involved with licensing and credentials can intervene if any of these parties fall short of expected standards. And courts can assess which party is liable if patients are harmed.

An incoming wave of medical AI tools is set to blur these lines.

Current AI tools are being used mainly as assistants, backed up by human checks, much as with other medical devices. For instance, when AI models trigger an alert that a patient is at risk of sepsis, a clinician must review and confirm the physiological rationale before any treatment.

But in the next few years, AI-driven clinical tools are expected to advance from synthesizing data to acting on it, with less and less human input. They will make diagnoses, devise treatment plans and make patient-management

decisions that clinicians play little or no part in. Whereas the outputs of many existing tools are understandable – sepsis models, for instance, work by combining defined physiological variables according to a transparent formula[1] – more-advanced tools can involve black-box deep-learning processes. Clinicians will no longer be able to fully follow the tools' reasoning, even for algorithms that provide some explanations for their judgements[2].

Such tools land between accountability rules. Their black-box reasoning makes it hard to determine when they are defective. And when decisions are made jointly by humans and an opaque AI model, it becomes difficult to assess responsibility when people are harmed while receiving health care. Existing medical-liability frameworks do not address this crossover point[3].

This legal uncertainty means that people who are harmed could fall into liability gaps in which no one has clearly broken a rule[3]. Hospitals are likely to avoid AI for potentially useful functions because of concerns about legal risks – concerns that are repeatedly cited by physicians as a barrier to safe AI adoption[4–6] and that are seen by some as grounds not to use black-box tools already available. And AI vendors might shirk their obligation to monitor safety once a tool reaches the market, confident that it will be hard to attribute blame when things go wrong.

Clear liability frameworks are needed.

Here we outline how to achieve that, by defining seven levels of AI capabilities on the basis of three factors: autonomy, automation and operational scope. With aircraft and self-driving cars, regulators already use graded levels to specify exactly what tasks systems must perform, how independently, and when humans are expected to take over[7–9]. A similar set of levels for medical AI systems would help regulators, policymakers, governments, courts and professional health-care bodies to plan for the incoming tools.

Three properties, seven levels

Three properties dictate how AI is used in the clinic.

First, autonomy: how independently the system reasons, without the need for human input. Existing sepsis risk models are low-autonomy AI tools[1]. Deep-learning systems, by contrast, might arrive at diagnoses from patterns across thousands of variables, making inferences that physicians cannot retrace[10].

Second, automation: what tasks a medical AI system can execute on its own. Such a system might range from a low-automation robot that a surgeon controls, to a system that sends chemotherapy prescriptions directly to a pharmacy.

Third, operational scope: the boundaries in which the system is allowed to run. A retinal-imaging system deployed by a specialist ophthalmology service to double-check clinical decisions is less risky than the same tool used in a primary-care clinic to make an autonomous decision about whether to refer a patient to a specialist.

KLAUDIA RADECKA/NURPHOTO VIA GETTY

A scanner being trialled in Krakow, Poland, uses AI to help to detect breast cancer.

The following seven levels could form a basis for defining risks around medical AI tools. A given tool might sit at different levels depending on where and how it is used, and by whom.

Level 0: Informational. An AI system with no autonomy, variable automation and narrow operational scope. Such systems store, move and display data according to rules set by humans. Examples include algorithms for managing electronic health records or streaming data from wearable devices that track vital signs. Algorithmic behaviour is transparent and any output is independently verifiable.

This is well-trodden ground. Regulators can assess whether such tools are reliable and offer sufficient cybersecurity and data protection. Courts can apply conventional standards to assess liability.

Level 1: Assistive. AI systems with low autonomy, low-to-moderate automation and narrow operational scope. Such systems can perform a single, clearly defined task, overseen by a clinician – such as flagging suspicious rhythms on a heart monitor[11] or highlighting possible lung nodules on a scan[12]. The tools' algorithmic reasoning might be opaque, but the output can be verified through clinical review of patient data, or by 'explainability techniques' that check correlations between the model's input and output data.

These tools are regulated as diagnostic aids, similar to existing computer-aided detection systems for mammography and colonoscopy. Hospitals are held responsible for ensuring that staff understand their limitations. Users are expected to accurately judge whether alerts are meaningful.

Level 2: Decision support. AI tools with low autonomy, low-to-moderate automation and moderate operational scope. These systems combine multiple streams of data – perhaps including symptoms, previous medical history, imaging data and current guidelines. Existing tools at this level can produce a diagnosis and treatment plan, draft detailed discharge summaries and generate risk scores for triage decisions. As with level 1 tools, they are designed to be advisory – their reasoning might be opaque but is verifiable to some extent, and clinicians are expected to make the final decision.

Regulators should ask for robust evidence that these systems improve care. Hospitals must train clinicians to challenge outputs instead of accepting them by default, and courts must decide whether a clinician's reliance on a particular recommendation was reasonable in the circumstances. There's a risk that clinicians under time pressure will over-rely on advisory tools. There will be no easy way to determine how much a clinician was swayed by a machine's recommendation.

Level 3: Supervised automation. AI systems with moderate levels of autonomy and automation, and moderate operational scope. The system, rather than the clinician, is the decision maker. These tools are currently rare (see 'Classifying current medical AI'). In a tightly defined domain, they can dynamically

respond to treatment needs – for example, devices for insulin delivery automatically adjust doses according to continuous glucose readings, using embedded algorithms and predefined limits[13]. The clinician selects which patients are placed on the system and intervenes only when something seems wrong.

Regulators should focus on the quality of instructions for using and maintaining these systems and on information about the risks inherent in providing care. Health-care organizations should furnish guidance for informed-consent protocols that respond to the new types of risk inherent in AI-driven treatment, and providers should discuss these risks with patients.

With these systems, physicians can be criticised for enrolling an unsuitable patient, failing to respond to device alarms or not understanding system limitations. Organizations might be judged on how they selected, validated and monitored the technology. Manufacturers might be liable if flawed designs or updates contribute to harm. But conventional liability assessments can be unfit for purpose, if physicians and organizations acted reasonably, and harm is caused by an algorithmically determined decision that is not obviously wrong.

Level 4: AI-triggered human oversight. AI tools with moderate-to-high levels of autonomy and automation, and broad operational scope. These systems manage whole episodes of therapy in areas such as intensive care. They continuously monitor patients and adjust treatments – for instance balancing ventilator and sedation settings against several physiological targets at once. They trigger back-up protocols such as summoning a rapid-response team when something seems wrong. Clinicians are a safety net and act only when the system sounds the alarm. Level 4 deployments are yet to break through in medicine, although the technical components needed are emerging.

For these systems, regulators should require strong evidence that including a human in the loop would reduce safety. Given that clinicians would have no realistic opportunity to review the actions of these systems, courts might struggle to determine whether an adverse outcome reflects negligence, a hidden defect or simply the inherent risks of medicine.

Level 5: Autonomous clinician. These AI systems have high levels of autonomy and automation, with broad operational scope. Level 5 is the medical equivalent of fully autonomous – a system licensed to handle the full spectrum of diagnosable conditions and routine treatments, across a range of settings. Level 5 tools could take histories, arrange and interpret investigations, prescribe and titrate drugs and manage complications. Humans retain only the authority to intervene when needed. No real-world system currently comes close to this.

Regulators should require clear limits on what conditions such a system can tackle and the populations on which it is deployed, because the absence of routine human review removes a major safety net. Conventional ideas of negligence completely unravel here. If the system has been properly approved, deployed for its prescribed indications and used as intended, and if its recommendation aligns with current guidelines, then an adverse but unforeseeable outcome might not be the result of a defect or anyone's fault.

"Hospitals are likely to avoid AI for potentially useful functions because of concerns about legal risks."

Level 6: Self-teaching medical AI. Such systems have extremely high autonomy and automation, with broad operational scope. They both deliver fully autonomous care and update their own practices as fresh data and research emerge. Such tools – if ever developed – might discover disease subtypes, design drug regimens or devise therapies that have never been trialled in humans.

Whether this should even count as clinical AI, rather than as a system to be used for research and development, is an open question. If someone were harmed by an innovative treatment pathway with no human involvement, it's unclear who should be liable: the developers; the hospital activating the self-teaching system; or the broader health system that benefited from its successes.

Next steps

To prepare for an influx of tools at level 3 and above, the following changes are needed.

Regulators should begin to treat automation, autonomy and operational scope as discrete concepts. They might provide a tool entering the market with its initial grading, working from its intended uses and the evidence the developer submits. Many tools are multipurpose and should receive a separate level for each stated function.

This nuanced grading system will allow regulators to further refine their rules around AI, including frameworks from the US Food and Drug Administration (FDA) and European Union regulators concerning AI as a medical device, and cross-sectoral regulations on AI technology, such as the EU AI Act. A seven-level grading system could be used to set approval conditions, post-market monitoring obligations and clinician-oversight requirements that better respond to risks, depending on the role of the AI tool.

To guard against developers pressuring regulators to assign their tools the lowest plausible grade, a tool's overall classification should be at the highest level that any of its intended or reasonably foreseeable uses would warrant. Reclassification downwards should require evidence from the developer that the lower-risk use is the only one realistically possible for that deployment.

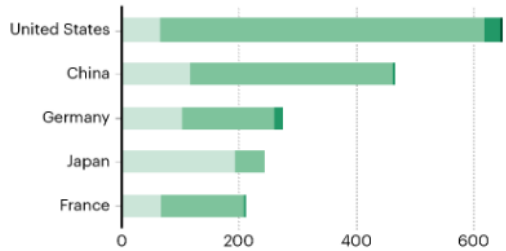
Regulators should also specify when and how humans must review tools, and should detail what meaningful review looks like for a high-autonomy system, for which a clinician cannot reconstruct an algorithm's reasoning. This might involve checking that the tool is being deployed in appropriate settings, auditing performance and analysing how often humans need to override the system. Changes in where and how a tool is used – for example, if a tool originally approved for specialist centres is to be made available in remote clinics – should trigger fresh scrutiny.

Health-care organizations will need clear guidance from regulators and licensing bodies around system selection, monitoring and governance. At a minimum, health-care providers should validate a tool's performance in their own populations, educate staff on AI limitations, maintain logs of AI-driven decisions and establish mechanisms to monitor and intervene when systems develop problems. They should also audit the contexts in which a tool is used, to determine the level at which they are using it. This information will then inform any safety review or liability case.

Lawmakers and courts should recognize that current rules make it difficult to assess

CLASSIFYING CURRENT MEDICAL AI

Among 2,825 currently approved medical devices* that have artificial-intelligence features, none ranks higher than level 3 in a seven-level taxonomy of AI capabilities. The breakdown below covers the five countries with the most device approvals.



Definition of levels

- **Level 0:** Informational – Able to store, move and display data
- **Level 1:** Assistive – Performs one task, overseen by a clinician
- **Level 2:** Decision support – Can combine data streams and suggest diagnoses and treatment plans
- **Level 3:** Supervised automation – Manages specific treatment regimens, such as insulin delivery

*Data from the US Food and Drug Administration, the European Database on Medical Devices and China's National Medical Products Administration.

SOURCE: ANALYSIS BY K. LAM *ET AL.*

whether a black-box system is defective. Harms involving AI might require changes that enable a legal assessment of a tool's 'behaviour'. The approach should consider how much control each party — clinician, manufacturer, health-care organization — has over the tool and whether that party is in a position to take actions that prevent future harm. Lawmakers could also require all stakeholders to be insured against AI-driven harm, and require private and public insurers to integrate AI coverage into product, professional liability and general commercial policies.

The opportunity to shape safe and responsible use of medical AI is now, while AI deployments remain at relatively low levels and before health systems become dependent on highly autonomous tools. Delaying action until level 3 and 4 systems are in routine use could lead to ad hoc crisis responses from courts, legislators and policymakers, and make it harder to draw clear lines of responsibility. Acting fast will enable the responsible development, regulation and deployment of AI technologies that can improve patient care.

The authors

Kyle Lam is an academic clinical lecturer in the Department of Surgery and Cancer, Imperial College London, UK. **Mindy Nunez Duffourc** is assistant professor in the Faculty of Law, Maastricht University, the Netherlands. **Jiankai Sun** is a PhD candidate in the School of Engineering, Stanford University, California, USA. **Eric Topol** is director and founder at the Scripps Research Translational Institute, La Jolla, California, USA. **Jianing Qiu** is assistant professor of personalized medicine in the Biological and Life Sciences Division, Mohamed bin Zayed University of Artificial Intelligence, Masdar City, United Arab Emirates.
e-mail: Jianing.Qiu@mbzuai.ac.ae

1. Usman, O. A., Usman, A. A. & Ward, M. A. *Am. J. Emerg. Med.* **37**, 1490–1497 (2019).
2. Saporta, A. *et al. Nature Mach. Intell.* **4**, 867–878 (2022).
3. Duffourc, M. N. *Ill. J. Law Tech. Pol.* **2023**, 1–49 (2023).
4. American Medical Association. *2026 Physician Survey on Augmented Intelligence* (2026).
5. Stroud, A. M. *et al. JMIR Ment. Health* **12**, e64414 (2025).
6. Duffourc, M., Møllebæk, M., Druedahl, L. C., Minssen, T. & Gerke, S. *Ann. Surg. Open* **6**, e542 (2025).
7. Monkhouse, H., Habli, I., McDermid, J., Khastgir, S. & Dhadyalla, G. in *Proc. 2017 IEEE SmartWorld* https://doi.org/10.1109/UIC-ATC.2017.8397595 (IEEE, 2017).
8. Anderson, E., Fannin, T. & Nelson, B. in *Proc. 2018 IEEE/AIAA 37th Digit. Avion. Syst. Conf. (DASC)* https://doi.org/10.1109/DASC.2018.8569280 (IEEE, 2018).
9. SAE. *SAE International Recommended Practice, Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles* (SAE, 2021).
10. Sahashi, Y. *et al. NEJM AI* **2**, AIoa2400948 (2025).
11. Poterucha, T. J. *et al. Nature* **644**, 221–230 (2025).
12. Setio, A. A. A., Jacobs, C., Gelderblom, J. & van Ginneken, B. *Med. Phys.* **42**, 5642–5653 (2015).
13. Nimri, R. *et al. Nature Med.* **26**, 1380–1384 (2020).

The authors declare no competing interests.

Computers built using human brain cells need proper donor consent

Nada S. Salem, Jianping Fu, Feng Guo, Jonathan Kadmon, Omer Revah, Orly Reiner & Insoo Hyun

Researchers are exploring whether cultures of human tissue can be used in AI computing applications. But a key ethical issue has been overlooked.

The workings of the brain are often compared to those of a computer. Some researchers are testing this analogy to its limits, attempting to build computers containing laboratory-grown cultures of brain cells. In doing so, they hope to circumvent the physical limitations of the silicon-based computing hardware used for today's artificial intelligence[1,2].

The promise is there. Brain tissue combines memory and computation in the same substrate, requiring fewer components than conventional computers do. Theories of neural networks, based on neurons and synapses, have inspired AI algorithms. Brains consume much less energy than AI machines do[3] and don't require vast cooling systems.

Such biocomputing does have distinct ethical challenges. Discussions have raged over the morality of using lab-grown brain tissue and the potential for consciousness[4]. Yet one key issue has been absent from debates so far: the lack of explicit consent for the use of human neural tissue in biocomputing.

Research in biocomputing relies on voluntary donations of cells, often from people undergoing medical procedures who give them altruistically to aid biomedical and basic biological research. Yet, consent forms don't ask donors what sort of research they would like their cells to be used for in the future — or, more importantly, all the possible ways that they wouldn't like them to be used. Donors who want to support the development of cancer treatments might not imagine that their tissues could be used to create biocomputers, possibly with commercial uses.

A more explicit and nuanced approach to donor consent is needed for biocomputing research that uses lab-grown brain tissue. In our view, researchers should avoid using cell lines for which consent was given only for biomedical research. A system must be established for seeking consent retrospectively, or for reviewing the ethics of projects when obtaining consent again (known as re-consent) is not possible.

Using brain cells to compute

Brain-based biocomputing relies on human cortical neurons because they are larger and form more complex circuits than do those from rodents or other primates. These features are thought to contribute to the remarkable information-processing capabilities of the human brain[5,6], from recognizing a face to performing abstract reasoning.

BRAIN-BASED BIOCOMPUTING

A 3D neuronal culture called an organoid receives electrical inputs and generates outputs. Activity in the network can adapt over time, enabling simple learning behaviours.