HealthManagement
Promoting Management and Leadership

# AI Risks in Healthcare: From Bias to Existential Threats

Artificial intelligence is already deeply embedded in healthcare systems, shaping decision-making processes from diagnosis to treatment recommendations. The emergence of foundation model-based chatbots and large language models has intensified the integration of AI into clinical workflows. However, alongside the growing adoption of AI comes an expanding landscape of risk. While many current discussions focus on near-term concerns such as fairness, privacy and transparency, these same risks could evolve into existential threats. Misaligned algorithms, data vulnerabilities, automation bias and monopolistic control of AI tools not only compromise patient outcomes but could also destabilise healthcare systems and society at large. The convergence of these factors demands urgent attention, particularly in the context of healthcare, where the stakes are inherently high.

## AI Misalignment and Systemic Inequities

AI alignment involves ensuring that a system's goals correspond to the ethical and operational values of humans or institutions. Misalignment, whether through incorrect goal specification or inconsistent application, poses a significant danger. In healthcare, misaligned AI systems have already demonstrated their potential for harm. For instance, algorithms designed to identify patients needing additional care have exhibited racial bias, favouring less ill white patients over sicker black patients. Such disparities also emerge in diagnostic tools, where underdiagnosis in low-income and minority populations has been documented. These outcomes not only perpetuate existing social inequalities but risk reinforcing them at scale.

As AI becomes more deeply entrenched in healthcare infrastructure, its missteps could trigger broader public distrust, sparking civil unrest or systemic collapse. Mitigating such risks requires robust oversight mechanisms. Scalable human oversight, mechanistic interpretability and inverse reinforcement learning offer avenues to monitor and adjust AI behaviour. However, effective oversight hinges on the development of more intuitive and transparent interfaces between humans and AI systems. Addressing these alignment challenges is essential to prevent a future where healthcare becomes a vehicle for embedded structural biases.

***Must Read:*** ***Balancing Innovation and Risk: Managing AI in Healthcare Cybersecurity***

## Safety, Privacy and the Cost of Overtrust

The ability of adversarial actors to circumvent safety measures in large language models illustrates the fragility of current AI defences. Models can be manipulated to generate inappropriate or harmful content, reproduce personally identifiable information or act unpredictably due to Trojan attacks. The Dinerstein v. Google case highlights the persistence of privacy threats, even when using de-identified patient data. These breaches not only threaten individual confidentiality but raise the spectre of large-scale surveillance or exploitation. In the worst-case scenario, such vulnerabilities could be exploited for bioterrorism or mass destabilisation.

In parallel, the phenomenon of automation bias—or excessive trust in AI systems—amplifies these risks. Medical professionals may over-rely on AI for critical decisions, bypassing necessary human judgement. This is particularly troubling when AI models generate highly convincing misinformation or unexpected outputs, such as detecting race in medical images, which raises concerns about hidden biases. While explainability techniques can clarify system reasoning, they also risk fostering overconfidence in flawed models. Misinformation, particularly during global crises like the COVID-19 pandemic, can undermine public compliance with health measures. Therefore, strategies such as hierarchical reinforcement learning, red teaming and AI debates are essential to mitigate the consequences of overtrust and enhance system robustness.

## Transparency, Monopolisation and Societal Disruption

A small number of corporations currently dominate the development of powerful AI models, restricting access and reducing transparency. Proprietary systems like ChatGPT or Gemini operate with limited public scrutiny, increasing the risk of unethical application, including patient exploitation or biosecurity threats. In molecular biology, for instance, advanced AI could be harnessed to engineer novel pathogens. Without transparency, such scenarios are difficult to anticipate or prevent.

By contrast, open-sourcing AI technologies enables broader evaluation, fosters innovation and reduces dependency on monopolistic players. Models such as Meta's LLaMA have contributed to a more inclusive AI ecosystem, lowering barriers to entry and encouraging community-wide scrutiny. Nonetheless, the proliferation of open-source tools also necessitates strong regulatory frameworks to address ownership rights, misuse and safety compliance. In the absence of such governance, AI development may accelerate unchecked, increasing the likelihood of societal harm. Economic displacement and enfeeblement, whereby humans lose the ability to make autonomous decisions, are real possibilities. Unregulated AI integration could deepen inequalities, marginalise vulnerable communities and centralise power in dangerous ways. These outcomes not only threaten public health but also the societal structures that support it.

AI holds vast potential to enhance clinical care, from early diagnosis to personalised treatment plans. Yet, as these technologies evolve, so do the risks they pose. Misalignment, privacy breaches, overtrust and monopolisation are not isolated concerns—they are interconnected threats that can compound into existential dangers. Healthcare, as one of the most sensitive and critical domains, is uniquely vulnerable to these developments. To preserve both patient wellbeing and societal stability, a coordinated global response is essential. This includes creating diverse datasets, strengthening oversight mechanisms, promoting transparency through open-source models and enforcing robust governance. Failing to act on these fronts may compromise not only the promise of AI in healthcare but the future of healthcare itself.

**Source:** **BMJ Health & Care Informatics**

**Image Credit:** **iStock**

Published on : Tue, 13 May 2025