SPRITE+ Explainer #006

Series Editor: Mark Elliot

Trust, Identity, Privacy, Security and Safety (TIPSS) in Healthcare AI

By Abir Awad, Edward Apeh, Yang Lu and Pejman Saeghe

This Explainer focuses on Healthcare AI from the perspective of digital Trust, Identity, Privacy, Security and Safety (TIPSS). It explains what Healthcare AI is and highlights key possibilities for developments, outlines potential methods and implications.

What is AI in Healthcare?

Healthcare AI refers to the use of computational systems and algorithms including machine learning, deep learning and increasingly generative AI to support clinical decision-making, streamline administrative workflows with the aim of improving patient outcomes. Not yet a universal breakthrough, a balanced reading of evidence suggests notable advances in specific tasks (e.g., image analysis, documentation support, and workflow optimisation) when systems are properly validated and embedded in clinical workflows. Examples of current and emerging applications include:

- Medical diagnostics (e.g. radiology, dermatology and pathology)
- Predictive analytics for prognosis, and population health
- Precision medicine (genomics interpretation, cohort selection)
- Operational optimisation (bed management, theatre scheduling, supply chain)
- Drug discovery and development
- Robotic assisted surgery
- Virtual health assistants (see Topol, 2019; Helen, 2024 and Tetteh, 2025 for more information).

Table 1 in the Appendix provides a comparative overview of how AI is used in healthcare in different countries, highlighting regional priorities, implementation examples, and strategic focus areas. This integration of AI into healthcare has resulted in its increasing application in areas such as clinical decision support, patient engagement, and operational processes. Nonetheless, the rapid expansion of these technologies has prompted ongoing concerns regarding governance, security, safety and ethics.

Critical issues for TIPSS

The adoption of this technology introduces risks that must be critically assessed. A useful lens for evaluation is TIPSS which examines five key domains: Trust, Identity, Privacy, Security, and Safety.

The rapid integration of AI into healthcare systems introduces critical challenges across all five TIPSS domains.

Below we suggest how each might be strengthened by clarifying typical risks, providing concrete healthcare examples, and outlining practical mitigations that align with regulators and assurance bodies.

Trust

To foster trust among clinicians, patients and regulators, healthcare AI systems must be transparent, explainable and accountable. Opaque decision-making and

lack of interpretability erode confidence and hinder adoption.

Building trust requires explainable AI techniques, tailored to clinical contexts, clear audit trails for decision making, and participatory design involving end users (Leslie, 2020). Evidence generation should extend beyond laboratory accuracy to multi-site trials and ongoing monitoring, ensuring that systems perform reliably in real-world settings.

Identity

Robust identity management is essential to prevent unauthorised access, ensure traceability and maintain system integrity. Weak authentication mechanisms can expose patient records and allow unauthorised changes to AI models.

To mitigate these risks, healthcare systems are adopting federated identity systems, biometric authentication, and zero-trust architectures (Auth0, 2023). These approaches reduce identity fraud and support interoperability across distributed care environments and ensure secure access to sensitive systems.

Privacy

Dynamic consent platforms and robust governance models further empower patients while ensuring compliance with legal and ethical standards.

AI systems often require large datasets, raising significant concerns about data protection, consent and reidentification, particularly when data crosses national borders. Privacy-by-design approaches combined with technical measures like differential privacy (Dankar & El Emam 2013) and synthetic data generation (Gonzales et all 2023), can reduce these risks. For generative AI tools, additional safeguards like data residency controls and prevention of model memorization are critical. Dynamic consent platforms (Morley et al., 2021) and robust governance models empower patients while protecting patient data, ensuring legal compliance and supporting the ethical deployment of AI.

Security

Healthcare AI systems are vulnerable to cybersecurity threats such as adversarial attacks that manipulate inputs and data poisoning that corrupts models. Ransomware incidents targeting hospital health systems (IBM 2024) highlight the vulnerability of interconnected AI services.

To strengthen security measures organisations should implement adversarial testing, secure deployment protocols, and continuous monitoring either human-led or automated (Red Hat, 2023). Incident response plans and strategies such as fallback to human-only workflows during incidents will help maintain system integrity and support resilience.

Safety

Ensuring AI systems do not cause harm is a foundational requirement in healthcare. This goes beyond initial validation; it requires ongoing vigilance. Bias and performance drift (Bayram et al 2025) can lead to inequitable or unsafe outcomes, while generative models may also introduce unpredictable errors. Safety can be promoted through human-in-the-loop oversight (Kumar et al 2024), rigorous model validation, and real-world performance monitoring (Mennella et al., 2024). Prospective hazard analysis (Potts et al 2014) and structured change-control plans help manage adaptive software updates, ensuring that AI systems remain reliable and aligned with clinical standards throughout their lifecycle.

Given the complexity and potential impact of AI in healthcare, it is essential to conduct ongoing risk assessments and implement comprehensive mitigation strategies within the TIPSS framework. By addressing these dimensions holistically, stakeholders

can ensure that AI is deployed in a manner that is secure, ethical, and aligned with the values of modern healthcare systems.

The regulation and assurance landscape

Regulatory bodies such as the Medicines and Healthcare products Regulatory Agency (MHRA) are actively developing frameworks to ensure that AI technologies are deployed responsibly and safely within healthcare settings (NHS Confederation, 2021) and several key developments are shaping this:

- **Regulatory Innovation:** The MHRA's AI Airlock sandbox enables controlled testing of AI powered medical devices prior to full scale deployment (GOV.UK, 2023).
- **Global Collaboration:** Initiatives like the HealthAI Global Regulatory Network aim to harmonise international standards for AI safety, efficacy and interoperability (NHS Confederation, 2021).
- **Ethical AI Strategies:** National health bodies are advocating for coordinated strategies to ensure that AI is used responsibly and equitably across healthcare systems (Health Foundation, 2023).
- **Technical Advancements:** Emerging tools such as model cards, hazard tracking systems, and secure APIs are being developed to support transparency and traceability and system integrity (Red Hat, 2023).
- **Identity Evolution:** New identity solutions including passkeys, biometrics, and federated identity systems are being integrated to secure access and protect sensitive patient data (Auth0, 2023).

A Potential Future

Over the next decade, healthcare AI is poised to develop from experimental deployments to deep systemic integration across care pathways. Emerging innovations — such as generative AI for clinical documentation, predictive analytics for population health and real time decision support will make healthcare systems more personalised, proactive, and efficient.

As digital health ecosystems mature, AI will increasingly serve as the connective tissue of national health infrastructures, linking hospitals, community care, and research in real time. As part of this, it is vital that healthcare organisations embed Trust, Identity, Privacy, Security, and Safety principles throughout the entire AI lifecycle — from data design and model training to clinical deployment and post-market monitoring.

In a positive scenario we might imagine that:

- **Trust** is strengthened through transparent, explainable AI models and public reporting of outcomes.
- **Identity** is secured through strong authentication, verifiable digital credentials, and role-based access to clinical data and AI systems.
- **Privacy** is reinforced through federated learning, secure data enclaves and data-based methods to protect sensitive patient information.
- **Security** is enhanced as AI systems are used not only to deliver care but also to detect and respond to cyber-threats in real time.
- **Safety** becomes an adaptive process, with AI continuously learning from clinical incidents to prevent harm and improve reliability.

If realised, this pathway would position AI not only as an enabler of healthcare innovation but also as a guardian of its ethical and operational boundaries – ensuring that technological progress aligns with the core values of care equity and safety.

References

Auth0 (2023) Securing Trust in AI: Identity Guide for Developers. Available at: https://auth0.com/blog/securing-trust-in-ai/.

Bajwa, J., Munir, U., Nori, A. and Williams, B. (2021) Artificial intelligence in healthcare: transforming the practice of medicine. Future healthcare journal, 8(2), pp.e188-e194.

Bayram, F., & Ahmed, B. S. (2025). Towards trustworthy machine learning in production: An overview of the robustness in mlops approach. ACM Computing Surveys, 57(5), 1-35.

China Briefing (2025) China's AI Healthcare Market (Part I): Growth Trends, Drivers, and APAC Comparison. Available at: https://www.china-briefing.com/news/chinas-ai-healthcare-market-growth-trends-drivers-and-apac-comparison/.

Cuggia, M. and Combes, S. (2019) The French Health Data Hub and the German Medical Informatics Initiatives: two national projects to promote data sharing in healthcare. Yearbook of medical informatics, 28(01), pp.195-202.

Dankar, F. K., & El Emam, K. (2013). Practicing differential privacy in health care: A review. Transactions on Data Privacy, 6(1), 35-67.

NHS England. (2025) Artificial intelligence (AI) and machine learning. Available at: https://www.england.nhs.uk/long-read/artificial-intelligence-ai-and-machine-learning/.

Gonzales, A., Guruswamy, G., & Smith, S. R. (2023). Synthetic data in health care: A narrative review. PLOS Digital Health, 2(1), e0000082.

GOV.UK (2023) UK MHRA Leads Safe Use of AI in Healthcare. Available at: https://www.gov.uk/government/news/uk-mhra-leads-safe-use-of-ai-in-healthcare.

Health Foundation (2023) Priorities for an AI Strategy in Healthcare. Available at: https://www.health.org.uk/publications/long-reads/priorities-for-an-ai-in-health-care-strategy.

Helen, D., et al. (2024) Generative AI in healthcare: Opportunities, challenges, and future perspectives. Revolutionizing the Healthcare Sector with AI, pp.79-90.

IBM (2024) Ransomware on the rise: Healthcare industry attack trends 2024. Available at: https://www.ibm.com/think/insights/healthcare-industry-attack-trends-2024.

Kumar, S., Datta, S., Singh, V., Datta, D., Singh, S. K., & Sharma, R. (2024). Applications, challenges, and future directions of human-in-the-loop learning. IEEE Access, 12, 75735-75760.

Leslie, D. (2020) Understanding Artificial Intelligence Ethics and Safety. The Alan Turing Institute.

Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. Heliyon. 2024 Feb 15;10(4):e26297. doi: 10.1016/j.heliyon.2024.e26297. PMID: 38384518; PMCID: PMC10879008.

Morley, J., Floridi, L., Kinsey, L. and Elhalal, A. (2021) 'From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices', Science and Engineering Ethics, 27(1), pp. 1–31.

NHS Confederation (2021) AI in Healthcare: Opportunities and Challenges. Available at: https://www.nhsconfed.org/publications/ai-healthcare.

Oumaima, H. and Aziz, S. (2024) Digital health system and e-health in the following countries: United Kingdom, Norway, Sweden, Denmark, Germany and United States. Journal of Theoretical and Applied Information Technology, 102(1), pp.167-176.

Perez, K., Wisniewski, D., Ari, A., Lee, K., Lieneck, C. and Ramamonjiarivelo, Z. (2025). Investigation into application of AI and telemedicine in rural communities: a systematic literature review. In Healthcare (Vol. 13, No. 3, p. 324). MDPI.

Potts, H. W., Anderson, J. E., Colligan, L., Leach, P., Davis, S., & Berman, J. (2014). Assessing the validity of prospective hazard analysis methods: a comparison of two techniques. BMC health services research, 14(1), 41.

Red Hat (2023) Foundations of Security, Safety and Transparency in AI. Available at: https://www.redhat.com/en/resources/security-safety-transparency-ai-whitepaper.

Smith T. M. (2024) The future of AI and precision health: What stands in the way. Available at: https://www.ama-assn.org/practice-management/digital-health/future-ai-and-precision-health-what-stands-way.

Tetteh, S.G. et al (2025) Artificial Intelligence in Healthcare: A Systematic Review of Virtual Healthcare Assistants. Asian Journal of Probability and Statistics, 27(7), pp.43-62.

Topol, E. (2019) Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

Zuhair, V., Babar, A., Ali, R., Oduoye, M.O., Noor, Z., Chris, K., Okon, I.I. and Rehman, L.U. (2024) Exploring the impact of artificial intelligence on global health and enhancing healthcare in developing nations. Journal of primary care & community health, 15, p.21501319241245847.