npj | digital medicine

Perspective

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02073-4

A surgical approach to building impactful artificial intelligence

Simon C. Williams[1,2,5] ✉, Danyal Z. Khan[1,2,5], Sophia Bano[2], Ann Blandford[2,3], Matthew J. Clarkson[2], Enrico Costanza[3], Evangelos B. Mazomenos[2], Danail Stoyanov[2], Peter McCulloch[4] & Hani J. Marcus[1,2]

Two recent studies in surgical AI video analysis show divergent outcomes based on user expertise and AI explainability. While both demonstrate improved accuracy with AI-assistance, differences in trust and transparency critically shaped impact. This commentary explores human-computer interaction factors such as trust, usability, and explainability, and argues for structured evaluation frameworks to ensure effective, safe integration of AI into surgical practice.

Two recent studies describing AI-driven surgical video analysis revealed contrasting outcomes, highlighting the critical interplay between human factors and machine intelligence. Both studies—Khan et al. in *npj Digital Medicine*[1] and Williams et al. in *Annals of Surgery*[2]—explored how surgeons, across a spectrum of experience, performed when partnered with AI-assistance. In the first study, participants identified critical anatomy during the sella phase of pituitary surgery. In the second, clinicians determined whether a cerebral aneurysm was present in the operative microscope field. Whilst AI-assistance improved accuracy across the board in both studies, subgroup analysis revealed striking differences. Khan et al. found that novices benefitted the most, with their accuracy improving from 66% to 79%, while experts' gains were more modest (73% to 75%). Williams et al., however, flipped this dynamic—expert neurosurgeons saw a marked improvement in accuracy (77% to 92%), outpacing the novice's improvement (75% to 86%). Why did such a discrepancy in AI-assistance occur? Why, in one study, did experts trust, adopt, and integrate AI-assistance into their decision making, whilst in the other experts were largely unaffected by the AI's recommendation?

These contrasting patterns reveal an essential truth: the impact of AI in the operating room isn't just about accuracy; it's about how humans perceive, trust, and use AI. Simply deploying an accurate algorithm is insufficient if core human-computer interaction (HCI) factors—such as trust, explainability, usability, and perceived workload—are not deliberately addressed. This Commentary explores these themes, showcased through the narrative of Khan's and Williams' research. Finally, we touch on a roadmap for future research in this area—the path from AI-assistance to improved patient outcomes demands a structured approach. As proposed in frameworks such as IDEAL[3] and DECIDE-AI[4], researchers and developers must systematically calibrate human–AI "alignment" via iterative refinements and rigorous outcome measurements.

Explainability, trust, and expertise

In Khan's pituitary surgery study, participants faced a challenging task: navigating the intricate bony labyrinth seen in endoscopic endonasal pituitary surgery and outlining the sella. Bordered by complex loops of the internal carotid artery and the optic nerves, the sella represents the anatomical safe zone for entry in pituitary tumour surgery. After drawing the sella, participants were shown an outline of what the AI predicted to be the sella. Participants could adjust their decision or stick with their original outline. No one knew why the AI had made its choice. There was no information about its training data or reasoning, just a silent, unexplained recommendation. It is unsurprising, therefore, that those with the least baseline knowledge, medical students, placed the most trust in the AI. In *every case*, they changed their decision to more closely align with the algorithm, achieving an impressive 13% boost in accuracy.

Compare this to Williams et al.'s aneurysm study. Participants were once again asked to make predictions, in this case, whether an aneurysm was visible in a surgical frame. But this time, they were armed with more than just the AI's answer. Alongside predictions, participants received accuracy metrics for the model and heatmaps showcasing locations of the input image the AI was focusing on most to make its prediction, essentially, giving a glimpse of where the AI was 'looking'. The 'black box' had been partly opened.

For experts, who draw on years of experience and deeply ingrained heuristics, presenting additional information regarding AI model context, rationale and performance, was a game-changer. The heatmaps helped them validate their instincts on straightforward cases while nudging them to trust the AI in tougher scenarios, pushing their accuracy from 77% to 92%, notably higher accuracy than the AI platform alone. Novices, without such domain-specific intuition, depended more heavily on the AI regardless of explanation quality, and were seemingly more willing to trust it. It is unsurprising, therefore, that the novices gravitated to the AI-benchmark

[1]National Hospital for Neurology and Neurosurgery, London, UK. [2]The UCL Hawkes Institute, University College London, London, UK. [3]University College London Interaction Centre, University College London, London, UK. [4]Nuffield Department of Surgical Sciences, University of Oxford, John Radcliffe Hospital, Oxford, UK. [5]These authors contributed equally: Simon C. Williams, Danyal Z. Khan. ✉e-mail: rmapwil@ucl.ac.uk

(81% accuracy). Prior experience, familiarity with AI, and cognitive load all shape acceptance, with explanations most valuable when they reduce uncertainty without adding to mental burden[5,6]. In Williams et al., this balance of trust and autonomy translated into a 14% gain for experts.

It should be noted that both AI platforms had comparably high accuracy, the studies examined different tasks (anatomical segmentation vs. classification) within different neurosurgical subspecialties (neurovascular vs. pituitary skull base). These differences represent potential confounding factors, and any conclusions should therefore be interpreted with caution. Nevertheless, valuable lessons can still be drawn from these case studies.

AI is not a static tool; it is a complex intervention shaped by the environment in which it is placed and a range of human factors (Fig. 1). Trust and explainability are not optional, they're foundational, and must be at the heart of early stage surgical-AI design and evaluation, drawing upon validated scales (e.g. Hoffman Trust Scale) and principles of explainable AI (xAI) such as transparency and decision understanding[7].

These conclusions are echoed by the growing evidence base in this area, supported by several use cases. In radiology, interfaces that expose model rationale (e.g., heatmaps) increased clinicians' agreement with AI and shaped trust, underscoring the influence of explanation design on uptake[8].

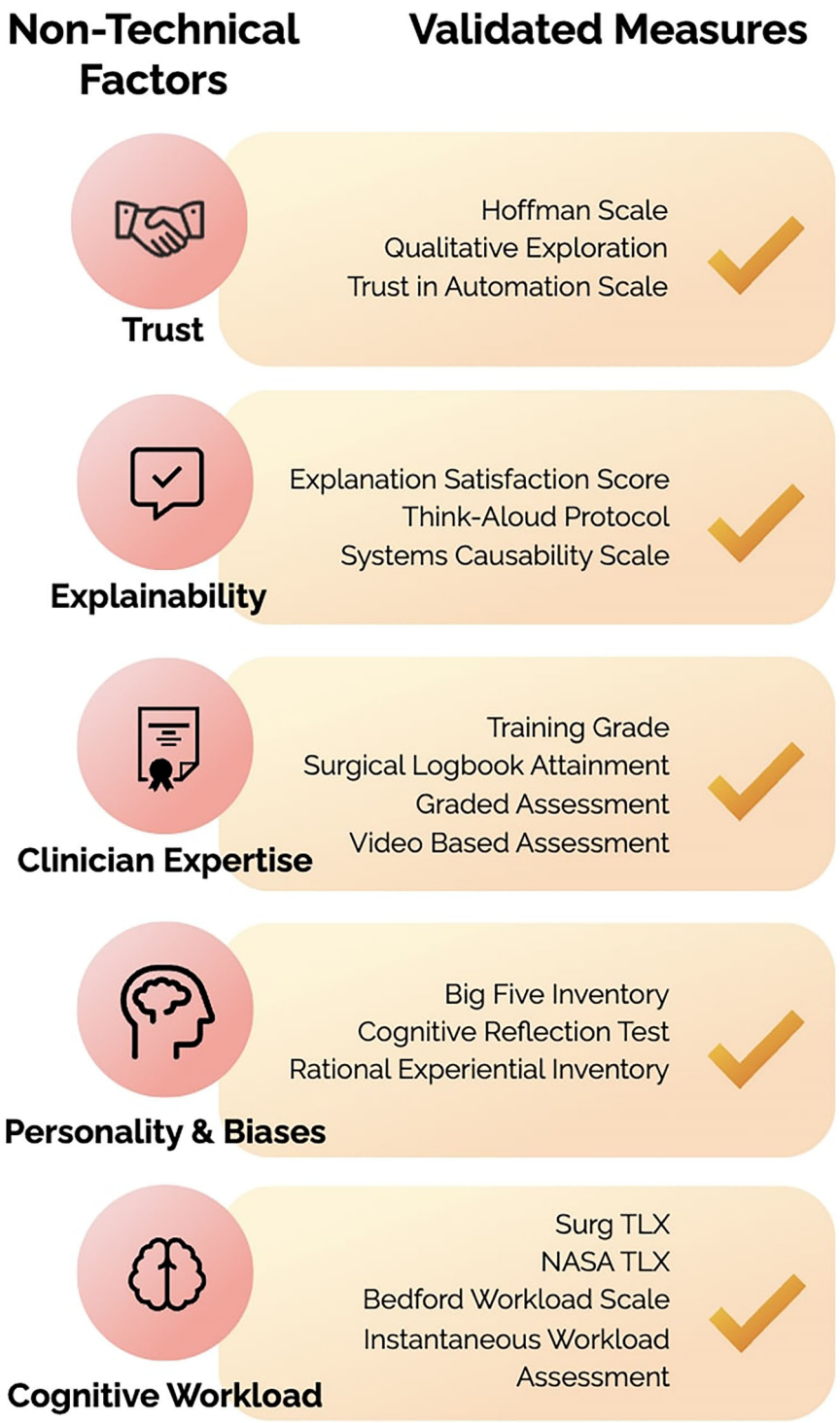

Fig. 1 | Non-technical performance factors influencing real-world performance: key non-technical factors influencing AI performance in healthcare, with validated measures used in evaluation. These metrics align with IDEAL and DECIDE-AI frameworks to ensure rigorous assessment in clinical implementation.

Outside of surgery, dermatology studies report that human-AI collaboration outperforms either alone and that class-activation map insights can guide better human decisions[9]; HCI toolkits for pathologists likewise increased diagnostic utility and trust without sacrificing accuracy[10]. Aman et al. go further, and argue that explainability in AI is a top priority, and carries ethical, legal, and clinical implications[11]. Frameworks, such as Markus et al.'s, have been developed to assist clinicians and engineers in selecting explainability tools[12].

Whilst the literature base demonstrating the value of improved HCI to clinicians is vast, there is a paucity of evidence demonstrating improved clinical outcomes secondary to improved explainability. Rezaeian et al. assessed how radiologists' trust, cognitive load, and accuracy were affected by AI outputs accompanied by saliency maps and confidence scores when diagnosing potential breast cancer. They found that high confidence in AI systems, even with explainable features built in, could lead to reduced performance, as clinicians sometimes over-relied on incorrect AI outputs when explanations appeared convincing[13]. Patterns of overreliance on AI leading to impaired performance have been reported in ophthalmology too[14]. These studies highlight the need for real-world outcomes assessment during an innovations life cycle. Frameworks for this include DECIDE-AI and IDEAL.

Lessons from DECIDE-AI and IDEAL

Building trust and fostering explainability in AI systems requires deliberate, structured efforts from both clinicians and engineers. Frameworks like IDEAL and DECIDE-AI offer a roadmap for systematically evaluating AI during its most critical stages of development—when innovations transition from the lab to the clinical environment.

Published in 2022, DECIDE-AI sought to address a gap in existing evaluative frameworks. Much attention has been given to pre-clinical (STARD-AI[15], TRIPOD-AI[16]) and comparative AI evaluation (CONSORT-AI[17], SPIRIT-AI[18]). However, frameworks for early-stage first-in-human evaluation of AI, arguably when the technology is at its most iterative, were absent. Following a two-stage expert Delphi process, the DECIDE-AI reporting standard was generated for this phase, with the aim of improving standards to aid reproducibility and scalability[4].

A key insight was the necessity of prioritising HCI factors from the outset, incorporating assessments of user trust, workload, and cognitive alignment into preclinical studies (IDEAL Stage 0). However, even the most thorough preclinical analyses cannot fully predict the complexities of human–AI interactions in real-world clinical settings. This underscores the importance of explainability and trust during early-phase evaluations, ensuring that systems are adaptable to user needs and are ready for safe and effective deployment. Moreover, trust in AI technology is not solely important in the pre-clinical stage, nor is it static. Its evolving nature should be systematically studied over time. Longitudinal studies could help map typical trust trajectories, which may rise, fall, or fluctuate in response to events, and should inform the design and interpretation of comparative studies such as randomised controlled trials. The importance of structured evaluation of HCI in AI-healthcare innovations is a sentiment echoed in numerous publications[11,19], but a framework solely dedicated to this remains absent.

Designing for trust and explainability

To bridge the trust gap and maximise AI's potential, future efforts should prioritise the following:

- **Measure human-computer interaction factors throughout the entire life cycle of an innovation:** Employ mixed methods approaches to measure key factors such as explainability and trust. Validated frameworks targeting discrete stages of the life cycle can be employed, such as DECIDE-AI[4], STARD-AI[15], TRIPOD-AI[16], CONSORT-AI[17], or SPIRIT-AI[18]. Other frameworks, such as IDEAL or Al-Ansari's xAI pillars[20], are designed to evaluate facets such as human factors throughout typical life-cycle progression. The type of approach used will be device-specific and study-specific and should be designed in conjunction with a human-computer interaction or human factors specialist. For example, early studies where there is rapid design iteration may benefit more from in-depth qualitative surveys and interviews, whilst later stage studies may benefit more from validated quantitative scales and behavioural measures (e.g. how frequently do participants accept correct or incorrect AI suggestions).
- **User-centred design feedback loop**: Incorporate human-computer interaction factors feedback from diverse users to ensure AI systems meet the needs of both users of various clinical and technological experience levels. In early studies this may result in version design changes, whereas later in the process this may result in implementation or process changes for a more stable device version. This approach aligns with the principles set out in the Chartered Institute of Ergonomics and Human Factors (CIEHF) White Paper on Human Factors in AI for Healthcare, which emphasises the critical role of user-centred design in ensuring safety, usability, and adoption.
- **Incorporating explainable AI principles where possible**: Provide clear metrics (e.g. input similarity or output confidence) and explanations (e.g. saliency maps) to promote well-calibrated human-AI alignment. Real time applications will need to carefully balance this additional information against cognitive workload (e.g. overload) and safety metrics (e.g. distraction and workflow disruption).
- **Clinical outcome incorporation**: Ultimately, when well-calibrated human-computer alignment is achieved in a pre-clinical setting, it must be further calibrated against real-world clinical and patient-reported outcomes.

Furthermore, designing for these human factors throughout an innovation's life cycle may indirectly benefit the regulatory process. Evaluation of human–computer interactions would provide evidence for formative and summative usability feedback. Consequently, a more granular understanding of human-computer interactions may emerge, leading to more effective risk analysis, improved risk control measures, and ultimately, safer medical devices.

Incorporating HCI into the evaluation of AI in healthcare is not without practical challenges. Time constraints of healthcare personnel, availability of expert HCI personnel, and the additional cost required to fund HCI evaluation are important considerations. Dedicated HCI teams may be incorporated into early-stage analysis to address these challenges[21,22].

Conclusion

Surgical AI's ultimate success hinges on more than accurate algorithms; it requires thoughtful integration of human-computer interaction factors, particularly explainability and trust. Khan's and Williams's studies underscore the need to tailor AI support to user expertise and ensure transparency in decision-making. By leveraging frameworks like IDEAL and DECIDE-AI, developers and clinicians can address these factors throughout the lifecycle of AI devices.

Data availability

No datasets were generated or analysed during the current study.

Received: 2 July 2025; Accepted: 8 October 2025;
Published online: 21 November 2025

References

1. Khan, D. Z. et al. Artificial intelligence assisted operative anatomy recognition in endoscopic pituitary surgery. *npj Digit. Med.* **7**, 1–7 (2024).
2. Williams, S. C. et al. Artificial intelligence assisted surgical scene recognition: a comparative study amongst healthcare professionals. *Ann. Surg.* https://doi.org/10.1097/SLA.0000000000006577 (2024).
3. Marcus, H. J. et al. IDEAL-D framework for device innovation: a consensus statement on the preclinical stage. *Ann. Surg.* https://doi.org/10.1097/SLA.0000000000004907 (2021).

4. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat. Med.* **27**, 186–187 (2021).
5. Hoffman, R.R. et al. Metrics for explainable AI: challenges and prospects. ArXiv abs/1812.04608 (2018).
6. Sujan, M. et al. Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care Inf.* **26**, e100081 (2019).
7. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **23**, 18 (2020).
8. Rainey, C. et al. Reporting radiographers' interaction with Artificial Intelligence—how do different forms of AI feedback impact trust and decision switching?. *PLOS Digit. Health* **3**, e0000560 (2024).
9. Tschandl, P. et al. Human-computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).
10. Cai, C et al. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY,USA, Paper 4, 1–14. https://doi.org/10.1145/3290605.3300234
11. Amann, J. et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **20**, 310 (2020).
12. Markus, A. F., Kors, J. A. & Rijnbeek, P. R. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inform.* **113**, 103655 (2021).
13. Rezaeian, O., Bayrak, A. E., & Asan, O. (2025). Explainability and AI Confidence in Clinical Decision Support Systems: Effects onTrust, Diagnostic Performance, and Cognitive Load in Breast Cancer Care. International Journal of Human–Computer Interaction, 1–21.https://doi.org/10.1080/10447318.2025.2539458
14. Carmichael, J. et al. Diagnostic decisions of specialist optometrists exposed to ambiguous deep-learning outputs. *Sci. Rep.* **14**, 6775 (2024).
15. Sounderajah, V. et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* **11**, e047709 (2021).
16. Collins, G. S. et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **11**, e048008 (2021).
17. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* **370**, m3164 (2020).
18. Cruz Rivera, S. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit. health* **2**, e549–e560 (2020).
19. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
20. Al-Ansari, N., Al-Thani, D. & Al-Mansoori, R. S. User-Centered Evaluation of Explainable Artificial Intelligence (XAI): a systematic literature review. *Hum. Behav. Emerg. Technol.* **2024**, 4628855 (2024).
21. Sujan, M., Pool, R. & Salmon, P. Eight human factors and ergonomics principles for healthcare artificial intelligence. *BMJ Health Care Inf.* **29**, e100516 (2022).
22. Chartered Institute of Ergonomics & Human Factors. *Human Factors and Ergonomics in Healthcare AI*. (2021).

Acknowledgements

We would like to thank Professor Peter McCulloch for his contributions on Trust in this work.

Author contributions

SCW—writing – original draft preparation; writing – review and editing. DZK—writing – original draft preparation; writing – review and editing. SB—writing – review and editing. AB—writing – review and editing. MJC – writing – review and editing. EC – writing – review and editing. EBM—writing – review and editing.DS – writing – review and editing. PMC—writing – review and editing. HJM—conceptualisation, writing—review and editing.

Competing interests

No specific funding was received for this piece of work. SCW is supported by the Amethyst Healthcare Group. HJM is supported by the NIHR Biomedical Research Centre at University College London. DZK is supported by the Cleveland Clinic London. DS is an employee of Digital Surgery, Medtronic. HJM is an employee of and has shares in Panda Surgical. DS has shares in Panda Surgical, Odin Vision, EnAcuity, and Helico Medical.

Additional information

Correspondence and requests for materials should be addressed to Simon C. Williams.