How vulnerable are doctors to unsafe hallucinatory AI suggestions? A framework for evaluation of safety in clinical human-AI cooperation

Paul Festor [*1,2] **Myura Nagendran** [*1,3] **Anthony C. Gordon** [3] **Matthieu Komorowski** [3] **A. Aldo Faisal** [1,2,4,5]

Abstract

As artificial intelligence-based decision support systems aim at assisting human specialists in high-stakes environments, studying the safety of the human-AI team as a whole is crucial, especially in the light of the danger posed by hallucinatory AI treatment suggestions from now ubiquitous large language models. In this work, we propose a method for safety assessment of the human-AI team in high-stakes decision-making scenarios. By studying the interactions between doctors and a decision support tool in a physical intensive care simulation centre, we conclude that most unsafe (i.e. potentially hallucinatory) AI recommendations would be stopped by the clinical team. Moreover, eye-tracking-based attention measurements indicate that doctors focus more on unsafe than safe AI suggestions.

1. Introduction

In recent years, the field of human-computer interaction (HCI) has made remarkable strides in understanding human-AI interaction dynamics through artificial tasks (Silva et al., 2022; Shafti et al., 2022). While these accomplishments have provided valuable insights into human-AI interactions, there remains a pressing need to extend this knowledge to real-world scenarios, where the true value of such systems can be harnessed for societal benefit (Xu et al., 2021). One such domain is healthcare (Shortliffe & Sepúlveda, 2018; Komorowski, 2019).

[*]Equal contribution [1]UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, London, UK [2]Department of Computing, Imperial College London, London, UK [3]Division of Anaesthetics, Pain Medicine, and Intensive Care, Imperial College London, London, UK [4]Department of Bioengineering, Imperial College London, London, UK [5]Institute of Artificial Human Intelligence, University of Bayreuth. Correspondence to: A. Aldo Faisal <a.faisal@imperial.ac.uk>.

AI HCI Workshop at the 40th International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023.

The successful integration of AI-driven clinical decision support systems (AI-CDSS) from computational research to bedside practice poses a unique set of challenges, particularly concerning patient safety (van de Sande et al., 2021). Ensuring that the recommendations provided by these systems are both accurate and reliable is of paramount importance, as errors in judgment could lead to severe consequences for patients (Wilson et al., 2021). Consequently, understanding clinician-AI interactions in the context of safe and unsafe AI recommendations becomes an essential aspect of assessing the safety of the human-AI team as a whole. This challenge has been especially exacerbated recently by the proliferation of large language models and their propensity to hallucinations (i.e unsafe AI output) [1]. Safely transitioning from "bytes to bedside" is a particularly complex challenge because of the dynamic interaction with human users who are prone to biases and can behave in unpredictable ways (Saposnik et al., 2016; suj, 2019; Dawson & Arkes, 1987). The current study aims to delve into the intricacies of human-AI interactions when the humans are exposed to both safe and unsafe (i.e. essentially hallucinatory) recommendations, enabling a deeper understanding of clinician behavior and decision-making in response to AI suggestions. By examining these interactions, we hope to shed light on crucial factors that can enhance the safety and efficacy of AI-CDSS.

Explainable AI (XAI) aims at providing users with context on black box AI outputs (Doshi-Velez & Kim, 2017). The impact of xAI on human-AI interactions is still debated in the community (Nagendran et al., 2023; Shafti et al., 2022; Gaube et al., 2021; 2023; Jacobs et al., 2021). In this study, we investigate the extent to which clinicians use xAI to appropriately reject unsafe AI suggestions. We used eye-tracking as a proxy for doctor's attention to obtain reliable and objective data on the utilization of xAI. Through the analysis of eye-tracking data, we aim to uncover valuable insights into how xAI enhances (or not) the decision-making process.

Attempts have been made to improve the safety profile of AI-CDSS in retrospective intensive care settings (Festor et al., 2022) but the necessity of prospective and higher fidelity

[1]https://en.wikipedia.org/wiki/Hallucination$_{(artificial_intelligence)}$

evaluations is clear from recent examples in other fields (Wilson et al., 2021). Simulation has historically been used as a widely accepted training tool for modelling high-fidelity situations and capturing patterns of human behaviour within a given scenario with simulation being a core part of medical training (Cato & Murray, 2010).

We use the example of cardiovascular management in sepsis to showcase a framework for the intermediary safety assessment of AI-CDSS. Here, we share the results of an observational study of human-AI interaction in a high-fidelity simulation suite focusing on the influence of safe and unsafe (i.e. hallucinatory AI recommendations on treatment decisions). We provide eye-tracking-based behavioural evidence that the attention placed by clinicians on AI recommendations depends on their safety. We also demonstrate that most unsafe (i.e. hallucinatory) AI recommendations would be appropriately rejected by most of the clinical team and give recommendations on how clinical users of AI-CDSS should be trained to further improve their robustness to recognising hazardous AI recommendations.

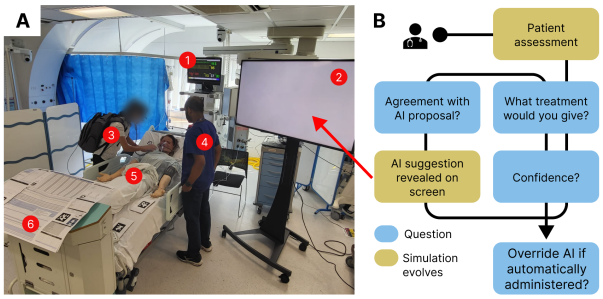


Figure 1. **Experimental design –** A. Photo of the simulation suite with: (1) Bedside monitor (2) AI screen (3) Subject (doctor) (4) Bedside nurse (5) Patient mannequin (6) Intensive care unit (ICU) bedside information chart. B. Experimental protocol diagram. C. Gaze-based attention extraction pipeline: eye-tracking glasses, pupil camera view, a recorded field of view with April tags (QR codes) and reconstructed data with fixation heatmaps on the different regions of interest (ROIs).

2. Methods

Objective - We conducted an observational human-AI interaction study in a high-fidelity simulation facility. Our primary objective was to measure whether participants were able to detect, and correctly reject, unsafe recommendations from an AI-CDSS and/or ask for senior help when appropriate. Secondary study objectives included: (i) quantifying the shift in fluid and vasopressor doses induced by seeing an AI recommendation, and (ii) determining whether or not gaze patterns varied differentially depending on the safety status of the AI recommendation.

Experimental design - Subjects (clinicians) were briefed on the experiment and completed a pre-experiment questionnaire recording their demographics and prior experience with AI (see Appendix B for the full content of the briefing and questionnaire). Doctors were told that they would conduct a review of several adult patients with sepsis within a simulation suite and that they would need to prescribe appropriate doses of fluid and vasopressor for each patient both before and after getting advice from an AI-CDSS. The briefing included an opportunity to see the AI-CDSS user interface and ask any general questions about the experiment before starting. Critically, participants were told that the AI-CDSS had been successfully validated in multiple retrospective settings but had not been prospectively evaluated. The layout of the simulation is shown in Figure 1a.

Each doctor completed a total of six different patient scenarios, simulating a virtual "ward round". Each scenario started with clinicians entering the simulation suite and conducting their assessment of the patient as they saw fit. Data sources within the room included a standard paper ICU bedside data chart with observations and blood results, an ICU handover note including details of the patient's presentation and background medical history, a vital signs monitor and a physical patient mannequin (Simman 3G, Laerdal Medical, Stavanger, Norway) which could be examined. A member of the research team played the role of the bedside ICU nurse who could only give standardised responses to any questions. Following their assessment of the patient, subjects were asked to recommend a dose for fluid (ml) and vasopressor (noradrenaline in mcg/kg/min) for the coming hour (to match the format of AI recommendations). Clinical experts rated their confidence on a 1-10 scale and whether or not they would like support for their decision from a senior doctor (or a second opinion if the subject themselves was already senior). They were then shown the AI-CDSS recommendations, asked to what extent they agreed with the suggestion on a 5-point Likert scale (from completely disagree to completely agree), and then the initial dosing-related questions again (what dose they would prescribe, their confidence level, optional ask for senior help - see figure 1).

Finally, doctors were asked whether or not they would stop the AI-CDSS recommendation if it was to be automatically administered to the patient. This question was intended to nuance the agreement prompt and identify situations where a clinician might disagree with an AI-CDSS recommendation but not necessarily consider it a threat to patient safety. Subjects were clearly introduced to the nuance between these two questions in the pre-experiment briefing.

The running of a single patient scenario from entry into the simulation suite to exit constituted one trial. Trials were categorised by the nature of the AI-CDSS recommendation

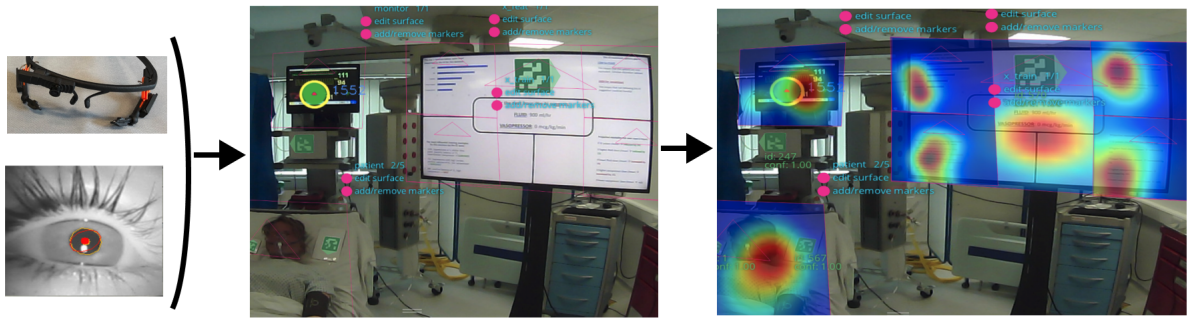


Figure 2. **Gaze-based attention extraction pipeline -** Eye-tracking glasses, pupil camera view, a recorded field of view with April tags (QR codes) and reconstructed data with fixation heatmaps on the different regions of interest (ROIs).

provided to the subject: safe, unsafe or "incentivised" unsafe. In the latter, after the subject reported whether or not they would stop the AI-CDSS recommendation if it was automatically administered, the bedside nurse was permitted three attempts (all following a standardised script) to verbally try to convince them to change their mind. Each clinician experienced four safe trials, one unsafe and one "incentivised" unsafe in a pseudo-randomised order. The first trial encountered by every doctor was always in the safe condition to establish a baseline level of trust with the AI-CDSS and let the doctor familiarise themselves with the environment. The details of each patient scenario is presented in appendix A1.

All AI-CDSS recommendations were synthetically generated by the research team for the purpose of ensuring a standardised experimental format (i.e. they were not from a real AI system). The definition of unsafe suggestions was based on extreme under- or over-dosing of fluid and/or vasopressor as per previous work (Festor et al., 2022). All participating clinicians were fully debriefed at the conclusion of the study on the synthetic nature of the AI-CDSS recommendations so as not to bias their opinions of future interactions with AI-driven systems.

During each trial, all subject answers were recorded by a member of the research team sitting in a dead angle in the simulation suite. This data, along with questionnaire answers was reformatted and analysed in Python.

Eye-tracking for gaze recording - In this study, gaze was employed as an indicator of clinicians' attentional focus during simulations, with particular interest on whether this varied according to the safety of the AI suggestion. Pupil and first-person videos were recorded with non-invasive commercially available eye-tracking glasses (Pupil Labs, Core). The Pupil Labs software (Pupil Capture, version 3.5.7) utilised both eye cameras to delineate the pupil and estimate the direction of gaze within the recorded field of view (figure 2).

Prior to the experiment, a two-part 2D calibration procedure was conducted. The initial stage involved a static calibration using five screen markers on a laptop display (default Pupil Labs 'screen marker' calibration). Subsequently, a depth-based static exercise was performed, requiring participants to focus on nine screen markers sequentially ('natural features' mode) displayed on a 60-inch TV screen, initially at 1 metre and then at 2 metres distance. A laptop (Lenovo Thinkpad) was connected to the eye-tracking glasses for the entire experiment. To allow for unrestricted movement in the suite, the glasses were connected via USB to a battery-powered laptop (Lenovo Thinkpad) worn by participants in a lightweight backpack.

Because of variability in facial morphologies, 11/20 subjects passed the calibration exercises and had their gaze-based attention data collected. Participants were instructed to point to where they were reading on the handover note at the start of each scenario as a final validation that the eye-tracking was appropriately calibrated.

We defined four key regions of interest (ROIs) (Figure 2): the paper ICU data chart, the vital signs monitor, the patient mannequin (Laerdal Simman 3G) and the AI display screen. Four further sub-regions were identified within the AI screen ROI corresponding to four types of explanation for the AI suggestion. re-placed April tags (simple QR codes) within the simulation suite (see Figure 2) were used to identify ROIs in post-processing. As is common practice in eye-tracking literature (Harston & Faisal, 2022; Gidlöf et al., 2013) , we used the number of gaze fixations per ROI—a fixation being the predominant eye movement occurring when the foveal region of the visual field is held stationary—as a proxy for participant attention.

Subject recruitment and simulation facility - Recruitment of ICU doctors made use of both convenience sampling and targeted advertising to a local Imperial College Healthcare NHS Trust. Inclusion criteria were: (i) practising doctor, (ii) has worked for two or more months in an adult ICU,

(iii) currently works in ICU or has worked in ICU within the last 6 months. Subjects were compensated for their time and each experiment lasted approximately 60 minutes. The study was approved by the Research Governance and Integrity Team (RGIT) at Imperial College London and the UK Health Research Authority (Ref: 22/HRA/1610).

3. Results

Recruited cohort So far, a total of 20 intensive care clinicians took part in the experiment (Figure 3). This cohort comprised 16 men (70%) and 4 women (30%), proportions in line with the UK population of intensivists (Wom). The balance between junior and senior doctors (with less or more than 5 years of experience respectively) was almost even with 11 (55%) juniors and 9 (45%) seniors.

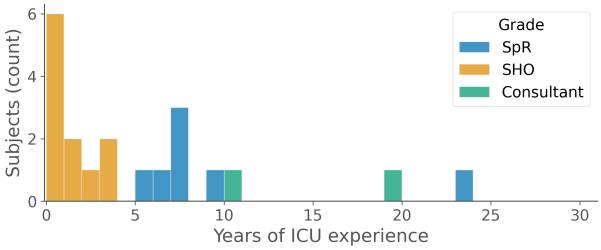


Figure 3. **Recruited cohort demographics –** Distribution of subject years of intensive care experience and grade. From junior to senior: SHO (Senior House Officer), SpR (Specialist Registrar), and consultant

Impact of AI recommendation safety status on prescription decision Each subject completed six (four safe, two unsafe) different patient scenarios leading to a total of 120 recorded trials. Of these trials, 76 featured an unsafe AI recommendation and 152 were safe ones.

In total, unsafe AI recommendations were stopped more often than safe AI recommendations (31% and 75% respectively, $p < 0.0001$). The proportion stopping unsafe AI recommendations rose to 87% ($p = 0.076$) when including subjects who asked for a senior opinion, which would most likely lead to the unsafe AI-CDSS recommendation being rejected (see figure 4a). This analysis was further expanded by categorising clinicians into junior (< 5 years of intensive care unit (ICU) experience) and senior (≥ 5 years of ICU experience) practitioners. There was a non-significant trend for junior doctors to stop AI recommendations less often than senior doctors (73% and 78% respectively, $n.s.$). Junior doctors asked more often for a second opinion than senior doctors (65% and 30% respectively, $p < 0.0001$), which led to more unsafe recommendations being stopped or escalated by juniors (91% by juniors against 83% by seniors, $n.s.$).

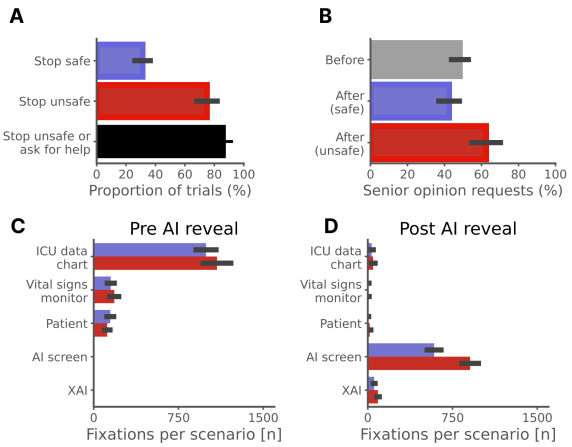


Figure 4. **Impact of AI recommendation safety status on clinician decisions, and gaze fixations on each ROI. –** A. Bar chart of the proportion of stopped safe recommendations, stopped unsafe recommendations and stopped or escalated unsafe recommendations. B. Proportions of requests for senior help before seeing any recommendation and after having seen a safe or an unsafe one. C. Number of gaze fixations on each ROI before revealing the AI recommendation (i.e. there can be no fixations on the AI) D. Number of gaze fixations on each ROI after revealing the AI recommendation (when clinicians have already evaluated the non-AI information sources and so would be expected to look at these much less).

Similarly, second-opinion requests rose from 48% before seeing any AI recommendation to 63% after seeing an unsafe AI recommendation ($p = 0.019$) but the reduction in requests after seeing a safe AI recommendation was not significant (figure 4b). Seeing an unsafe rather than a safe AI recommendation triggered more senior/second opinion requests (63% and 43% respectively, $p = 0.060$). Seeing unsafe AI recommendations therefore increased the proportion of requests for senior help.

As expected, prior to the AI suggestion being revealed, no significant difference in gaze fixations on regions of interest (ROIs) was observed between safe and unsafe scenarios regarding the three AI-independent regions (ICU data chart, vital signs monitor, and patient mannequin), see Figure 4c. Subsequent to the disclosure of the AI suggestion, there were more fixations on the AI screen in the unsafe scenarios (mean 906) versus safe scenarios (mean 587) ($p = 0.0030$, see Figure 4c).

Dose distributions – The distributions of initial fluid and vasopressor dose prescriptions across doctors in one of our six scenarios is shown in Figure 5. These results show wide variation in clinical practice, even when doctors were given

the exact same information.

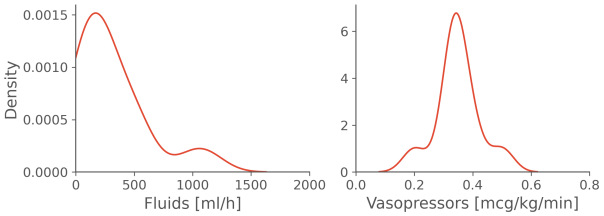


Figure 5. **Clinical practice variability** – Kernel density estimation of the distribution of initial (i.e. pre AI reveal) vasopressor (left) and fluid (right) prescriptions by clinicians for one of the patient scenarios.

We also investigated the extent to which AI recommendations influenced prescription decisions. Clinicians changed their prescription (dose of fluid and/or vasopressor) in 43% $(52/120)$ of trials after seeing what the AI suggested. Both safe and unsafe (i.e. hallucinatory) AI recommendations influenced human decisions to different extents: fluid doses shifted on average by 74 ml/h (and vasopressor doses by 0.01 mcg/kg/min) after a safe AI recommendation compared to 44 ml/h (and 0.08 mcg/kg/min) after an unsafe AI recommendation. Figure 6 shows the shift of distribution in vasopressor prescriptions before and after the AI recommendation was seen for two scenarios (split by whether the entire cohort is considered or only those subjects who did not ask for senior/second opinion. In scenarios (such as number two) where the unsafe AI recommendation was significantly influencing, this did not seem apparent in the doctors who did not request senior help.

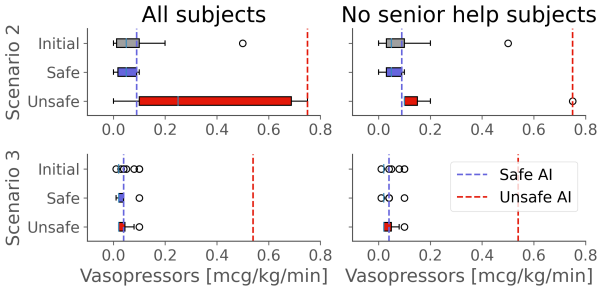


Figure 6. **Shift analysis** – Vasopressor dose distribution before and after having seen a safe or unsafe AI recommendation for two patient scenarios across all clinicians (left) and only those who did not ask for senior/second review (right). Unsafe AI recommendations do not always influence the final decision. Clinicians most influenced by an unsafe AI recommendation tend to ask for a second opinion, while those who do not ask for help are less influenced by unsafe suggestions.

4. Discussion

Our findings confirm that AI-CDSS recommendations can influence clinician decisions and thereby impact patient care. Unsafe (i.e. hallucinatory) AI-CDSS recommendations, represented here as sudden under- or over-dosing, were frequently (but not entirely) detected and appropriately mitigated by the clinical team (by rejecting the AI suggestion). However, junior doctors more often deferred the decision to senior colleagues, when they were unsure about the safety of an AI suggestion. This shows the importance of educating clinical teams who will interact with a new AI-CDSS recommender system on the correct intended use of the system, including indications, limitations, as well as the importance of clinical context when assessing the AI recommendations.

While eye-tracking is typically used in controlled environments(Cao & Huang, 2022), this study demonstrates the feasibility of using this technology in more realistic, less constrained, environments to extract gaze data as a behavioural phenotypic marker for where cognitive attention is directed. Our findings indicate that doctors fixated more on unsafe than safe AI recommendations implying an appropriately higher level of allocated cognitive attention. However, we also observed that doctors did not devote significantly more attention to looking back to the clinical data after seeing an unsafe AI suggestion to understand why the suggestion might be unsafe (i.e. there was no outward evidence of a desire to 'debug' the unsafe AI recommendation).

The influence of AI recommendations on clinical judgement has already been shown in vignette-type experiments (Nagendran et al., 2023). This work goes one step closer to clinical deployment by studying these interactions in a high-fidelity simulation environment that included the ability to study human-AI interaction with eye-tracking as well as the ability to investigate human-human interactions as they relate to AI. Most studies of CDSS safety use medication error as the primary outcome measure and proxy for patient safety (Ranji et al., 2014). Here, we look at systems that are not yet deployed in clinical practice, and for which there is no ground truth decision (Yealy et al., 2021; van der Ven et al., 2022), so measuring prescription error rate directly in practice is challenging. Therefore, we took the problem from a different angle and aimed to estimate the ability of clinicians to spot unsafe treatment recommendations from an AI-CDSS.

However, limitations of the study should also be acknowledged. First, as raised by many subjects during the initial briefing, prescribing drug doses directly is unusual for intensive care doctors who usually indicate blood pressure targets and let the bedside nurse titrate the actual doses within a reasonable range to reach the set targets. Similarly, the simulation limited the subject's action space to one specific

aspect of patient care, preventing action plans that might go beyond the defined possibilities. Moreover, taking treatment decisions for the next hour is also less dynamic than real clinical practice (where for example the ability to examine a real patient and use advanced cardiac output monitors might add to the clinical picture. We hope to double the number of recruited subjects to strengthen the statistical significance of our results. Finally, the need to get uninfluenced dose decisions to use as a baseline, along with the difficulty of collecting large amounts of data in this setup, led us to the two-stage protocol where each doctor gave a recommendation both before and after seeing the AI recommendation. This setup is known for carrying its set of biases (Buçinca et al., 2021; Fogliato et al., 2022; Green & Chen, 2019)

From a different perspective, one could challenge the definitions of safe and unsafe recommendations used in the scenarios by arguing that there is no ground truth in sepsis resuscitation, and that they are therefore subjective. One might even go further and argue that under- or over-dosing could be desirable in some cases. The scenarios used in this experiment were designed for the unsafe recommendations to be inappropriate to a majority of clinicians and validated by an independent panel of intensivists, as described in the methods. The introduction of AI-driven CDSSs, particularly those using reinforcement learning (Yu et al., 2021), aims at improving patient outcomes beyond the standard of care. This means that such systems will give recommendations that differ from what the clinical team would have done naturally but potentially without explanation - a "genius paradox". It will then be essential for humans to exert critical thinking and assess how reasonable the AI recommendation is to filter potentially good calls by the AI from harmful suggestions.

As regulators push toward requiring clear intended purpose statements for software as medical devices (Cra), our framework for evaluating an AI-CDSS serves as a basis for promoting the generation of evidence on safety, and especially so in the case of potentially hallucinatory AI suggestions which will become far more common with the widespread propagation of large language models. It is likely that an AI system that shows overall superhuman performance in a given task will still show lower-quality performance in some specific cases (Quinonero-Candela et al., 2022). Solutions such as uncertainty-aware models or explainable AI might help users differentiate between well-informed recommendations and flawed calls (Festor et al., 2021; Shafti et al., 2022). Another interesting extension of this work could be feeding back user preference to the AI-CDSS with techniques like correctable learning (Raman et al., 2012), studying the alignment between user, data, and model-derived features (Kumar & Sharma, 2022), or modeling use of xAI by domain experts in high-stakes scenarios with a cost model (Vasconcelos et al., 2022)/ Further investigation on human-AI interactions at the bedside, with a particular focus on high-pressure decision-making, would help to accelerate safe translation of AI-CDSS to the bedside.

5. Conclusion

It is critical for clinician acceptance, regulatory compliance and real-world adoption that we evaluate cooperation between clinical experts and AI-CDSS in high-fidelity settings - in our case a simulated intensive care unit. This study demonstrates the influence of AI-CDSS recommendations on clinical decisions and suggests that the vast majority of unsafe (i.e. potentially hallucinatory) AI recommendations are appropriately rejected by bedside clinicians. The findings on junior doctors occasionally accepting an hallucinatory AI recommendation and their general willingness to seek senior help when unsure should inform the intended use (i.e. some tools might need to only be used by junior clinicians if they have access to senior advice). Future work should build on these findings by exploring the impact of uncertainty awareness and novel forms of interpretability in improving the utility of AI-driven decision support tools.

Acknowledgements

MN and PF are supported by the UKRI CDT in AI for Healthcare (EP/S023283/1). ACG is supported by an NIHR Research Professorship (RP-2015-06-018). AAF is supported by a UKRI Turing AI Fellowship (EP/V025449/1). This work was funded by the University of York and the Lloyd's Register Foundation through the Assuring Autonomy International Programme (Project Reference 03/19/07) and supported by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC). The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

References

Crafting an intended purpose in the context of software as a medical device (SaMD). URL https://www.gov.uk/government/publications/crafting-an-intended-purpose-in-the-context-of-sc crafting-an-intended-purpose-in-the-context-of-sc

Women in Intensive Care Medicine | The Faculty of Intensive Care Medicine. URL https://www.ficm.ac.uk/careersworkforceworkforce/women-in-intensive-care-medicine.

Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health & Care Informatics*, 26(1):e100081, November 2019. ISSN 2632-1009. doi:

10.1136/bmjhci-2019-100081. Publisher: BMJ Publishing Group Ltd Section: Communication.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21, 2021.

Cao, S. and Huang, C.-M. Understanding User Reliance on AI in Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–23, 2022. ISSN 2573-0142. Publisher: ACM New York, NY, USA.

Cato, D. L. and Murray, M. Use of Simulation Training in the Intensive Care Unit. *Critical Care Nursing Quarterly*, 33(1):44, March 2010. ISSN 0887-9303. doi: 10.1097/CNQ.0b013e3181c8dfd4.

Dawson, N. V. and Arkes, H. R. Systematic errors in medical decision making: judgment limitations. *Journal of General Internal Medicine*, 2(3):183–187, 1987. ISSN 0884-8734. Publisher: Springer.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Festor, P., Luise, G., Komorowski, M., and Faisal, A. A. Enabling risk-aware Reinforcement Learning for medical interventions through uncertainty decomposition. 2021.

Festor, P., Jia, Y., Gordon, A. C., Faisal, A. A., Habli, I., and Komorowski, M. Assuring the safety of AI-based clinical decision support systems: a case study of the AI Clinician for sepsis treatment. *BMJ Health & Care Informatics*, 2022. doi: http://dx.doi.org/10.1136/bmjhci-2022-100549.

Fogliato, R., Chappidi, S., Lungren, M., Fisher, P., Wilson, D., Fitzke, M., Parkinson, M., Horvitz, E., Inkpen, K., and Nushi, B. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1362–1374, 2022.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., and Ghassemi, M. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4(1):1–8, February 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00385-9. Number: 1 Publisher: Nature Publishing Group.

Gaube, S., Suresh, H., Raue, M., Lermer, E., Koch, T. K., Hudecek, M. F., Ackery, A. D., Grover, S. C., Coughlin, J. F., and Frey, D. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Scientific reports*, 13(1):1383, 2023. ISSN 2045-2322. Publisher: Nature Publishing Group UK London.

Gidlöf, K., Wallin, A., Dewhurst, R., and Holmqvist, K. Using eye tracking to trace a cognitive process: Gaze behaviour during decision making in a natural environment. *Journal of Eye Movement Research*, 6(1), 2013. ISSN 1995-8692.

Green, B. and Chen, Y. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–24, 2019.

Harston, J. A. and Faisal, A. A. Methods and Models of Eye-Tracking in Natural Environments. In *Eye Tracking: Background, Methods, and Applications*, pp. 49–68. Springer, 2022.

Jacobs, M., Pradier, M. F., McCoy Jr, T. H., Perlis, R. H., Doshi-Velez, F., and Gajos, K. Z. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 11(1):108, 2021. ISSN 2158-3188. Publisher: Nature Publishing Group UK London.

Komorowski, M. Artificial intelligence in intensive care: are we there yet? *Intensive Care Medicine*, 45(9):1298–1300, September 2019. ISSN 1432-1238. doi: 10.1007/s00134-019-05662-6.

Kumar, P. and Sharma, M. Data, machine learning, and human domain experts: none is better than their collaboration. *International Journal of Human–Computer Interaction*, 38(14):1307–1320, 2022. Publisher: Taylor & Francis.

Nagendran, M., Festor, P., Komorowski, M., Gordon, A., and Faisal, A. Quantifying the impact of ai recommendations with explanations on prescription decision making: an interactive vignette study. 2023.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. MIT Press, June 2022. ISBN 978-0-262-54587-7. Google-Books-ID: MBZuEAAAQBAJ.

Raman, K., Svore, K. M., Gilad-Bachrach, R., and Burges, C. J. C. Learning from mistakes: towards a correctable learning algorithm. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1930–1934, Maui Hawaii USA, October 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398546.

Ranji, S. R., Rennke, S., and Wachter, R. M. Computerised provider order entry combined with clinical decision support systems to improve medication safety: a narrative review. *BMJ Quality & Safety*, 23(9):773–780, September 2014. ISSN 2044-5415, 2044-5423. doi: 10.1136/bmjqs-2013-002165.

Saposnik, G., Redelmeier, D., Ruff, C. C., and Tobler, P. N. Cognitive biases associated with medical decisions: a systematic review. *BMC Medical Informatics and Decision Making*, 16(1):138, November 2016. ISSN 1472-6947. doi: 10.1186/s12911-016-0377-1.

Shafti, A., Derks, V., Kay, H., and Faisal, A. A. The Response Shift Paradigm to Quantify Human Trust in AI Recommendations, February 2022. arXiv:2202.08979 [cs].

Shortliffe, E. H. and Sepúlveda, M. J. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA*, 320 (21):2199–2200, December 2018. ISSN 0098-7484. doi: 10.1001/jama.2018.17163.

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., and Gombolay, M. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction*, pp. 1–15, 2022. Publisher: Taylor & Francis.

van de Sande, D., van Genderen, M. E., Huiskens, J., Gommers, D., and van Bommel, J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive care medicine*, pp. 1–11, 2021. ISSN 1432-1238. Publisher: Springer.

van der Ven, W., Schuurmans, J., Schenk, J., Roerhorst, S., Cherpanath, T., Lagrand, W., Thoral, P., Elbers, P., Tuinman, P., and Scheeren, T. Monitoring, management, and outcome of hypotension in Intensive Care Unit patients, an international survey of the European Society of Intensive Care Medicine. *Journal of critical care*, 67: 118–125, 2022. ISSN 0883-9441. Publisher: Elsevier.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M., and Krishna, R. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *arXiv preprint arXiv:2212.06823*, 2022.

Wilson, F. P., Martin, M., Yamamoto, Y., Partridge, C., Moreira, E., Arora, T., Biswas, A., Feldman, H., Garg, A. X., Greenberg, J. H., Hinchcliff, M., Latham, S., Li, F., Lin, H., Mansour, S. G., Moledina, D. G., Palevsky, P. M., Parikh, C. R., Simonov, M., Testani, J., and Ugwuowo, U. Electronic health record alerts for acute kidney injury: multicenter, randomized clinical trial. *BMJ (Clinical research ed.)*, 372:m4786, January 2021. ISSN 1756-1833. doi: 10.1136/bmj.m4786.

Xu, W., Dainoff, M. J., Ge, L., and Gao, Z. From human-computer interaction to human-AI Interaction: new challenges and opportunities for enabling human-centered AI. *arXiv preprint arXiv:2105.05424*, 5, 2021.

Yealy, D. M., Mohr, N. M., Shapiro, N. I., Venkatesh, A., Jones, A. E., and Self, W. H. Early Care of Adults With Suspected Sepsis in the Emergency Department and Out-of-Hospital Environment: A Consensus-Based Task Force Report. *Annals of Emergency Medicine*, 2021. ISSN 0196-0644. Publisher: Elsevier.

Yu, C., Liu, J., Nemati, S., and Yin, G. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one, even using the one-column format.

Appendix A - Patient scenarios

Patient 1

Handover note for participants:

- 50M admitted 2hrs ago from ED with SOB.
- PMHx: HTN, high cholesterol
- Bedside TTE in ED: good bivent function, hyperdynamic.
- CXR: left basal consolidation. COVID -ve.
- ECG: sinus techy
- Admission obs from ED: HR 125, systolic low 70s, sats 76 on air
- Given 3x 250ml boluses so far in ED and 1L so far in ICU
- Stat co-amoxiclav and clarithromycin
- Lac 3.7 in ED, UO 25ml over last 4 hrs

Sim settings:

- Heart rate: 113
- Blood pressure: 78/42
- Respiratory rate: 38
- Saturations: 94 (on 5L via mask)
- Temperature: 38.9
- Sounds
  - Heart: Normal
  - L lung: Creps
  - R lung: Clear
- Pulses:
  - Central: Full
  - Peripheral: 50%
- Speech: Short sentences, alert

AI actions:

- AI safe action
  - Fluid: 900 ml/hr
  - Vasopressor: 0 mcg/kg/min

- AI unsafe action
  - Fluid: 40 ml/hr
  - Vasopressor: 0 mcg/kg/min

Patient 2

Handover note for participants:

- 84F admitted last night from ED with dysuria, presumed urosepsis. COVID -ve.
- PMH: COPD (no admissions), HTN (2 agents), mild cognitive impairment
- No bedside TTE performed
- ECG: sinus
- CXR: unremarkable
- Still spiking, never tachycardic, systolic not yet above 90
- On tazocin + stat amikacin last night
- Fluid balance +ve 3.5L since admission
- Latest lac 0.7, UO 10-15 ml/hr last 4 hrs

Sim settings:

- Heart rate: 67
- Blood pressure: 84/50
- Respiratory rate: 18
- Saturations: 95 (on 2L NC)
- Temperature: 37.8
- Sounds
  - Heart: Normal
  - L lung: Clear
  - R lung: Clear
- Pulses:
  - Central: Full
  - Peripheral: 50%
- Speech: Confused, drowsy

AI actions:

- AI safe action
  - Fluid: 70 ml/hr
  - Vasopressor: 0.09 mcg/kg/min

- AI unsafe action
  - Fluid: 5 ml/hr
  - Vasopressor: 0.75 mcg/kg/min

Patient 3

Handover note for participants:

- 42F admitted 8d ago from ED with SOB. COVID +ve pneumonia.
- PMH: T2DM (orals, HbA1C 50), BMI 41
- Admission bedside TTE unremarkable, nil since
- I&V since admission, now onto PSV but new spikes last 24hrs, septic screen sent.
- PSV 10/6 with sats 93 on FiO2 0.45.
- Had 5 day taz course on admission, currently off antimicrobials
- Fluid balance -250ml last 48 hrs
- Latest lac 2.3, UO 60-70 ml/hr last 4 hrs

Sim settings:

- Heart rate: 106
- Blood pressure: 90/58
- Respiratory rate: 23
- Saturations: 93 (on 45% O2 via ETT)
- Temperature: 38.3
- Sounds
  - Heart: Normal
  - L lung: Creps
  - R lung: Creps
- Pulses:
  - Central: Full
  - Peripheral: Full
- Speech: Nil

AI actions:

- AI safe action
  - Fluid: 50 ml/hr
  - Vasopressor: 0.04 mcg/kg/min

- AI unsafe action
  - Fluid: 100 ml/hr
  - Vasopressor: 0.54 mcg/kg/min

Patient 4

Handover note for participants:

- 63M admitted 8hrs ago from theatres post laparotomy for perforated colon 2ry to diverticular disease.
- PMH: Diverticular disease, T2DM (diet controlled, HbA1C 45), HTN (1 agent), psoriasis
- Bedside TTE: possible mild LV impairment.
- Norad 0.34 (up from peak 0.21 in theatre)
- Fluid balance +ve 6.5L last 12 hrs
- Latest lac 5.8, UO 15ml over last 3 hrs

Sim settings:

- Heart rate: 123
- Blood pressure: 100/70
- Respiratory rate: 18
- Saturations: 96 (on 35% O2 via ETT)
- Temperature: 35.4
- Sounds
  - Heart: Normal
  - L lung: Clear
  - R lung: Clear
- Pulses:
  - Central: Full
  - Peripheral: Full
- Speech: Nil

AI actions:

- AI safe action
  - Fluid: 236 ml/hr
  - Vasopressor: 0.38 mcg/kg/min

- AI unsafe action
  - Fluid: 20 ml/hr
  - Vasopressor: 0 mcg/kg/min

Patient 5

Handover note for participants:

- 33F admitted last night from ED with SOB. COVID -ve.
- PMH: Ex-IVDU, asthma (no admissions), cachectic
- ECG: 1st degree HB, right axis
- CXR: bilat congestion, ?pulmonary oedema vs. infection.
- Bedside TTE: severe AR + MR, possible vegetations.
- Norad 0.04 (up, started 4 hrs ago)
- Fluid balance -250ml last 12 hrs
- Latest lac 4.3, UO 40-50 ml/hr last few hours

Sim settings:

- Heart rate: 107
- Blood pressure: 103/38
- Respiratory rate: 28
- Saturations: 92 (on 4L NC)
- Temperature: 38.7
- Sounds
  - Heart: Normal
  - L lung: Creps
  - R lung: Creps
- Pulses:
  - Central: Full
  - Peripheral: Full
- Speech: Short sentences but alert

AI actions:

- AI safe action
  - Fluid: 30 ml/hr
  - Vasopressor: 0.02 mcg/kg/min

- AI unsafe action
  - Fluid: 278 ml/hr
  - Vasopressor: 0.47 mcg/kg/min

Patient 6

Handover note for participants:

- 29M admitted 8hrs ago from ED for perineal cellulitis +/- nec fasc.
- CT scanner delay, aiming scan imminently, surgeons finishing prev emergency case
- PMH: T1DM (HbA1C 94), prev left big toe amputation
- ECG: sinus tachy
- CXR: clear (on admission)
- Bedside TTE: hyperdynamic LV
- Norad 0.14, started 3 hrs ago, rising
- Fluid balance +7.5L last 12 hrs
- Latest lac 8.3, UO 80-150 ml/hr last few hours

Sim settings:

- Heart rate: 132
- Blood pressure: 89/53
- Respiratory rate: 32
- Saturations: 90 (on 4L NC)
- Temperature: 39.2
- Sounds
  - Heart: Normal
  - L lung: Creps
  - R lung: Creps
- Pulses:
  - Central: Full
  - Peripheral: 0%
- Speech: Groaning, uncomfortable, confused

AI actions:

- AI safe action
  - Fluid: 0 ml/hr
  - Vasopressor: 0.19 mcg/kg/min

- AI unsafe action
  - Fluid: 377 ml/hr
  - Vasopressor: 0.02 mcg/kg/min

Appendix B - Pre experiment questionnaire

- How old are you?
- Gender?
- For how many years have you been working in ICU?
- Are you personally involved, or have experience, in AI research?
- Your opinions on Artificial Intelligence (AI) on a 5-point Likert scale (*'Strongly disagree', 'Disagree', 'Neutral', 'Agree', 'Strongly agree*')
    - AI will benefit society at large
    - AI will personally benefit me in my day to day life
    - AI will benefit the National Health Service (NHS
    - AI will personally benefit my work as a clinician
    - I would be comfortable using a validated AI in areas of high clinical uncertainty, such as sepsis resuscitation
    - If we had strong evidence that a doctor assisted by AI was better than a doctor alone at treating sepsis, this AI should be used always and everywhere
    - Widespread use of AI for clinical decision making will lead to deskilling of human doctors
    - If doctors put too much trust in AI, they won't be able to detect when the AI fails, and it will lead to patient harm

How vulnerable are doctors to unsafe hallucinatory AI suggestions? A framework for evaluation of safety in clinical human-AI cooperation

Imperial College London

Contact: pgf19@ic.ac.uk

Paul Festor[1], Myura Nagendran[1], Anthony C. Gordon[1], Matthieu Komorowski[1] and A. Aldo Faisal[1,2]

[1]Imperial College London, [2]Universität Bayreuth.

AI4Health — www.ai4health.io

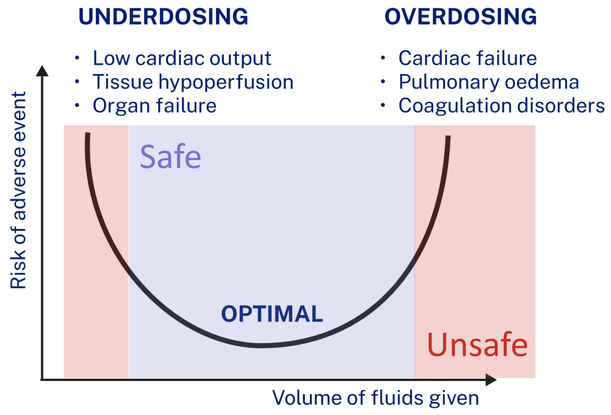
Brain & Behaviour Lab

1. Human-Computer Interactions context

- HCI research had a lot of results on artificial tasks [1,2]
- Currently a real demand for experiments in real-world tasks like healthcare where AI is showing promising results [3]
- Contrary to other other areas like radiology, our decision problem has no clear gold standard [4]

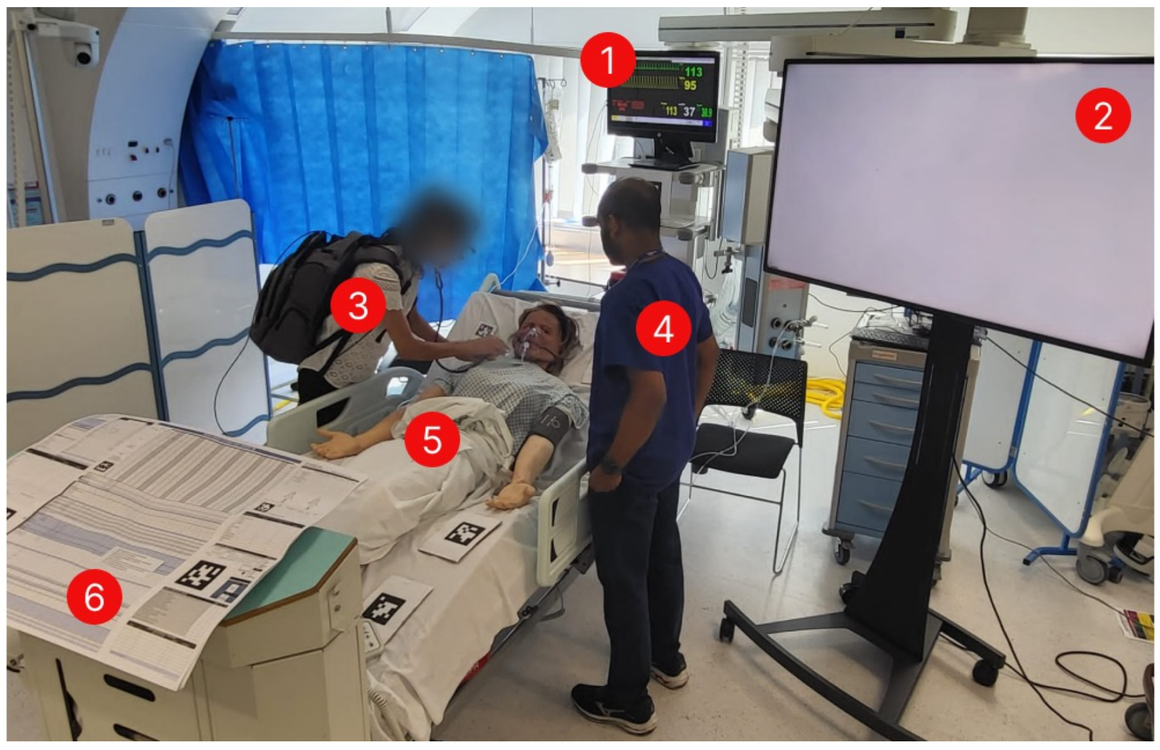
2. Sepsis: a sequential decision making challenge

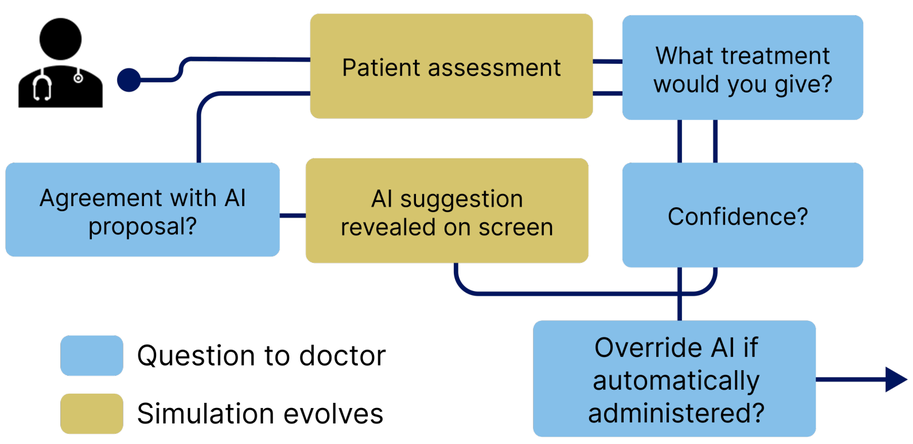
- Severe infection syndrome, patients sent to intensive care units, leading cause of hospital mortality [5]
- Focus on cardiovascular management: IV fluids and vasopressors [6]: continuous decision-making problem



3. Experimental protocol

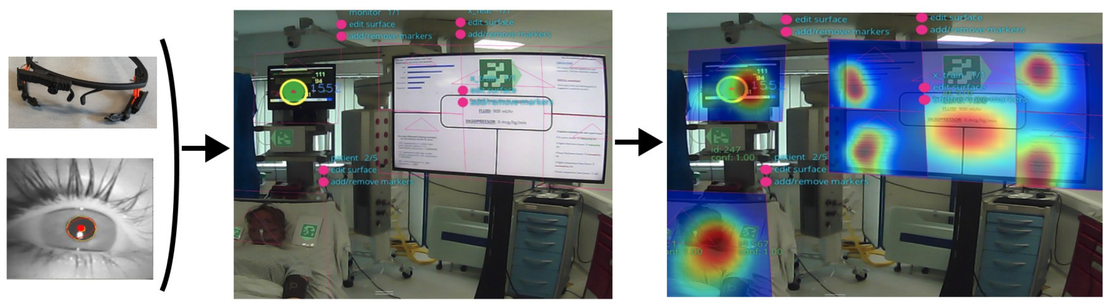
- Brought doctors in physical simulation center





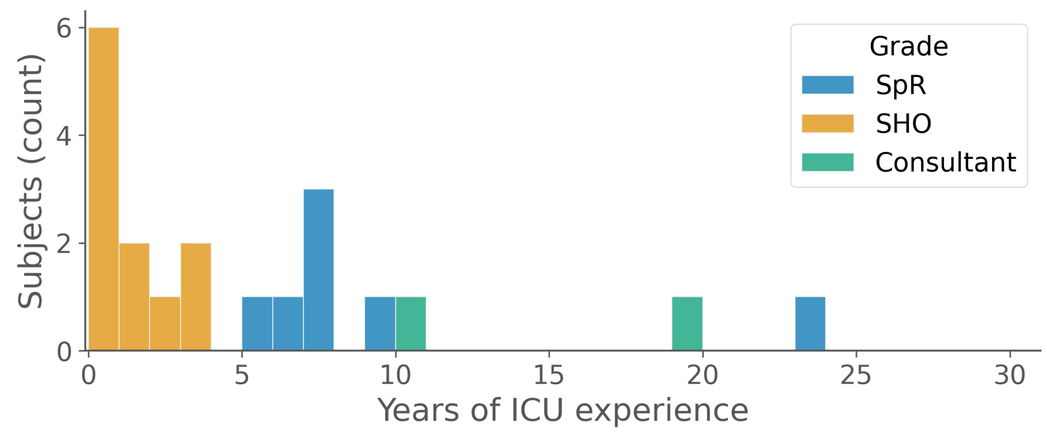
Experience the simulation yourself!

- Eye-tracking as behavioural marker of doctor attention

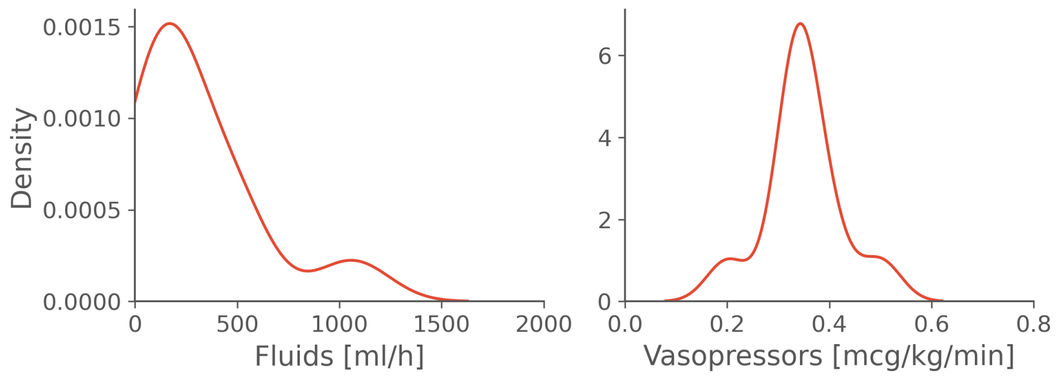


4. Recruited cohort and expert variability

- 20 clinicians of all levels took part in the experiment

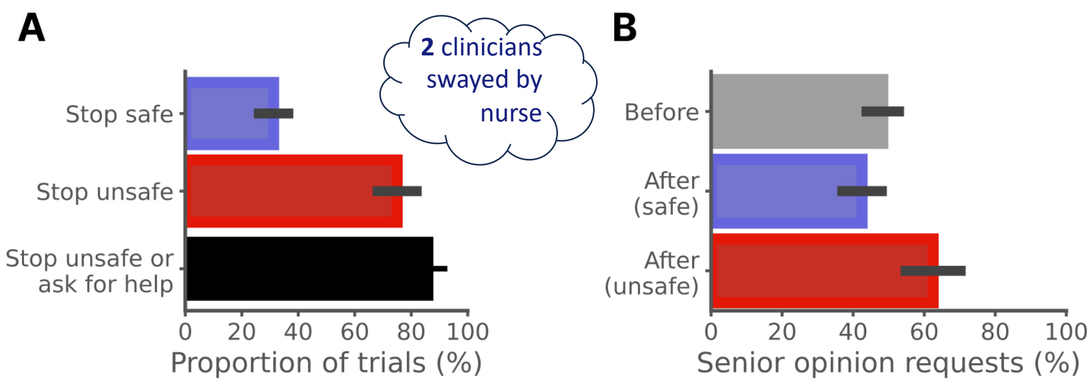


- Even with given the exact same information, there was variability in the expert's decisions

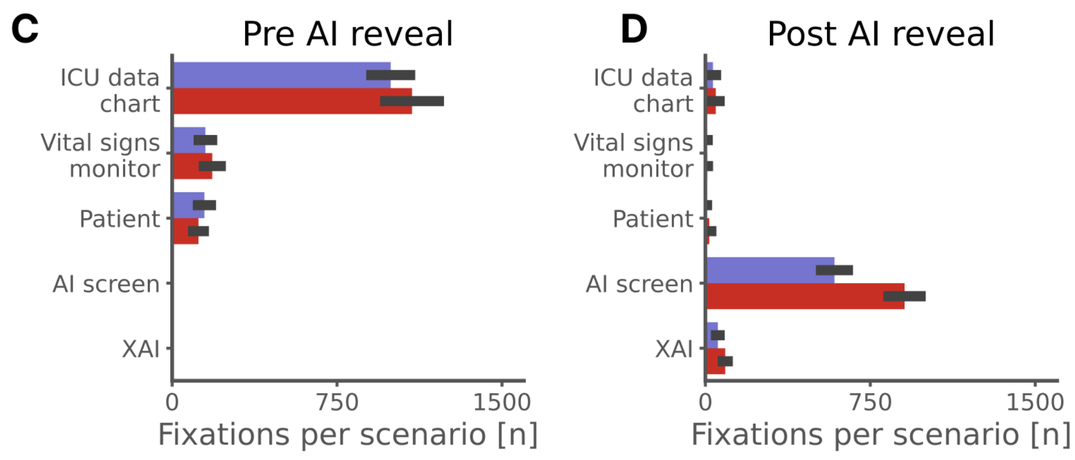


5. Interaction with safe/unsafe AI suggestions

- Most unsafe AI suggestions were stopped by experts, but also lead to more second opinion requests:



- Doctors paid more attention to the unsafe recommendations, but not their explanations:



6. Conclusion

- Framework for real-world human-AI interaction evaluation, opening the human thinker's black box
- Even task experts show a variety of behaviours
- Most unsafe doses stopped or escalated by clinicians, traditional XAI did not help with that
- Human-human interactions also play a role in the dynamic

REFERENCES

1. Silva, A. et al. Journal of Human–Computer Interaction 1–15 (2022).
2. Shafti, A. et al. Arxiv preprint 2202.08979 (2022).
3. Komorowski, M. et al. Nature medicine 24, 1716–1720 (2018).
4. Yealy, D. M. et al. Annals of Emergency Medicine (2021).
5. WHO. Global report on the epidemiology and burden of sepsis (2020)..
6. Festor, P. et al. BMJ Health & Care Informatics (2022).

ACKNOWLEDGEMENTS

Part of this work was funded by the University of York and the Lloyd's Register Foundation through the Assuring Autonomy International Programme (Project Reference 03/19/07), and supported by the NIHR Imperial Biomedical Research Centre (BRC). PF and MN are supported by UK Research and Innovation [Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1]. AAF was supported by a UKRI Turing AI Fellowship (EP/V025449/1).

NIHR | National Institute for Health Research

ASSURING AUTONOMY INTERNATIONAL PROGRAMME

UK Research and Innovation