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

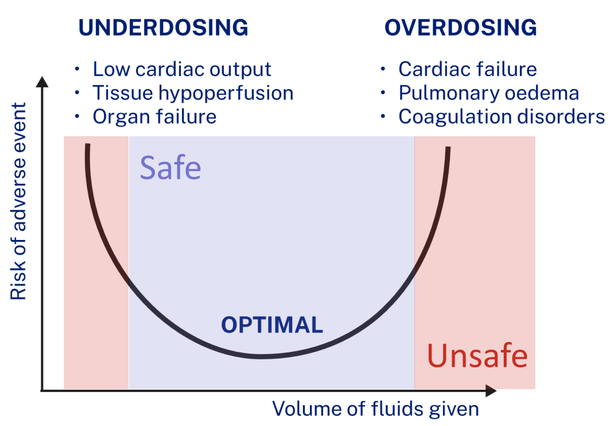
Brain & Behaviour Lab

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]

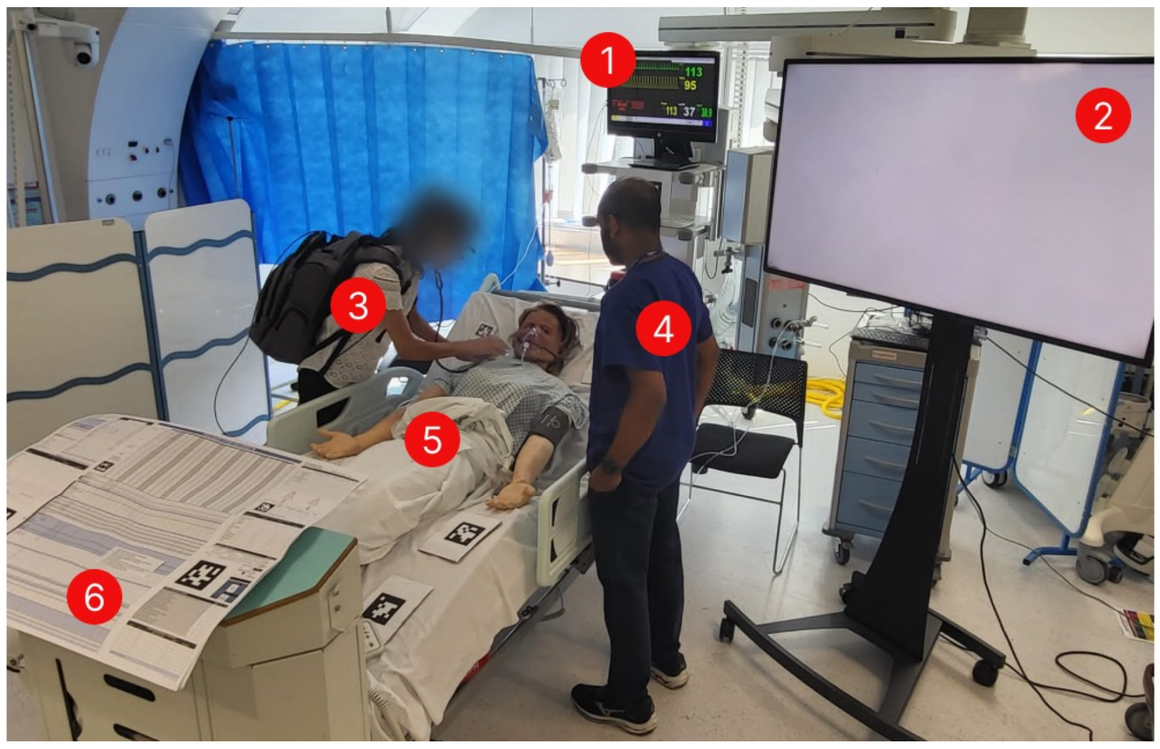
2. Sepsis: a sequential decision making challenge

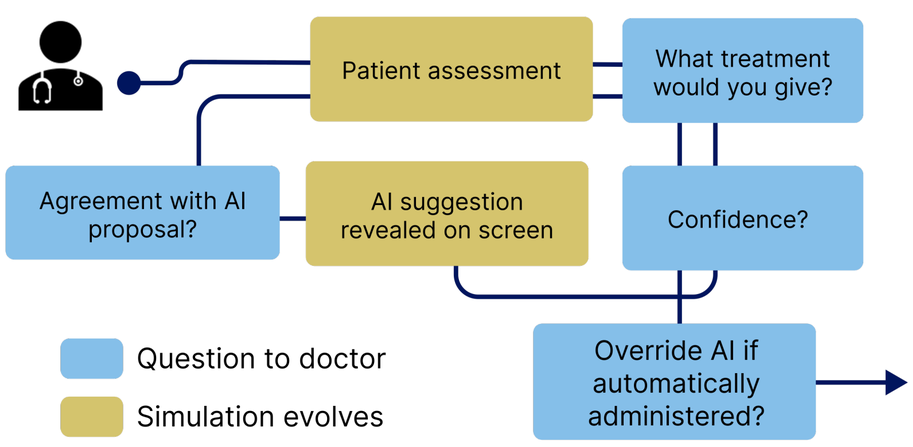
- Severe infection syndrome, patients sent to intensive care units, leading cause of hospital mortality [5]
- Focus on cardiovascular management: IV fluids and vasopressors [6]: continuous decision-making problem



3. Experimental protocol

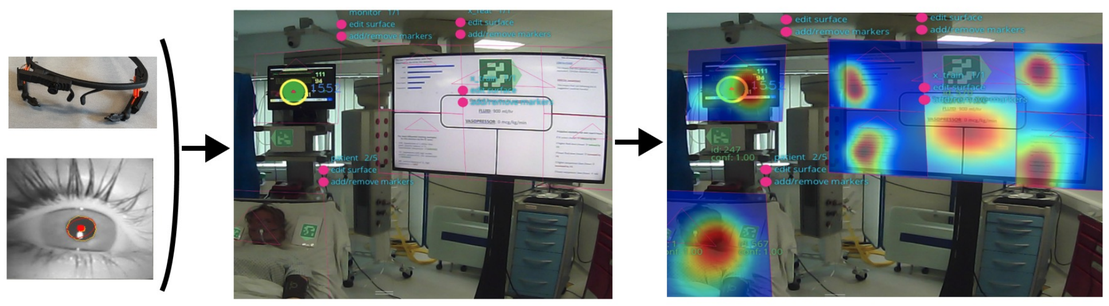
- Brought doctors in physical simulation center





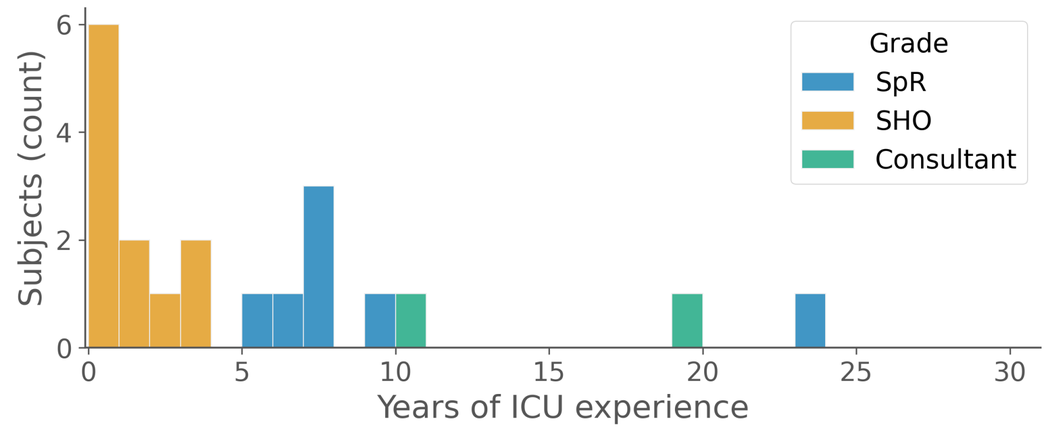
Experience the simulation yourself!

- Eye-tracking as behavioural marker of doctor attention

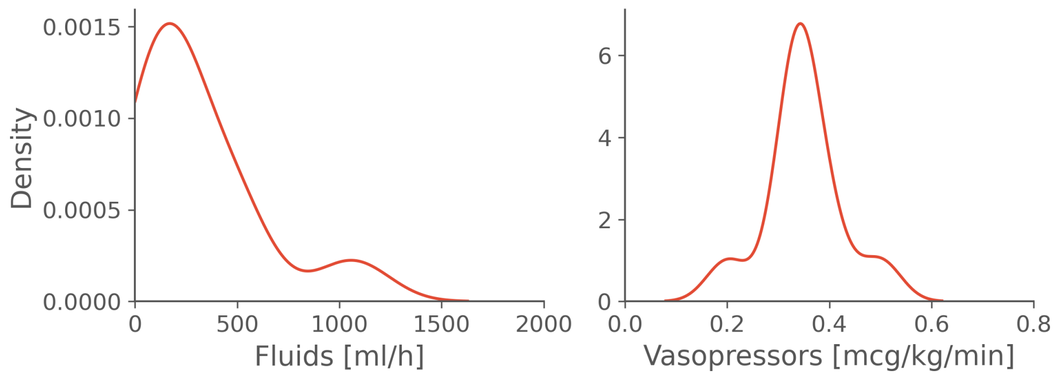


4. Recruited cohort and expert variability

- 20 clinicians of all levels took part in the experiment

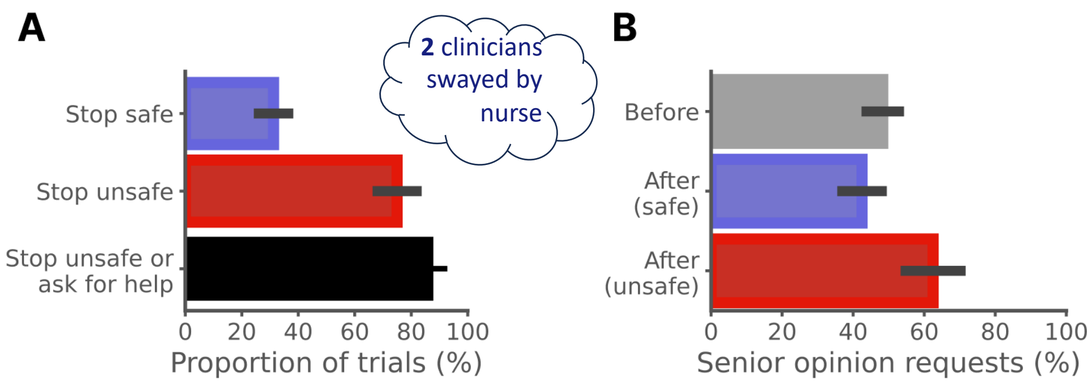


- Even with given the exact same information, there was variability in the expert's decisions

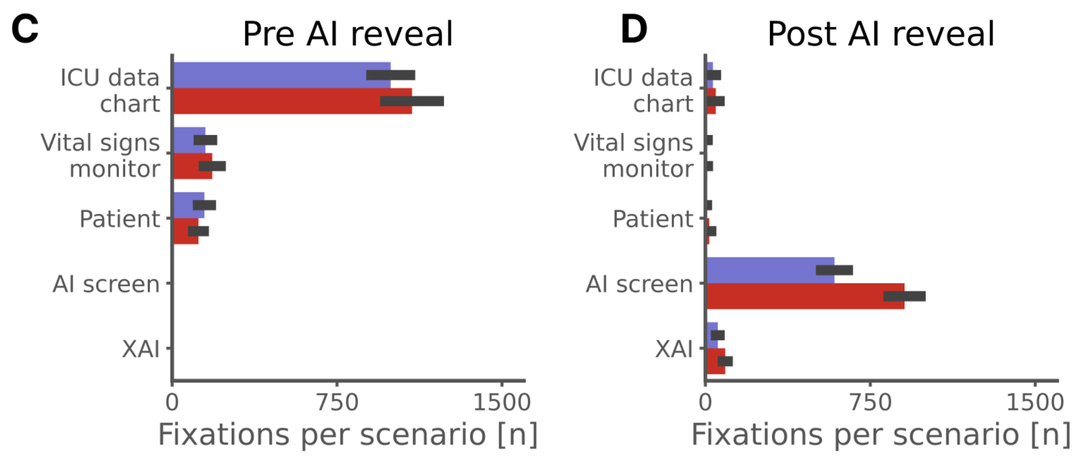


5. Interaction with safe/unsafe AI suggestions

- 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