Evaluating the explainability of AI-driven clinical decision support tools

Myura Nagendran

Supervised by Prof. A Aldo Faisal & Prof. Anthony C Gordon

Department of Computing

Imperial College London

Submitted in part fulfilment of the requirements for the degree of

Doctor of Philosophy of Imperial College London

March 2025

Statement of Originality

I hereby confirm that the work presented in this thesis is my own and that appropriate references have been used to denote the work of others. Any parts of this work that have been published have been duly acknowledged and referenced.

Myura Nagendran

March 2025

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International Licence (CC BY-NC-ND).

Under this licence, you may copy and redistribute the material in any medium or format on the condition that; you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

I would like to thank both my supervisors, Professor Aldo Faisal and Professor Anthony Gordon, for their enthusiasm, wisdom and guidance throughout the thesis. I am of course also extremely grateful to have had the support of many colleagues, friends and family throughout the last four years. Thank you to one and all.

"An explanation of cause is not a justification by reason"

C. S. Lewis

Abstract

Most clinical artificial intelligence (AI) will be realised as part of a decision support system rather than an autonomous agent. Keeping the physician in the loop means that evaluating human-AI interaction behaviour is critical to improving adoption and impact at scale, which has hitherto been lacking. Explainable AI (XAI) aims to not only provide recommendations but also to justify the rationale of the AI to clinicians and so achieve its effect via an influence on human decision-making, positively or negatively. Successful deployment of XAI is therefore as much a problem of clinicians' influenceability, cognition and psychology as one of machine learning. This thesis investigates the impact of XAI on clinical decision-making in intensive care settings, focusing on sepsis management - a domain characterised by uncertainty and practice variation.

Through innovative experiments in low- and high-fidelity settings, this work provides insights into physician-XAI interaction behaviour and shows that AI recommendations have a varied influence on prescription decisions depending on the experimental setup and XAI type. Eye-tracking in high-fidelity simulation settings was feasible and valuable for objectively assessing physician attention and cognitive load. Importantly, a consistent lack of correlation was found between physician self-reported XAI usefulness and its actual influence on decision-making, challenging the reliability of self-reports as a sole metric for evaluating XAI. The final study evaluated large language models for dynamic XAI, facilitating bi-directional influence between physician and AI while providing the potential for enhancing the clinical reasoning of both.

The findings from this thesis have important implications for all healthcare AI stakeholders involved in attempting to bring XAI-driven decision support tools toward real-world clinical implementation. Specifically, this thesis supports an emphasis on high-fidelity simulation, neuro-behavioural objective markers of attention such as real-

time eye-tracking and a more nuanced understanding of the interaction behaviour between physicians and AI in high-stakes healthcare settings.

Table of contents

List of figures **19**

List of tables **23**

1 Background **1**

1.1 AI-driven decision support tools in healthcare 1

1.2 Reinforcement learning solutions 3

1.2.1 Introduction to reinforcement learning 3

1.2.2 Reinforcement learning approaches in the ICU 4

1.2.3 AI Clinician: a case study 4

1.3 Explainable AI 5

1.3.1 Terminology and definitions 5

1.3.2 Why is explainability important? 6

1.3.3 Explainable AI (XAI) 6

1.3.4 Examples of XAI in non-clinical RL settings 8

1.3.5 Challenges in applying XAI to the ICU 11

1.4 Physician behaviour 12

1.4.1 Clinical decision support systems 12

1.4.2 Response to data from peer physicians 13

1.4.3 Response to AI recommendations 15

1.4.4 Response to XAI recommendations 17

1.5 Regulation of healthcare AI 18
1.5.1 Current regulatory landscape 18
1.5.2 Requirements for interpretability 19
1.6 Is explainability essential for building an artificial moral agent in healthcare? 21
1.6.1 The necessity of morality in healthcare 22
1.6.2 Explainability and the four ethical pillars 24
1.6.3 Building trust and accountability in AI 26
1.7 Patient and public involvement (PPI) 27
1.7.1 Rationale for PPI 27
1.7.2 PPI methodology 27
1.7.3 Insights from PPI 28
1.8 Research questions and goals in this thesis 30

2 Quantifying the impact of XAI recommendations on prescription decision making 33

2.1 Introduction 33
2.2 Methods 34
2.2.1 Data source and AI clinical decision support system 34
2.2.2 Vignette experiment and conditions 37
2.2.3 Participant recruitment and experiment conduct 41
2.3 Results 42
2.3.1 Impact of arms on dosing shift 43
2.3.2 Impact of arms on practice variation 43
2.3.3 Association of clinician factors with adherence to AI suggestions 45
2.3.4 Clinician opinions on AI and the explanations 45
2.4 Discussion 46
2.4.1 Summary of results 46

2.4.2 Limitations 47
2.4.3 Implications and comparison with other studies 48
2.4.4 Conclusion 50

3 Development of an eye-tracking pipeline to evaluate user interaction with XAI 59

3.1 Introduction 59
3.1.1 Alternatives to self-reports 59
3.1.2 How eye-tracking works 61
3.1.3 Eye-tracking in clinical settings 63
3.2 Methods 64
3.2.1 Simulation suite and XAI experiment 64
3.2.2 Equipment and software 65
3.2.3 Tagging regions of interest 66
3.2.4 Calibrating eye-tracking glasses to the subject 68
3.2.5 Post-processing of videos using proprietary Pupil Labs software 69
3.2.6 Post-processing of gaze data using Python 70
3.3 Results 72
3.3.1 Gaze 72
3.3.2 Fixations and saccades 73
3.3.3 Visual real-estate adjustment 74
3.3.4 Blink rate 78
3.3.5 Pupil diameter 78
3.4 Discussion 79
3.4.1 Summary of results 79
3.4.2 Limitations 82
3.4.3 Conclusion 85

4 Eye tracking of physicians interacting with safe and unsafe XAI 87

4.1 Introduction . . . 87

4.2 Methods . . . 89

4.2.1 Piloting of simulated sepsis cases . . . 89

4.2.2 Simulation experiment . . . 90

4.2.3 Experiment conditions . . . 91

4.2.4 Explanation modalities . . . 93

4.2.5 Participant recruitment and experiment conduct . . . 97

4.3 Results . . . 97

4.3.1 Eye-tracking metrics . . . 97

4.3.2 Clinical practice variation among physicians . . . 101

4.3.3 Influence of AI and simultaneously provided XAI . . . 102

4.3.4 Rejection of unsafe AI suggestions and second opinions . . . 108

4.3.5 Effect of nurse challenge . . . 109

4.3.6 Self-reported XAI utility . . . 111

4.3.7 Comparison to a low-fidelity non-simulation setting . . . 111

4.4 Discussion . . . 111

4.4.1 Summary of results . . . 111

4.4.2 Limitations . . . 113

4.4.3 Implications and comparison with other studies . . . 115

4.4.4 Conclusion . . . 117

5 Large language model derived explanations 119

5.1 Introduction . . . 119

5.1.1 The advantages of dynamic explainability . . . 119

5.1.2 Large language models . . . 120

5.1.3 Explainability with LLMs . . . 121

5.2 Methods . . . 122

5.2.1 Patient scenario selection . . . 122
5.2.2 Experiment format . . . 123
5.2.3 Generating LLM-derived explanations . . . 124
5.2.4 Integration with commercially available LLMs . . . 126
5.2.5 Additional analyses . . . 128
5.2.6 Participant recruitment and experiment conduct . . . 129
5.3 Results . . . 130
5.3.1 Eye-tracking metrics . . . 130
5.3.2 Clinical practice variation among physicians . . . 132
5.3.3 Influence of AI and XAI . . . 140
5.3.4 Self-reported utility of XAI . . . 141
5.3.5 LLM conversations . . . 141
5.4 Discussion . . . 151
5.4.1 Summary of results . . . 151
5.4.2 Limitations . . . 152
5.4.3 Implications . . . 153
5.4.4 Conclusions . . . 155

6 Discussion 157
6.1 Summary of findings from experiments . . . 157
6.2 Strengths and limitations . . . 159
6.3 Overarching implications and future directions . . . 162
6.4 Conclusions . . . 164

References 165

Appendix A Thesis outputs 189

Appendix B Chapter 1: Background 193

Appendix C Chapter 2: Vignette experiment **203**

Appendix D Chapter 3: Eye-tracking data **207**

Appendix E Chapter 4: Simulation suite experiment **211**

Appendix F Chapter 5: Large language model experiment **223**

List of figures

2.1 Coverage of patient scenarios over the MIMIC-III state space 35
2.2 Feature importance for a toy example with two patients 36
2.3 Feature importance for an example patient at time block 1 38
2.4 AI and XAI screens 39
2.5 Trial matrix for vignette experiment 40
2.6 Example of dose shift metric 41
2.7 Experiment protocol 42
2.8 Dose variability by intervention arm 44
2.9 Aggregate dose shift by intervention arm 51
2.10 Principal component analysis of attitude to AI 52
2.11 Impact of AI attitude on adherence to AI suggestions 52
2.12 Impact of duration of clinical experience on adherence to AI suggestions 53
2.13 General post-experiment questions 53
2.14 Post-experiment questions related to XAI utility 54
2.15 Boxplots of prescription dose distributions by trial arm 1/3 55
2.16 Boxplots of prescription dose distributions by trial arm 2/3 56
2.17 Boxplots of prescription dose distributions by trial arm 3/3 57

3.1 Simulation suite 65
3.2 Eye-tracking glasses and pupils 66

3.3 April tags 67
3.4 Video post-processing 73
3.5 Gaze distance and speed 74
3.6 Demonstration of fixations, saccades and scanpath 75
3.7 Fixations on various surfaces 76
3.8 Fixation and saccade metrics stratified by trial safety status and experience level 77
3.9 Fixation to saccade duration ratios stratified by trial safety status and experience level 78
3.10 Position of AI screen ROI over 500 frames 79
3.11 Gaze proportion per ROI 80
3.12 Ratio of actual to chance gaze per ROI 80
3.13 Blink rate stratified by trial safety status and experience level 81
3.14 Pupil diameter over time for a single subject 82

4.1 Example of instance-based XAI in imaging 95
4.2 Consistency assessment for instance-based XAI 96
4.3 Physician experience and attitudes 98
4.4 Gaze fixations per scenario by safety status of scenario 99
4.5 Fixations per scenario by safety status of scenario and experience level of doctor 100
4.6 Gaze fixations per scenario 101
4.7 Mean fixation duration and blink rate by region of interest 102
4.8 Correlation between practice variation and eye tracking metrics, for both fluid and vasopressor 103
4.9 Blink rate by influence of AI for both fluid and vasopressor and for both safe and unsafe AI suggestions 104

4.10 Number of gaze fixations by influence of AI for both fluid and vasopressor and for both safe and unsafe AI suggestions 105
4.11 Blink rate by degree of adherence to AI for both fluid and vasopressor and for both safe and unsafe AI suggestions 106
4.12 Number of gaze fixations by degree of adherence to AI for both fluid and vasopressor and for both safe and unsafe AI suggestions 107
4.13 Physician responses regarding stopping AI suggestions and seeking second opinions 108
4.14 Usefulness rating for each type of explanation as well as overall rating . 109
4.15 Fixations on XAI vs self-reported usefulness 110
4.16 Comparison between prescription behaviour between vignette and simulation settings 112

5.1 Physician experience and attitudes 130
5.2 Fixations per scenario by AI or XAI condition of scenario 131
5.3 Mean fixation duration and blink rate by region of interest 132
5.4 Gaze proportion per ROI and ratio of actual to chance gaze per ROI . 133
5.5 Correlation between practice variation and eye tracking metrics, for both fluid and vasopressor 134
5.6 Blink rate by influence of AI/XAI for both fluid and vasopressor 135
5.7 Fixation duration by influence of AI/XAI for both fluid and vasopressor 136
5.8 Blink rate by degree of adherence to AI/XAI for both fluid and vasopressor 137
5.9 Fixation duration by degree of adherence to AI/XAI for both fluid and vasopressor 138
5.10 Boxplots of prescription dose distributions by trial arm comparing vignette and simulation (LLM) experiments 139
5.11 Overall self-reported XAI usefulness rating against eye-tracking metrics 142
5.12 Heatmap for the breakdown of specific XAI rating questions 143

E.1 Trial matrix for XAI simulation suite experiment 212

F.1 Initial explanation data in raw CSV format 230

List of tables

4.1 Nurse challenge verbal prompts 92

D.1 Description of fields in eye tracking gaze position data (here using the example region of interest of AI screen) 207

D.2 Description of fields in eye tracking gaze position data 208

D.3 Description of fields in eye tracking fixation data (here using the example region of interest of AI screen) 208

D.4 Description of fields in eye tracking data (surface positions AI) 209

D.5 Description of fields in eye tracking blink data 209

Chapter 1

Background

1.1 AI-driven decision support tools in healthcare

Uncertainty in sepsis treatment

Critically ill patients with sepsis are at risk of circulatory failure that can lead to septic shock and death [1]. Septic shock is treated using intravenous fluids and vasopressor drugs (commonly norepinephrine/noradrenaline) that maintain an optimal blood pressure for organ perfusion [2]. However, these treatments have varying risk-benefit profiles and it falls to the astute clinician to select the right treatment at the right dose for the right patient at the right time [3]. Traditional teaching for sepsis advised administering large volumes of resuscitation fluid up front before progressing to vasopressor drugs where necessary [4–6]. An increasing body of observational work now suggests that this approach might be potentially harmful and that excess fluid could increase morbidity [7, 8]. Accumulation of randomised trial evidence in this sphere is ongoing and should give us useful evidence on which to base adjustment to guidelines [9–11]. Ultimately, increased evidence in this sphere will still only inform the average treatment effect rather than what the optimal approach to any given individual patient might be. Therefore what is needed is a more individualisable

approach. Although there are tools that can provide individualised real-time data on cardiac output (waveform monitors and point of care echocardiography for example [12]) these tools also have limitations such as extensive training requirements or lack of integration with other patient data.

One important factor that restricts overall clinician performance is that human clinicians are limited in the total volume of patients that they will see over their career and therefore, to coin a machine learning term, limited in the amount of 'training data' available to them. Even an experienced consultant intensivist who has been working for two to three decades may only have seen in the region of several thousand patients rather than hundreds of thousands or millions that may have attended their hospital network during the same period, let alone the wider healthcare system. Furthermore, clinicians exhibit both recency bias and recall bias in that they typically remember recent and exceptional cases more frequently [13]. A big data approach that leverages the electronic health record (EHR) to form individualised treatment recommendations could improve clinical outcomes from sepsis [14]. The intensive care unit (ICU) is among the most data rich environments within the hospital and has consequently been the focus of many machine learning efforts thus far [15–17].

Intensive care staffing and logistics

Beyond improvement in clinical outcomes, there is also a pressing need to improve the efficiency of care delivery. ICU admissions are one of the most expensive aspects of secondary care [18]. The number of ICU beds per capita of the population is also lower in the UK compared to similar developed countries [19, 20]. This bed pressure can be a rate limiting factor for timely admission of critically ill patients and may impact on mortality rates [21]. This was especially apparent during the COVID-19 pandemic [22]. As an example, London has approximately 800-1000 intensive care beds for a population of 8.9 million people [23]. During COVID, an emergency field hospital

known as the 'Nightingale Hospital' was rapidly constructed in a conference centre with capacity for 4,000 critically ill COVID patients [24]. In actuality, the rate limiting factor in admitting patients, by an order of magnitude, was staffing rather than beds, drugs or equipment. AI decision support tools could lower the heavy burden on staff in two immediate ways. First, by reducing the time spent on some aspects of clinical decision making and thus freeing staff to perform other tasks and especially those requiring communication with patients and families [25]. Second, by paying constant 24/7 attention to small details of the patient's treatment that human clinicians do not have the logistics or bandwidth to keep constant surveillance on. A classic example from COVID was the frequency of minor ventilator setting changes made [26].

1.2 Reinforcement learning solutions

1.2.1 Introduction to reinforcement learning

The machine learning approaches toward solving ICU based problems can be broadly divided into two main categories: supervised learning approaches to prediction/diagnosis problems [27–29] and reinforcement learning (RL) approaches to treatment decisions [30] (although unsupervised learning has also been applied to discover new patient phenotypes [31, 32]). RL is an attractive option to use for sequential decision making problems such as those faced by clinicians and has been applied successfully to several non-clinical contexts such as games including Chess, Go and Starcraft II [33–35]. RL, in its simplest form, is a type of machine learning in which an artificial agent learns a set of rules (or a policy) by trial and error that teach it how to act in any given state so as to maximise a reward [36, 37].

1.2.2 Reinforcement learning approaches in the ICU

RL partially mimics the clinical context in which a clinician learns by experience (trial and error) which treatments to prescribe for any given situation so as to maximise a good patient outcome. RL approaches have been applied to several areas of healthcare including oncology chemotherapy regimens [38], human immunodeficiency virus (HIV) treatment regimens [39] and diabetes management [40]. Within the ICU, RL has been applied to the management of ventilation [41, 42], sedation [43], antimicrobial choice [44] and sepsis treatment [45–47] among others. However, there are no RL agents approved for general clinical use outside of a research setting as yet. This remains a novel field with little reliable information to guide safe clinical use of such systems.

1.2.3 AI Clinician: a case study

The 'AI Clinician' is an artificial intelligence clinical decision support system (AI-CDSS) developed at Imperial College London. It uses an RL agent to provide personalised treatment recommendations for patients with sepsis after learning an optimal treatment policy from EHR data [45]. The AI Clinician can produce real-time recommendations to clinicians on the optimal dose of fluid and/or vasopressor. Initial retrospective off policy analyses suggest that these recommendations may improve the likelihood of survival at 90 days [45], with further supporting retrospective evidence from external clinical databases in other geographical settings [48]. Prospective live testing is currently underway at various sites in London in an attempt to build up to a large scale cluster randomised trial of efficacy and safety. Further work has also continued in the background to advance the algorithm itself by incorporating partially observed Markov decision processes (POMDPs) and a continuous state and action space to better account for the variability in availability of data and the fact that clinicians tend to prescribe over a continuous range rather than in discrete bins [49, 50]. However, the

success of the AI Clinician and its scalability will depend on more than just the overall accuracy or safety of its recommendations.

For an AI-CDSS to have a real-world positive impact within the ICU requires that the recommendation is efficiently presented to the clinician and that the clinician trusts it enough to act on it [51]. An AI recommendation that is provably effective (for example from randomised trial evidence) but that is not followed by a human clinician is of no benefit to the patient or the wider healthcare system. This means that we need a better understanding of the interpretability of the AI output to the clinician and a better understanding of the factors that might influence their trust in, and use of, such an AI-CDSS [52]. Understanding these two facets will help to design an optimal system. This situation is made more problematic by the fact that as the complexity of RL algorithms (and especially deep RL) increases, explainability of RL agent behaviour becomes more difficult to ascertain [53].

1.3 Explainable AI

1.3.1 Terminology and definitions

The field of explainable AI (XAI) seeks to investigate and offer solutions to how we can discover the rationale for why an AI system works the way it does. There are many related terms to explainability in the literature including interpretability, transparency, comprehensibility, understandability and trust among others. There is not complete consensus even within the AI community on the nuances between definitions for these closely related terms [54]. After considering many related concepts and terms in their comprehensive review, Arrieta and colleagues propose the following global definition which I will use for the purposes of this thesis. They define explainable AI as follows: *"Given an audience, an explainable Artificial Intelligence is one that produces details*

or reasons to make its functioning clear or easy to understand" [54]. The emphasis on the audience here is a key factor which will be discussed later.

1.3.2 Why is explainability important?

Clinicians hold the medico-legal responsibility for the decisions they make, even those decisions supported by a CDSS. They must be able to justify and explain their decisions to patients, their peers and the regulators. If this decision was partially based on an AI-CDSS, then the AI-CDSS will need explainability to allow the clinician to make said justifications. This is not to say that a black box model cannot exist. The justification for using an opaque system might be to point out the strong evidence of efficacy of the AI's suggestions. However, high level evidence will take time to gather and there are other advantages to AI explainability including but not limited to its impact on trustworthiness, confidence, informativeness, fairness, establishing causality, interactivity and privacy awareness among others [54]. In the ICU for example, an explainable AI could engender more trust by presenting a causal rationale in a similar manner to how clinicians explain their rationale to other clinicians. Explainability might also strengthen the interactivity of an AI-CDSS by allowing clinicians to probe the system in cases where the AI suggestion appears unusual or counter to the status quo.

1.3.3 Explainable AI (XAI)

Biran et al. conducted a survey in 2017 of explanation and justification in machine learning [55]. They argued that an explanation provides the 'how' whereas a justification also provides the 'why' for a decision being a good one. A justification may not include a detailed mechanistic 'how' at all in fact. This does not seem as useful for our purposes as the justification in a clinical scenario usually stems from evidence that the decision is effective. So in the case of the AI Clinician, this would probably come from a

prospective clinical trial. The explanation of 'how' can still be useful though in the building of trust with users.

Doshi-Velez and Kim state that interpretability can also be useful when it is not possible to verify easily if a prediction is accurate [56]. In the ICU context, for any given clinical scenario, there is no gold standard consensus on the optimal amount of fluid or vasopressor to prescribe. Interpretability therefore allows some insight into whether the underlying reasoning might be plausible. The authors also point out a lack of broad agreement on the optimal way to evaluate interpretability for the purposes of bench-marking and so use their review to try and form a more definitive framework. The result is a three level taxonomy of evaluation approaches for interpretability: functionally grounded evaluation which uses no real humans and proxy tasks, human-grounded evaluation which uses real humans but simple tasks and application-grounded evaluation which uses real humans and real tasks [56]. It is clear that an AI-CDSS that will be used in the ICU will need a heavy, if not exclusive, focus on the third level of evaluation i.e. real clinicians testing the system's interpretability on a real task. Doshi-Velez and Kim, referring to application-grounded evaluation, make the important point that *"we as a community respect the time and effort involved to do such evaluations, and also demand high standards of experimental design when such evaluations are performed"* [56].

Doshi-Velez and Kim also discuss different spectra of task similarity with respect to explanations [56]. For example, some tasks might be better suited to local vs. global explanations. Some tasks might have time constraints for the decision in question which mean that an explanation has to be very brief and easily understandable. Or perhaps the nature of user expertise might affect cognitive chunk size and thus lead to a requirement for tailored explanations to take this into account. These are all aspects that can potentially be separated out, controlled and investigated in an experimental fashion.

A potential trap highlighted by Miller and colleagues is the importance of defining the appropriate target audience for an explanation before deciding on the best way to generate it [57]. The kind of explanation useful to a clinician, who may have no background or knowledge of AI, will need to incorporate clinical domain awareness while avoiding AI jargon. In contrast, an explanation aimed at a machine learning engineer for the purpose of debugging aberrant model behaviour would likely need the opposite properties. Miller et al. suggest that incorporating social and behavioural sciences into XAI approaches will ensure a more user-centred focus. From an experimental point of view, this might mean that we need both quantitative and semi-qualitative approaches to understand how clinicians react to explanations.

1.3.4 Examples of XAI in non-clinical RL settings

Conventional supervised learning was an initial source of inspiration for RL. There are many prominent examples of XAI in supervised learning, typically using popular techniques such as SHapley Additive exPlanations (SHAP) [58] and Local Interpretable Model-Agnostic Explanations (LIME) [59]. SHAP uses a game theoretical approach to assign each feature an importance value for a particular prediction. It then calculates Shapley values, which represent the average marginal contribution of a feature across all possible combinations of features, thus providing a unified measure of feature importance that is both consistent and locally accurate. LIME generates locally faithful explanations by learning an interpretable model around individual predictions. It functions by perturbing the input data and observing the impact on the model output and then fitting a simple, interpretable model (for example a linear regression) to this local behaviour which can then be used to explain the original prediction. AI within the ICU has used such techniques, e.g. an AI model predicting critical illness from EHR data with explanations provided via SHAP values [60]. The recognition that RL agents will become more commonplace is now extending this XAI focus to the RL

sphere as well [61]. However, there are specific challenges to XAI in RL. For example, RL decisions may be related to each other in a way that isolated supervised learning predictions are not. This could impact upon the complexity of the resulting explanation. Furthermore, RL does not always use training data in the same way as supervised learning (for example, an agent's training may occur as a result of interactions with the environment) [62].

The Dot-to-Dot approach by Beyret et al. uses hierarchical RL (HRL) in a robotic arm manipulation task [63]. A high-level RL agent divides the full task into smaller, easier actions (or sub-goals) that a low-level agent can learn. It is the dynamics of the environment and representations learned by the high-level agent that provide the explainability. The approach makes use of two other key aspects. First, a deep deterministic policy gradient (DDPG) which applies deep learning (DL) to a DPG, itself an actor-critic architecture where the actor performs an action and the critic evaluates with respect to goals using temporal difference learning. Second, hindsight experience replay (HER) is used to assess every trace whether the goal was reached or not in that trace. This allows for learning from mistakes and is an efficient way of speeding up training in situations with large state and action spaces. Their findings demonstrated that the agent learnt a notion of distance. It was able to create way-points to an end-goal by splitting a complex high-level task into several simpler low-level consecutive tasks and reusing learnt behaviours across these while providing some human level interpretability to this process [63]. The high-level agent valued different parts of the environment as varying in suitability for the low-level agent's sub-goals (displayed as a visual map in the paper). The authors state that this allowed interpretation into how the agent internally represents the environment and therefore a degree of insight into how the agent is making decisions [63].

Xu and collagues also used hierarchical RL to enhance interpretability of an RL agent though the task this time was the real time strategy game Starcraft II [64].

However, their sub-goal selection strategy makes use of human expertise as well, rather than pure end-to-end learning. They suggest that this is more efficient for problems with large state action spaces. As well as a suitable decomposition of the learning task, their approach also takes into account sequential relationships (i.e. some states might be prerequisites for others). This has obvious attraction to healthcare RL agents where the sequence of actions and states can hint at causality which may be a useful component of explanations to clinicians [65].

A different approach is taken by Madumal et al. in which they use a structural causal model during RL to encode causal relationships between variables of interest [66]. This allows 'why' and 'why not' explanations to be generated via a counterfactual analysis of the causal model. They applied this model to the game Starcraft II with 120 participants and found that the explanations improved understanding by humans of the agent's behaviour and rationale with superior explanation 'goodness' (explanation satisfaction and sufficiency) compared to both a baseline explanation model (state relevant variables visualised via a state-action graph) and a model with no explanation provided. However, a weakness of their approach is the requirement for the causal model upfront. In further work, the same authors deploy a structural causal (action influence) model and use it to learn opportunity chains of the form 'A enables B and B causes C' [67]. This model is used to extract reasons for action selection from a surrogate policy. The interpretable surrogate policy is learned in the form of a decision tree using batched replay data. The authors refer to this as a distal explanation and evaluate it in a human study of participants seeing explanations for Starcraft II RL agent behaviour. They find that the explanations were superior in certain scenarios (e.g. collaborative as opposed to adversarial or rescue scenarios) and conclude that further research is needed to understand the impact of distal explanations on other scenarios [67]. It is interesting that even within the same high-level task (playing Starcraft II) that certain scenarios have different explanation 'goodness' (i.e. explanation satisfaction and sufficiency) and

this may have consequences for the generalisability of approaches from contexts like Starcraft II to the clinical domain. It is likely that even within clinically deployed RL agents, there is no one size that fits all and an RL agent in the ICU, for example, may have very different explainability requirements to an RL agent in a surgical clinic.

All of the aforementioned XAI approaches to RL are promising but come with varying strengths and weaknesses and the degree to which they could be easily applied to the AI Clinician, if at all, is unclear. And if they can be applied, how might clinicians react to them?

1.3.5 Challenges in applying XAI to the ICU

The variability of the clinical environment makes it a challenging sphere in which to deploy AI and the ICU is certainly no exception. The clinical pace can vary rapidly from an unexpected resuscitation scenario requiring decisions within minutes to a slower paced ward round in which 30 minutes may be spent at the patient's bedside reviewing data and making plans or preparing for a multi-disciplinary team (MDT) meeting to discuss more complex cases. The decision types vary as well on a spectrum from frequent low-importance decisions (e.g. an individual fluid bolus) to rarer high-stakes decisions (e.g. whether to extubate a patient or transition them to an end of life care pathway). There is also the variability in experience between junior and senior doctors that may have a profound impact on the way in which they integrate data or interact with AI-CDSS tools, and even to the extent that they may be led astray by such systems. All of these factors and more will shape which XAI approaches are most suitable and in what context (for example it may be that different explanations are suited to junior versus senior doctors, or that one group is more likely to use explanations in their rebuttal of dangerous AI advice).

Existing work with user-studies has suggested that an increase in explanation complexity increases response time and decreases the subjective satisfaction with the

explanation [68, 69]. Another key study was conducted by Tonekaboni and colleagues in which they interviewed 10 clinicians (6 from the ICU and 4 from the emergency department) to seek their views on AI-CDSS and uncovered many useful points [70]. For example, *"some clinicians were interested in seeing similar samples to their patients because they were curious to know what actions were taken in those cases, and the associated outcomes of the interventions"*. Other clinicians highlighted that they viewed explainability as a means of justifying their decision-making (for example, to colleagues and patients). The clinicians also reported that explanations based on previous training examples were not useful in time-constrained scenarios.

1.4 Physician behaviour

1.4.1 Clinical decision support systems

Clinical decision support systems (CDSS) have been popular in medicine for several decades as a means of leveraging built up knowledge to positively influence clinicians in the course of their decision-making. However, the evidence base on the influence of CDSS on clinician behaviour is variable. A review of randomised trials on inpatient CDSS found no significant increase in adoption of desired behaviour, even when restricting studies to purely EHR based CDSS (as opposed to non EHR-based pager alerts and messages) [71]. A more comprehensive meta-analysis from 2020 reviewed the absolute level of improvement that might be expected from computerised CDSS across diverse clinical settings and tasks (including outpatient settings). Among the 68 trials focused on prescribing there was a 4.4% absolute improvement (95% confidence interval 2.6% to 6.2%) favouring CDSS over control, albeit with unsurprisingly large heterogeneity [72]. The authors classified this as a small to moderate effect size, suggesting the smaller-than-expected impact may be due to the lack of a comprehensive socio-technical implementation model. Such a model would consider factors such as staff

training, organisational culture, workflow, software/hardware infrastructure and alert fatigue among clinicians. Other work has looked to identify specific factors that might affect CDSS adoption, for example with antibiotic prescribing. One cross-sectional survey found that primary care doctors were more likely to perceive time constraints and patient preference as barriers to CDSS adoption (the former of which would also be important in ICU while the latter less so for everyday prescribing decisions) [73]. They also found that experienced clinicians were more likely to distrust CDSS due to concerns about impingement on their clinical judgment [73]. A systematic review of 63 articles on facilitators and barriers to medication related CDSS adoption reported that, once again, time constraints were viewed as a barrier with time-saving seen as a facilitator. Clinicians wanted simple systems and remarked that it was a hindrance to have to search for hard-to-find information. Furthermore, they emphasised relevance and usefulness as important modifiers of behaviour with concerns about alert fatigue [74].

Some have criticised non-AI driven CDSS for having fewer inferential connections of value to physicians than AI-derived systems (for example, most traditional CDSS are not knowledge or insight generating but serve more as reminders of already established best practice which ought to be followed). However, a recent systematic review refutes this claim. Vasey and colleagues found no substantial improvement in diagnostic performance from AI-based CDSS and specifically noted the limited attention given to human factors during implementation within the reviewed studies [75]. For instance, user feedback was reported in only four of the 37 studies [75].

1.4.2 Response to data from peer physicians

Human behaviour is heavily influenced by awareness of peer actions in similar contexts [76]. This harnesses social norms to promote behavioural moderation and reduce outlier tendencies [77]. Normative comparison feedback enables individuals to benchmark their

performance and in doing so adjust discrepancies in their self-assessment and metacognitive perceptions [78]. This principle also applies to physicians where understanding peer prescribing behaviour can guide and influence individual prescribing practices [79]. A clinical example which has been well studied is opioid prescribing by emergency department physicians where over-zealous prescribing can leave patients vulnerable to risks of longer term addiction [80]. Initial studies had combined provision of peer normative comparisons along with audit and more general feedback finding that average opioid prescription rates fell [81, 80, 82]. This was further highlighted in a randomised trial that compared audit and feedback alone to audit and feedback combined with peer norm comparisons with the aim of isolating the added benefit of the peer comparisons. The marginal improvement attributable to the peer norm comparisons was considerable with a 76% reduction in the opioid prescription rate for the combined peer norm and audit group compared to 53% for the group with only general purpose audit and feedback [83]. Human-human interactions are also commonplace in medicine and no doubt also have a more direct impact than just the provision of feedback alone (i.e. witnessing the behaviour of others might be influencing). For example, Yuan et al. performed a combined interview and observational study of EHR super users on two different wards and the effect of EHR uptake among clinicians overseen by these super users [84]. They found that super user behaviour significantly impacted implementation success, for better or worse. For example, super users who offered detailed explanations with proactive positive framing and shared information freely showed significantly greater improvement in clinician proficiency with the EHR system. Countering this, some studies suggest that decision-makers typically give less weight to external information compared to their own, and processing additional information can be cognitively taxing [85]. The nature of how certain peer information is presented might also modify this effect with evidence from the social sciences that people prefer to receive advice with

some notion of probability or likelihood and that advisors benefit from expressing themselves with confidence though without communicating false certainty [86].

When considering AI, comparing the source of advice (i.e. from AI or peer humans) has been assessed recently in a diagnostic setting. Gaube and colleagues found no significant difference in accuracy regardless of whether the source of advice given to task experts (radiologists) or task non-experts (emergency and internal medicine doctors) was from AI or fellow humans [87]. They were able to use accuracy as their outcome of interest because the diagnostic task had a gold standard to compare against which is absent in most ICU prescribing tasks. Of critical importance, Gaube and colleagues also found that expert physicians rated AI advice of lower value than peer human advice, yet this did not impact their reliance on it. This suggests a discordance (even having controlled for individual differences) between reported perceptions and actual behaviour which makes evaluating physician interaction with AI even more difficult (i.e. it suggests that simply asking people what they *think* may not provide an accurate picture of what they *do*).

1.4.3 Response to AI recommendations

A number of other studies have also looked at how clinicians interact with and respond to AI-CDSS recommendations. In a collaboration study with behavioural economists, Agarwal et al. studied radiologists and found that while AI support has potential benefits, the biases in clinician use of AI often negate these gains [88]. Despite the AI tool outperforming two thirds of radiologists, providing AI predictions did not typically improve radiologist performance unless the AI certainty was very high in either direction. The clinicians frequently underweighted AI predictions as well as erroneously believing that the AI suggestion and their own decision were independent. The result was suboptimal decision-making that took longer than had they been unsupported with the potential suggestion that there may be many situations where

AI and clinicians should work separately rather than together for optimal outcomes [88].

A survey study of 119 clinicians by Choudhury et al. on an AI-driven blood utilisation calculator (which suggested how many units of blood to transfuse to a patient based on their clinical data) found that most clinicians had a neutral to moderately high expectation of how AI would perform and how much effort it would require [89]. Higher expectations led to a lower perception of risk; i.e. when clinicians believed AI would work well, they saw it as less risky. Similar findings were apparent in a survey by Shamszare et al. of 265 clinicians from a variety of specialties aiming to understand perceptions toward integrating AI into healthcare [90]. Clinicians were more likely to trust and use AI if they saw it as potentially reducing their workload. Conversely, those who perceived greater risks with AI were less likely to adopt it. Interestingly, clinical experience was not a major factor in influencing use of the AI in decision making.

This observation that experience was not an important modifier of clinician behaviour with AI-CDSS was emphasised in a study by Yu and colleagues who studied the varying effects of AI support on a large sample of 140 radiologists across 15 chest radiograph diagnostic scenarios [91]. They noted that traditional factors such as familiarity with AI and length of clinical experience did not reliably predict the effect of AI support. Furthermore, they found no supporting evidence for the commonly posited theory that less well performing participants gain the most from AI support. Accurate AI predictions led to better treatment effects, while inaccurate predictions seemed to mislead radiologists (which occurred as the clinicians did not seem able to consistently distinguish between AI support that was accurate or inaccurate). The need for human users to better detect inaccurate AI advice is critical and the authors suggest that explainability could be useful here.

1.4.4 Response to XAI recommendations

How clinicians interact with XAI-CDSS is also being studied more frequently in recent years given the interest in the topic. In a follow-up to their original study, Gaube et al. investigated a radiology task again but this time with XAI support [92]. They observed that experts had higher confidence in their diagnoses when the advice was reported as originating from an AI, despite no change in accuracy. Meanwhile non-task experts accuracy improved significantly with explanations compared to experts whose baseline performance was higher and therefore may have had less room to increase [92]. Nonetheless, the authors suggest that XAI-CDSS could therefore be a useful tool to increase performance in non-experts (which has significant ramifications for equity of healthcare provision globally given the limited number of experienced healthcare practitioners relative to demand).

Schwartz et al. explored clinician trust in an AI-CDSS using a qualitative descriptive methodology based on semi-structured interviews of 17 participants [93]. They found that perceptions of AI-CDSS accuracy were influenced by the degree of concordance between the AI-CDSS predictions and what clinicians perceived of the patient's clinical condition. Clinicians' perceptions of system accuracy were influenced by how well the AI-CDSS predictions matched their own clinical impressions, while understandability was shaped by the quality of AI-CDSS explanations and training. Wysocki and colleagues performed a within-subject study of 23 clinicians who were exposed to an AI decision support tool assisting with diagnosis and disposition of oncology patients with suspected COVID [94]. Five cases were seen with only the AI recommendation alone and then five that were accompanied by a SHAP-based feature importance explanation. Clinicians recognised explanations as crucial for safety and trust in their comments to the investigators. However, communicating model uncertainty was regarded as an even stronger requirement than explanations per se. Moreover, explanations increased cognitive effort for some participants (between 1/3 and 1/6 of

clinicians in this experiment) and did not enhance error detection for 57% of them, sometimes causing over-reliance. Despite this, XAI did lead clinicians to spend longer making their decisions when they disagreed with the AI recommendation and in ambiguous cases.

Amann et al. analysed transcripts from four workshops with a range of stakeholders (including the developers of an in hopsital cardiac arrest prediction algorithm for the emergency department, clinicians, social scientists and computer scientists) [95]. The authors concluded that the added value of XAI-CDSS relied on several factors which will vary with any given use case. These include validation level of the XAI, time-constraints of the decision to be made, technical feasibility, clinical audience and the overall clinical workflow within which the XAI-CDSS sits. This recognition that the exact setting and use case are critical highlight that we must be careful about extending lessons from other clinical contexts to the ICU.

1.5 Regulation of healthcare AI

1.5.1 Current regulatory landscape

The high-risk nature of healthcare makes it a sector that is already heavily regulated [96]. For example, bringing a drug from inception to market typically takes 5-20 years and costs billions of dollars, with an estimated $120 million saved if a drug trial is shortened by a mere four months [97]. After approval, there are also stringent ongoing requirements for post-market surveillance. Similarly, the certification process for healthcare staff, such as doctors and nurses, involves years of study, formal degrees, post-graduate qualifications, and a requirement to document evidence of continuing professional development. Therefore, it is unsurprising that there has been little resistance to the principle of regulating healthcare AI [98]. However, achieving consensus on how to regulate AI has been more challenging than regulating therapeutic drugs

[99, 100]. Factors that make AI regulation more difficult include: (i) the rapid evolution of algorithms from one version to the next [101], (ii) lack of uniform standards and less developed regulatory practices [102], (iii) unclear transparency regarding mechanisms of action [103], (iv) data dependency, with private datasets restricting access for reproducing study findings [104] and (v) less developed outcome metrics, especially in cases where there is no pre-existing gold standard (i.e. sepsis management).

Transparency has been identified as a key attribute for reassuring regulators about the safety and efficacy of AI-driven decision support tools [101]. However, its presence is by no means guaranteed in the algorithms that have undergone the regulatory process so far. For instance, a study by McNamara *et al.* highlights the need for greater transparency in the clinical integration of FDA-cleared AI devices used for interpreting medical images [105]. The authors recommended training clinician users on the nature of the system's outputs, intended benefits and risks to prevent inadvertent misuse. However, of the 104 products reviewed, around 60% had no explainability (the majority of the remainder, around 40% of the total 104 studies, used localisation techniques such as bounding boxes or heatmaps) [105].

1.5.2 Requirements for interpretability

There has consequently been a push by regulators and related stakeholders to mandate the inclusion of explainability. The Artificial Intelligence Act from the European Union is perhaps the largest and most far-reaching of these efforts to date. It was initially proposed in 2021 and comes into effect in August 2024. The Act has various risk categories with healthcare falling into the high-risk category [106]. The transparency obligations state that *"AI systems must be designed to be understandable to users and stakeholders, ensuring they can comprehend how the system functions and makes decisions"* [107]. Following on from this (and facilitated by transparency) is the requirement for high-risk AI systems to allow for human oversight to intervene and

prevent harm or incorrect decisions. Detailed technical documentation explaining the AI system's development, deployment and functioning must also be available with clear information provided to end users [107]. Specifically, end users have a right to *"know how and why decisions are made"* [107]. The implications of this for the healthcare AI sector are significant as the requirement for transparency means that developers can not rely solely on demonstrating superior performance for a black box algorithm. While this might be 'safe', it would provide no recourse for end-users (clinicians and patients) to satisfy their right to *"know how and why decisions are made"*.

Overall, the transparency requirements in the EU AI Act offer several advantages, such as increasing trust by making AI systems more understandable and thereby increasing the salience of fostering user confidence and ethical compliance [108, 109]. They also theoretically improve accountability given that the transparency makes it clearer if the fault arises from, for example, the algorithm or the user who has not adhered correctly to the intended use [109, 110]. However, these requirements also present challenges including increased costs and complexity due to the need for extensive documentation and compliance measures which can reduce the pace of innovation [109]. The technical mission of achieving explainability, as discussed earlier, is non-trivial and can often (though this is disputed by some [111]) come at the cost of a trade-off with performance [108, 112]. Of critical importance to this PhD, the presence of explanations could actually worsen performance - there is a risk that end users may either misinterpret explanations or interact with them in a manner that increases automation bias or algorithmic aversion [113, 114]. Indeed, a study of non-clinicians (psychology students) performing a medically themed classification exercise found that they were influenced by a biased AI system. When later performing the task without AI assistance, they repeated the AI's errors, demonstrating that AI bias can be 'inherited' by human users [115].

Other international regulators including the US Food and Drug Administration (FDA), the UK Medicines and Healthcare products Regulatory Agency (MHRA) and the EU's European Medicines Agency (EMA) have also published guidance on the topic of regulation and the role of explainability therein [116–118]. There has also been a push to increase the presence of reporting guidance to increase reproducibility in the field of both AI-driven decision support studies more generally and transparency specifically [104, 119]. For example Chen and colleagues performed a review of XAI techniques in medical imaging and found that computational feasibility was frequently prioritised over user needs and that there was a general lack of formative user research and empirical validation [119]. They addressed this by creating the INTRPRT guideline to advocate for human-centred design principles. The INTRPRT guidelines compose six statements which recommend (i) specifying the clinical scenario, constraints, requirements, and end users, (ii) justifying the choice of transparency and determining the level of evidence required, (iii) clarifying how the model adheres to the transparency justification, (iv) determining effective communication methods with end users, (v) reporting the task performance and finally (vi) assessing the correctness and human factors of explainability [119].

1.6 Is explainability essential for building an artificial moral agent in healthcare?

A vision of a world in which artificial agents interact freely with human society has given rise to the field of machine ethics. This field focuses on developing ethical frameworks and mechanisms for AI agents to be capable of engaging in moral behaviour [120]. In this section, I tackle the issue of whether explainability is essential to building an artificial moral agent in healthcare, starting by examining why morality plays an

important role in healthcare and therefore why we might want a moral healthcare AI agent at all.

1.6.1 The necessity of morality in healthcare

First, we need to ask whether morality itself is necessary for agents in healthcare. Morality, defined for these purposes as a set of principles governing whether an action is right or wrong, is generally felt to be a vital component of healthcare. However, while not disputing this, Van Wynsberghe and Robbins argue that just because a technology exists within a healthcare environment for example, it does not necessarily follow that such an agent needs to reason ethically [121]. I would counter their argument by looking to human agents within healthcare and assessing the consequences when there is a clear departure from moral behaviour. The most egregious examples in the 20th century include human experimentation by the Nazi regime on concentration camp prisoners during the second world war [122]. Other disturbing experiments from the 20th century include the Tuskegee Syphilis study in the US in which African American men deliberately had antibiotic treatment withheld so as to observe the natural course of the disease [123]. The deception was such that the study lasted 40 years until being shut down in 1972 with the death of 128 participants overall. These examples sit atop many others that have shocked the medical community and reiterated the importance of having ethical frameworks and rigorous national enforcement structures in place to ensure oversight and protection of patients and the public (e.g. the Helsinki Declaration) [124].

However, this is the sharp end of the morality spectrum and represents the easy cases in which there is near universal consensus that the actions are wrong. What about more divisive healthcare scenarios? Should all patients have universal access to healthcare at the point of need such as in the UK? Should we focus on expensive novel cancer therapies that can prolong life for a few months or on water sanitisation

projects in developing countries that could benefit thousands? These may seem like high level questions that are directed towards governments and organisations rather than individual healthcare agents on the ground. So let us take a more concrete example. During the COVID-19 pandemic, many healthcare systems came close to being overwhelmed as massive demand outstripped supply [24]. In the UK, this necessitated diluting the usual standards of care so as to treat more patients. In a busy hospital, should a healthcare AI agent admit an independently living 90 year old to the intensive care unit knowing that the likelihood of survival is small or should they prioritise maintaining a higher nursing ratio for the existing patients?

A moral compass or 'north star' set of guiding principles is important for all healthcare staff and this will include AI agents even if operationalising cases in the grey zones is not clear cut. Clinicians in the UK are bound to principles outlined by our professional regulator, the General Medical Council (GMC), in a document known as 'Good Medical Practice' [125]. This guidance covers ethical principles which help to safeguard the best interests of patients. It seems reasonable to conclude that patients and the wider public will expect AI agents to be held to a similar moral standard. While it is true that many AI agents currently operate in a very limited advisory role, this is not likely to persist. We need to confront both the why and the how of moral AI agents well in advance of their widespread implementation, not after they are mainstream.

To AI researchers, transparency suggests that the model is in and of itself something that can be followed through from start to finish by a human. However, to patients transparency tends to mean openness i.e. not withholding information or being deceitful (see section 1.7). The ethical imperative for patients is often the intention behind a specific action rather than the exact mechanics of how information flows through an algorithm before an action is proposed. This distinction is emblematic of why we should base our discussion of explainability on how it applies to patients and clinicians, rather

than purely AI developers. The definition of explainability that we cited earlier by Arrieta et al. satisfies several criteria suggesting that explainability is indeed essential for a moral AI agent [54]. It talks of understanding, it takes into account the specific audience (in this case clinicians and patients) and it also places emphasis on ensuring things are clear or open and thus touches on the more patient centred definition of transparency.

1.6.2 Explainability and the four ethical pillars

Amann considers the imperative for explainability in a moral AI agent on the basis of the four key ethical pillars of healthcare: autonomy, justice, beneficence and non-malfeasance [126]. This seems a sensible approach and I will take each of these four principles in turn to discuss why explainability is a core tenet for moral healthcare delivery.

Autonomy relates to patients having informed consent and the ability to weigh up the risks and benefits of various management options before they decide, in conjunction with their clinician, what to do about their health. This is popularised in the maxim "no decision about me, without me" [127]. Here, the importance of explainability is in providing the context and rationale for a moral AI agent's reasoning such that it can be incorporated with other sources of information as well as the patient's own value preferences. For example, if an AI recommends an operation for a patient, why has it made this decision? Is it to maximise survival at the potential expense of quality of life? What does our patient place the greatest value on? An explainable AI will be able to provide them with the information they need, in the form they need (e.g. patient friendly language that avoids medical or technical jargon) so that the patient can exercise their autonomy. The clinician is also exercising their professional autonomy when they make decisions that they are medico-legally and ethically liable for (and for which they may well be called to defend themselves, potentially even in a court of law).

They too need to understand the rationale and motivations of an AI agent as well as some idea of performance.

London makes the insightful point that although explainability is valued, there will likely also be a subset of users (within both clinicians and patients) that may prioritise outcomes and performance above explainability [128]. Cutillo et al. on the other hand emphasise that one of the reasons explainability is important at this early stage in the development of AI moral agents is that we don't yet have a longstanding history of trust or performance and that, if anything, there is usually aversion to trusting machines in moral situations [129, 130]. They give the example in healthcare of trusting medications for which we do not fully understand the mechanism but for which we have very strong evidence to suggest good performance. This may come over time for AI agents but may not suffice early on in their use.

The second medico-ethical principle is justice and this describes the fair distribution of healthcare in a manner that avoids discrimination. As mentioned earlier, even human societies and organisations continue to debate the best ways to allocate resources in marginal cases [131]. However, there is an ethical duty on human clinicians to minimise discrimination where possible. This should also be required of a moral AI agent. There has been much press attention devoted to situations of potential AI discrimination as agents learn from biased human data [132]. Let us suppose that an AI recommends different treatment options for two patients, one of whom is white and one of whom is not. If challenged, how can the AI agent prove that it has recommended the treatment because it is the most effective clinically for each patient rather than on the basis of confounding or discriminatory intent. It must be able to justify its actions and this benefits from explainability.

The third medico-ethical principle is non-malfeasance. This places the onus on clinicians to avoid harming their patients. This is not limited to intentional direct harm but also includes over-diagnosis and overuse of medical interventions with corresponding

downstream iatrogenic harm. An AI agent could avoid missing all cancers by scanning every patient but this would expose many to unnecessary radiation risks which could itself lead to more cancer than is prevented overall. The rationale for AI generated recommendations should include both the intention and the associated confidence intervals for the uncertainty of the proposed action. This enables an approach that acknowledges harm that can arise from actions with higher uncertainties. A further point is that patients also need to have recourse to complain and seek legal remedy if they are wronged [133]. Such a case would be difficult to prosecute with an opaque system.

The final principle is beneficence which means acting to maximise the best interests of the patient. This will require discussion with the patient to ascertain what those best interests are and then to explain how the actions recommended by a moral AI agent help to make those interests a reality. Clinicians may have access to data that the AI agent does not and this emphasises why integration of data from multiple agents is critical. This will not be easy without explainability on the part of the AI. We also need to think about situations in which the AI agent and clinician may disagree on the best course of action for the patient. Carter makes the point that explainable AI may also have a role when it comes to such disagreements [134]. Core to any constructive discussion or debate is the ability to clearly explain a logical rationale with supporting evidence and this mandates some degree of explainability.

1.6.3 Building trust and accountability in AI

We can use the analogy of medical research studies to reinforce why explainability is crucial for a moral AI agent in healthcare. Initially we study a new treatment against a placebo. However, once a new treatment is proven, future treatments need to be compared against the new standard of care. To expose patients to anything less is regarded as unethical. So too with moral AI agents in healthcare, the minimum

standard they need to achieve is to be on par with clinicians. Human clinicians act in an ethical manner by following the four key principles of autonomy, justice, non-malfeasance and beneficence. All four require some degree of explainability from the clinician and there is therefore a strong argument that we should demand no less from AI agents moving forward.

1.7 Patient and public involvement (PPI)

1.7.1 Rationale for PPI

The AI Clinician is focused on an area of healthcare in which direct patient input (e.g. on drug choices or doses) is rarely sought. Often patients may be sedated and unconscious and the precise dosing details require a detailed understanding of many aspects of physiology and clinical medicine. This is in contrast to, for example, an RL algorithm that might be used in a clinic to help make a decision on a chemotherapy regime where there are varying side effect profiles that a patient might have very strong opinions on, even to the extent of choosing a less effective therapy with a better side effect profile [135]. Nonetheless, there is an increasing focus in recent years on involving patients and the wider public in AI related health endeavours as a means of building trust and ensuring that technology stays patient focused [136, 137]. I therefore designed and conducted a PPI session with two other ICU clinicians who were also investigating sepsis data and decision support tools.

1.7.2 PPI methodology

An initial research application was submitted to the patient charity ICUsteps (see appendix B) which was assessed by them for lay readability before being forwarded to their volunteer network. From this network, three former ICU patients were recruited for a formal PPI session and given a more detailed participant information sheet (see

appendix B). The information sheet covered three main AI related topics: (i) how they would feel about an AI algorithm suggesting treatments in the ICU, (ii) what sort of explanations might be helpful to them or their clinician and (iii) how they would feel if an AI recommendation turned out to be a mistake. The mix of themes was chosen to maximise the efficiency of the session. A group discussion with the three former patients lasting approximately 45 minutes was held on each of the topics via Zoom. Each session was led by one ICU clinician with the others taking notes before rotating. The participants were compensated for their time.

1.7.3 Insights from PPI

The full transcript for the PPI session is available in appendix B. A brief summary for each of the three main topics follows below:

- **Views generally on AI decision support tools**

  Participants were generally positive about the introduction of AI based decision support tools but felt that rigorous testing prior to implementation was definitely required. They would trust doctors to use the AI appropriately. When asked about incorporating patient wishes into decision support tools, participants were keen for doctors to be the main link between patients and the AI. They felt that clinicians needed to be able to interpret whether an AI suggestion was the right action for a patient.

- **Views on AI explainability**

  One participant felt that an explanation was not necessary; if the clinician was happy to use the device then they would trust the clinician. Another participant felt an explanation was important, while the third participant felt that the explanation should be very simple and straightforward. When asked about situations where the AI system is so complex that explanations are difficult to

obtain but that it is nonetheless effective at maximising longer term patient centred-outcomes, the participants generally felt that lack of explanation was okay but only if the efficacy was proven.

- **Views on AI mistakes/errors**

  When asked about how much uncertainty in treatment decisions they would like to be aware of if an AI decision support tool was being used, the general feeling was that participants would trust the doctor to make the right call in terms of how much detail to give the patient. One remarked that they are not typically consulted on every minor decision in any case so it wouldn't be that different to the current situation. Regarding errors, participants felt transparency was critical and especially that the learning of lessons to prevent similar mistakes should be shared widely. One participant worried that introduction of AI may decrease the volume of teaching and training provided to more junior staff.

Overall, the PPI work seemed to confirm no gross objections from the former ICU patients to the use of AI based decision support tools in the ICU. AI technologies outside of healthcare are increasingly prevalent and this may have contributed to patients being both more familiar and cognisant of the potential benefits that AI brings. However, the participants did stress that they expected rigorous testing and evaluation to take place before implementation, in a similar vein to the lengthy regulatory process applied to new pharmaceutical agents. It was interesting to hear fairly open views on the impact of potential AI mistakes and the effect on patient trust. The key factors the participants mentioned were (i) transparency and (ii) learning from mistakes. This aligns closely with recent regulatory guidance from the General Medical Council regarding duty of candour. There is now a statutory duty on clinicians and organisations to disclose mistakes to patients in a timely and transparent manner, with an apology provided and advice on how the mistake is to be investigated and prevented from recurring. Once AI-driven decision support tools are used in everyday clinical practice, the scope of

duty of candour will extend to cover such tools and explanations from AI suggestions may then be increasingly utilised from a medico-legal standpoint. Critical to this thesis, the view of patients on explanations was relatively mute. The participants felt that explanations could be useful but clearly recognised that the context of niche treatment decisions for critically ill patients with sepsis would not ordinarily be something they or their families would expect to be consulted on.

1.8 Research questions and goals in this thesis

This thesis aims to address several key questions regarding the implementation and evaluation of XAI in clinical decision support systems with a particular focus on the high-stakes clinical setting of intensive care. The overarching goal is to understand how XAI influences clinical decision-making and to develop more effective methods for evaluating its impact. The primary research questions are as follows:

1. **How does the presence of AI and XAI affect clinical decision-making in sepsis management?** This question aims to quantify the impact of AI and XAI on prescription decisions for fluid and vasopressors in sepsis (by, for example, comparing baseline, peer, AI, and XAI conditions across various experimental settings).

2. **Can novel methodologies, particularly objective behavioural markers such as eye-tracking provide insights into physician interaction with XAI systems?** This question seeks to validate eye-tracking as a methodology for evaluating attention and cognitive load in XAI interaction and to explore its potential in providing deeper understanding of the decision-making process beyond traditional self-report measures.

3. **How do different types of XAI modalities and the fidelity of the experimental environment impact clinical decision-making and attention patterns?** This

question explores the effectiveness of various XAI approaches ranging from simple feature importance to more dynamic and complex approaches. It also considers whether the experimental setting (from low-fidelity vignettes to high-fidelity simulations) is an important modifier of human-AI interaction behaviour.

By addressing these questions, this thesis aims to contribute to the development of more effective, physician-focused XAI systems for clinical decision support, as well as to advance the experimental methodologies for evaluating such systems in healthcare settings.

Chapter 2

Quantifying the impact of XAI recommendations on prescription decision making

This chapter is closely based on a manuscript published in 2023, on which I am the first author and my supervisors are co-authors:

Quantifying the impact of AI recommendations with explanations on prescription decision making, published in npj Digital Medicine [138].

2.1 Introduction

The previous chapter (see section 1.4) discussed the existing literature on how physicians can be influenced in their decision-making (by both peer humans and AI) but does not assess this either in an ICU setting or for a prescription task that has no 'gold standard' correct answer. Successful deployment of XAI-driven clinical decision support tools is as much a problem of clinicians' influenceability, cognition and psychology as one of machine learning alone [139–141]. I addressed these issues in this study by assessing how prescription decisions by ICU physicians may be influenced by additional

information more broadly, and how this influence can be modified by either the source of the information (human peers or AI) and the presence or absence of an AI explanation (here using simple feature importance). Furthermore, I investigated the extent to which self-reports of XAI usefulness actually correlate with the impact of XAI on prescription decision-making.

2.2 Methods

2.2.1 Data source and AI clinical decision support system

The 'AI Clinician' is an ICU clinical decision support system utilising reinforcement learning to provide semi-autonomous, continuous dosage recommendations for IV fluids and vasopressors [45]. This system was trained on data from 17,083 ICU patients sourced from the MIMIC-III database, as previously described [45]. MIMIC-III is an anonymised, open-access database containing information on over 60,000 ICU admissions from six teaching hospital ICUs in the USA, spanning 2001–2012 [17]. In brief, the patients selected for the AI Clinician's training were adults diagnosed with sepsis according to sepsis-3 criteria [142]. The patient data were divided into 4-hour intervals. For each interval across all 17,083 patients, the AI Clinician categorised the patient into one of 750 states and recommended a dose for IV fluids and norepinephrine/noradrenaline, the most frequently used vasopressor in septic shock [2].

Twenty-four patient scenarios were selected for my experiment. Twelve of these were 'expert-selected' to ensure representation from four broad categories: (i) three patients where both the fluid and vasopressor AI dose suggestions were similar to what human clinicians had done in MIMIC-III, (ii) three patients where only the AI vasopressor suggestion was similar to humans, (iii) three patients where only the AI fluid suggestion was similar to humans, and (iv) three patients where neither AI fluid nor vasopressor suggestions were similar to humans. These 12 patients also covered

scenarios where the patient was receiving anywhere from no vasopressor to a large dose (>0.5 mcg/kg/min of noradrenaline-equivalent) to ensure a representative patient mix. The remaining 12 patients were chosen by clustering the entire MIMIC-III sepsis dataset of 17,083 patients into 12 clusters and then selecting a patient within the closest percentile to the cluster centroid. This yielded 12 patients who were less sick (as defined by the proportion on vasopressor support and APACHE score) than the 'expert-selected' 12 but that were more representative of the MIMIC-III septic cohort. The amount of fluid and vasopressor support is shown in Figure 2.1, stratified by whether the patient was 'expert-selected' or 'cluster-derived'.

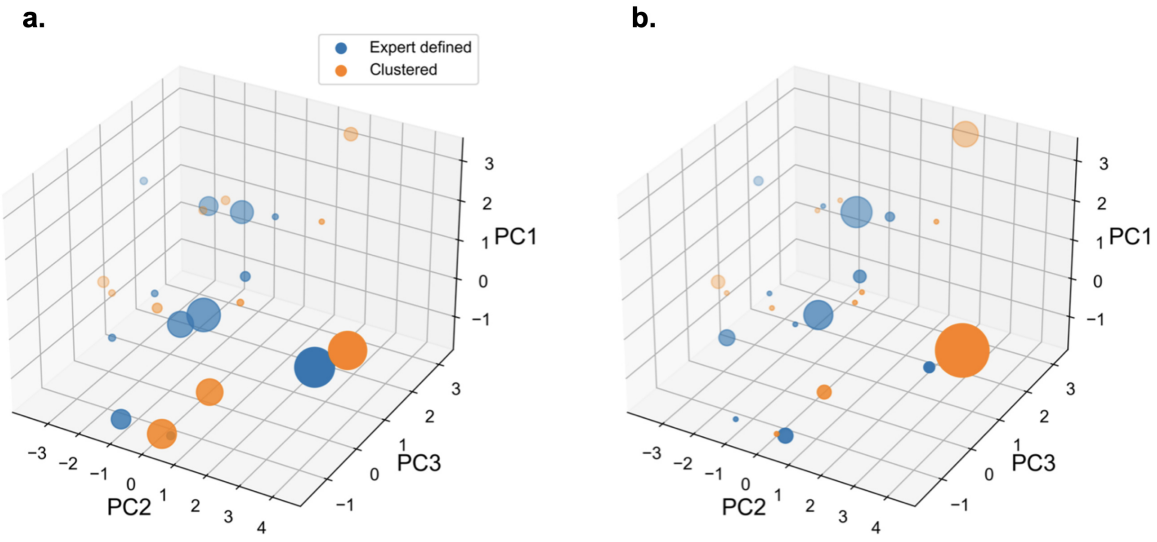


Fig. 2.1 **Coverage of patient scenarios over the MIMIC-III state space.** The patient scenarios are plotted below on a 3-component PCA model of MIMIC with dot size reflecting the dose of fluid in (a) or vasopressor in (b), where larger size means higher dose.

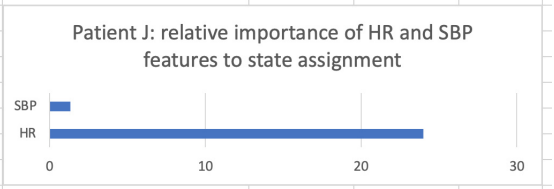
Feature importance explanation

When designing a simple, local feature based explanation mechanism, I focused on the importance of state membership which occurs prior to RL. The state space for the AI Clinician was constructed using a k-means clustering algorithm. After the algorithm

converges, the cluster centroids represent the average feature values for patients in that state/cluster. A new patient would be assigned to the state/cluster that minimises the distance from their feature values to the respective cluster centroid. Intuitively, with over 40 features, some features will be closer to the cluster centroid value than others for any patient assigned to a given state. I exploit this fact to rank features in terms of their proximity to the cluster centroid (or average state feature values) because the archetypal patient for whom an RL agent policy action most applies is a patient who is most typical of that state.

To take a toy example, see Figure 2.2. If the average heart rate (HR) and systolic blood pressure (SBP) for patients in state 3 is 120 and 60 respectively, consider two patients J and K who have both been assigned to the state 3 cluster. J has HR 115 and SBP 105 while K has HR 75 and SBP 55. The importance of HR for patient J being assigned to state 3 is higher than SBP. The reverse is true for patient K for whom their SBP is more important to them being assigned to state 3. Figure 2.2 shows this presented numerically with some simple mathematical calculations to convert this difference between state average value and patient value to a value on a scale from 0 to infinity that can be plotted graphically.

Patient J		
Variable	**HR**	**SBP**
State 3 average	120	60
Patient J	115	105
X = Patient J value / state average value	0.96	1.75
1 / (1 - X)	24.00	-1.33
Absolute value	**24.00**	**1.33**



Patient K		
Variable	**HR**	**SBP**
State 3 average	120	60
Patient K	75	55
X = Patient K value / state average value	0.63	0.92
1 / (1 - X)	2.67	12.00
Absolute value	**2.67**	**12.00**

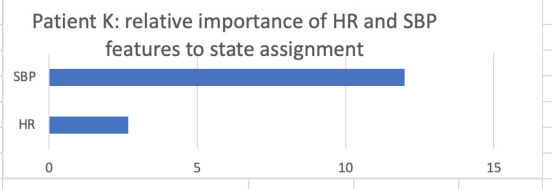


Fig. 2.2 **Feature importance for a toy example with two patients.** Patients J and K are both clustered to state 3. HR is more important than SBP for patient J being assigned to state 3 while SBP is more important than HR for patient K. The numeric scale is arbitrary - it is the relative difference between features that is of interest.

It is the relative difference between variables that is of interest. For example, patient J's HR has a higher arbitrary value than SBP and vice versa for patient K. This toy example ignores the fact features are on different scales with, for example, lactate typically varying from 0 to 20 while hourly urine output might vary from 0 to 500. Therefore all variables had standardised scaling applied before this process was undertaken. The result of taking this approach for real patients from the AI Clinician can be seen in Figure 2.3. This explanation does not answer the question of *why* the RL agent decided on a specific dose for patients in a specific state. It does however, give some idea of *what* was important in being assigned to that state. And we know that state membership is a critical precursor to the RL agent effectively looking up the optimal action it wants to pursue from its learned policy constructed during the training phase.

I considered using the full list of features as displayed in Fig. 2.3 as the explanation but worried that the number of features might overload participants and increase the amount of time spent on each scenario in the experiment. I therefore opted for an even simpler version in the pilot experiment where clinicians were simply given a list of the top 5 features in rank order as their explanation.

2.2.2 Vignette experiment and conditions

I conducted a human-AI interaction vignette experiment for doctors using a modified between-subjects design. The vignette design was chosen as it enables high standardisation of scenarios and easy recruitment of physicians (as participants do not necessarily have to physically attend to take part in the experiment). There were four experimental arms. In each arm, physicians were given patient data presented as a fixed variables table (e.g. age, gender, weight), an interactive graph displaying a limited set of time-varying features and a second larger table showing all time-varying features (see appendix C for screenshots). This setup was designed to resemble how

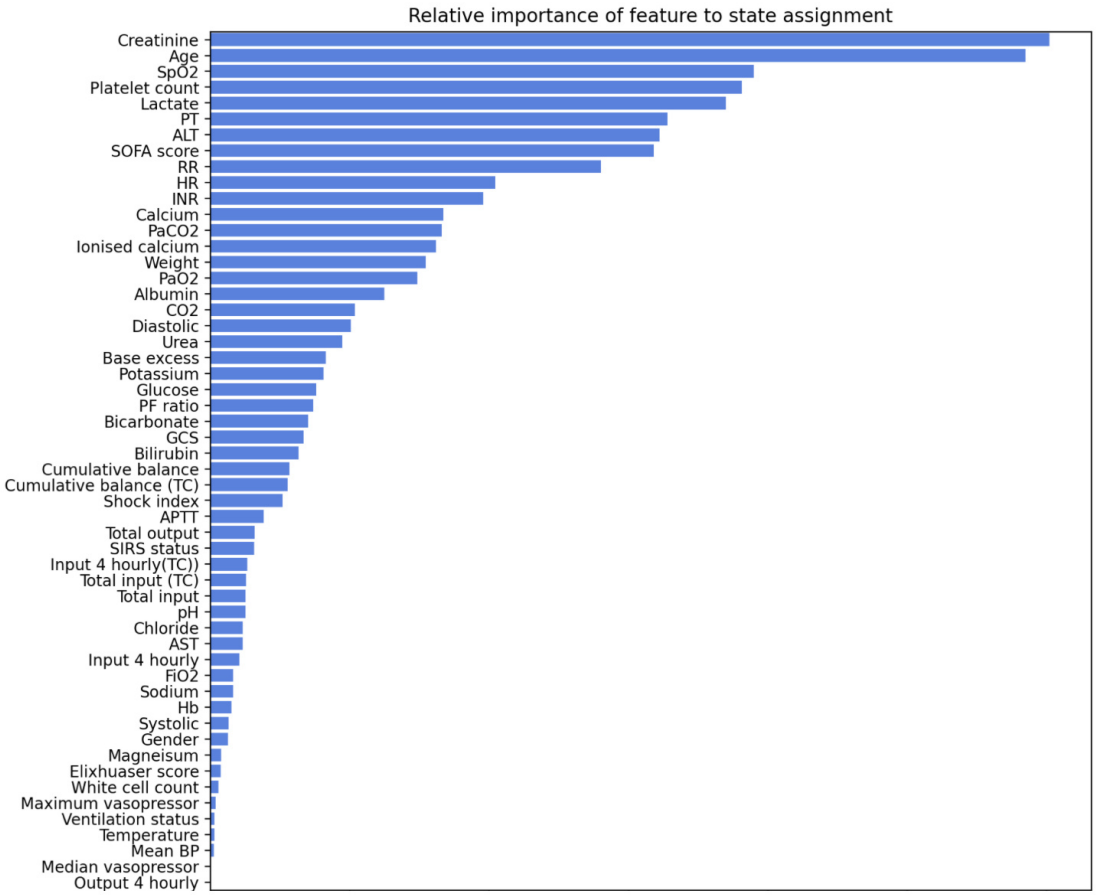


Fig. 2.3 **Feature importance for an example patient at time block 1.** The x-axis scale is arbitrary - it is the relative difference between features that is of interest.

most ICU doctors in the UK access patient data in their respective electronic health records (EHRs).

Each doctor performed 16 trials. The initial four trials were the same for all participants and served as a pre-training period. The subsequent 12 trials constituted the main experiment. In each trial, doctors were asked to select a dose for fluid and a dose for vasopressor to be administered over the next hour. I used a multi-factorial experimental design with four arms, wherein each doctor encountered all four arms across different subsets of the 24 patients. The four arms were: baseline with no additional AI or peer human information (baseline); additional peer human clinical information (peer, see description below); additional AI decision support system

information (AI); and additional AI decision support system with an explanation of the AI decision (feature importance, XAI). Examples of all four scenarios can be seen in the screenshots provided in appendix C.

In the baseline scenario, doctors viewed only the patient data. In the peer human clinician scenario, doctors were also shown the probability density function of IV fluid and vasopressor doses prescribed by other physicians in the MIMIC-III dataset for patients in the same state. This acted as a proxy for what peer clinicians had previously done for similar patients. These data were presented as a violin plot (a conventional box plot overlaid with a distribution derived via kernel density estimation). The rationale for including peer data as an experimental arm was to determine if doctors simply desire additional information or context to support their decisions (regardless of its source), or if there is something particular about an AI suggestion that is more or less persuasive than simply knowing what their peers typically do.

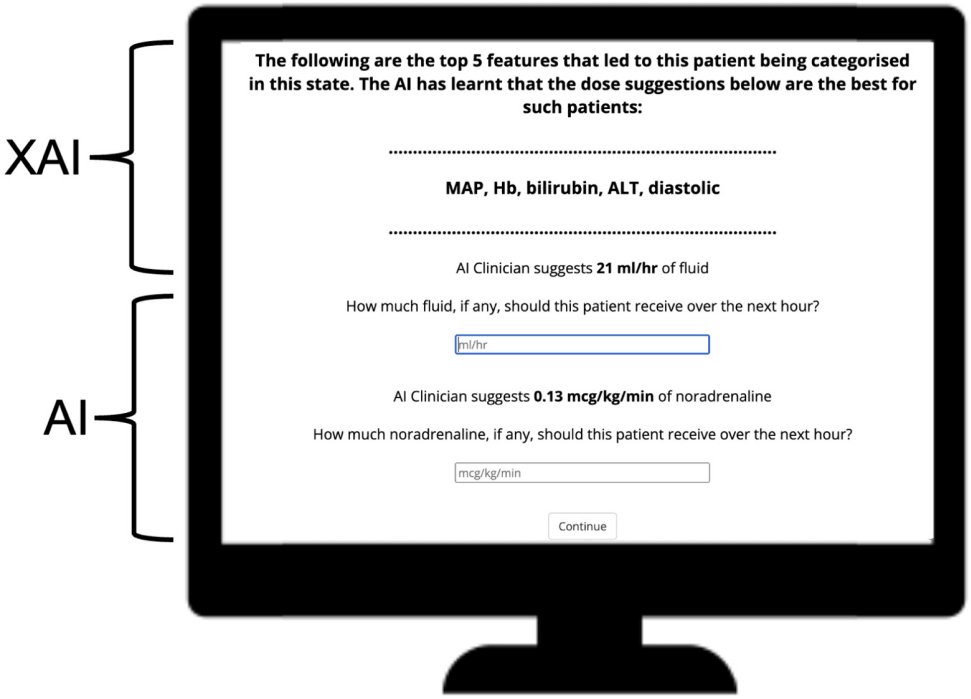


Fig. 2.4 **AI and XAI screens.** The AI and XAI components of the decision screen where participants input their prescription choice can be seen.

In the AI scenario, subjects were also shown the AI Clinician's suggested doses for fluid and vasopressor (in text form), see Figure 2.4. In the XAI scenario, subjects were shown the AI Clinician's suggested doses along with a feature importance based explanation, see Figure 2.4. Specifically, they were shown the top five ranked features contributing to the state assignment. Although feature importance might be considered a basic form of XAI, it is nonetheless widely used in medical studies [45, 143, 144, 42, 145, 146].

		REGION 1				REGION 2											
Subjects	Group	BA	PE	AI	XAI	BA	PE	AI	XAI	BA	PE	AI	XAI	BA	PE	AI	XAI
7	I	1	2	3	4	5	8	10	13	6	9	12	15	7	11	14	16
7	II	1	2	3	4	11	5	6	7	12	14	8	9	13	15	16	10
7	III	1	2	3	4	7	11	14	16	6	9	12	15	5	8	10	13
7	IV	1	2	3	4	13	15	16	10	12	14	8	9	11	5	6	7
6	V	1	2	3	4	10	7	5	14	9	6	15	12	8	16	13	11
6	VI	1	2	3	4	16	10	9	8	15	13	7	6	14	12	11	5
5	VII	1	2	3	4	8	16	13	11	9	6	15	12	10	7	5	14
5	VIII	1	2	3	4	14	12	11	5	15	13	7	6	16	10	9	8
Subjects	Group	BA	PE	AI	XAI	BA	PE	AI	XAI	BA	PE	AI	XAI	BA	PE	AI	XAI
5	IX	1	2	3	4	17	20	22	25	18	21	24	27	19	23	26	28
5	X	1	2	3	4	23	17	18	19	24	26	20	21	25	27	28	22
5	XI	1	2	3	4	19	23	26	28	18	21	24	27	17	20	22	25
5	XII	1	2	3	4	25	27	28	22	24	26	20	21	23	17	18	19
4	XIII	1	2	3	4	22	19	17	26	21	18	27	24	20	28	25	23
4	XIV	1	2	3	4	28	22	21	20	27	25	19	18	26	24	23	17
4	XV	1	2	3	4	20	28	25	23	21	18	27	24	22	19	17	26
4	XVI	1	2	3	4	26	24	23	17	27	25	19	18	28	22	21	20

Fig. 2.5 **Trial matrix for vignette experiment.** The numbers from 1 to 28 refer to patient scenarios. Patient scenarios 1 to 4 (region 1) were common to all participants and did not form part of the primary analysis (akin to a wash-in period). Region 2 was designed so that half the subjects saw a patient under one arm while the others encountered the same patient under a different arm, allowing estimation of between arm variability by controlling for the patient. The study was not randomised.

The trial matrix (see Figure 2.5) was designed to ensure that half of the doctors saw a patient under one arm while the other half encountered the same patient under a different arm, enabling the estimation of between-arm variability by controlling for the patient. The primary measure of interest was the difference in prescribed doses

to the same patient across the four different arms - effectively assessing the impact of each arm on clinical decisions by measuring the dose shift (Figure 2.6). The order of trials in the trial matrix was varied to counterbalance any learning effects. Statistical analyses included two-sided T-tests for comparing means (after confirming normality) and linear regression for assessing associations. Both analyses were performed without adjustment for multiple comparisons.

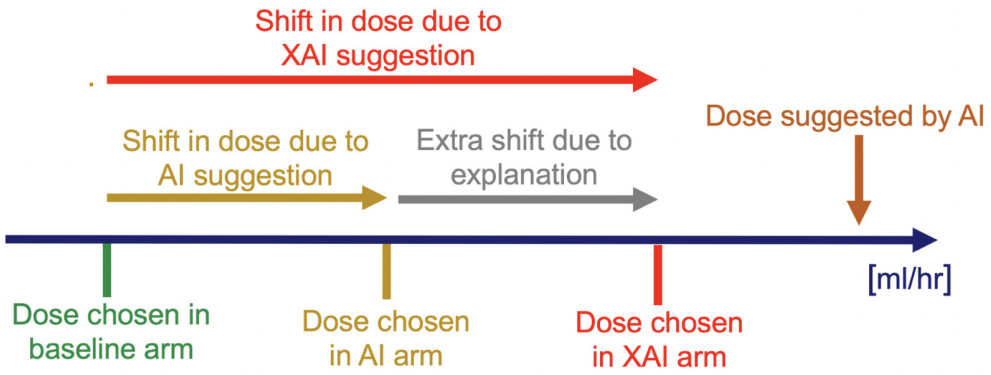


Fig. 2.6 **Example of dose shift metric.** This figure demonstrates the shift in dose relative to baseline occurring as a result of showing the AI suggestion. The extra shift between AI and XAI is the marginal shift attributable to the explanation.

2.2.3 Participant recruitment and experiment conduct

Physician demographics, experience, and affinity to AI were collected via a questionnaire before starting the main experiment (Figure 2.7). After the experiment, doctors completed a short post-experiment questionnaire (see appendix C). Data collected for each patient scenario included the physician's prescribed doses for fluid and vasopressor per patient scenario as well as the time taken per patient scenario.

A convenience sample of ICU physicians was recruited based on the following inclusion criteria: (i) practising doctor, (ii) has worked for at least 4 months in an adult ICU, (iii) currently works in ICU or has worked in ICU within the last 6 months. Physicians could participate remotely via Zoom or in person. Electronically

recorded informed consent was obtained from all physicians and each experiment lasted approximately 45 minutes. The study was approved by the Research Governance and Integrity Team (RGIT) at Imperial College London (ICREC reference 21IC7245). The institutional review board of the Massachusetts Institute of Technology (no. 0403000206) and Beth Israel Deaconess Medical Center (2001-P-001699/14) approved the use of MIMIC-III for research. Because this study made use of fully anonymised patient data, individual patient consent was not required.

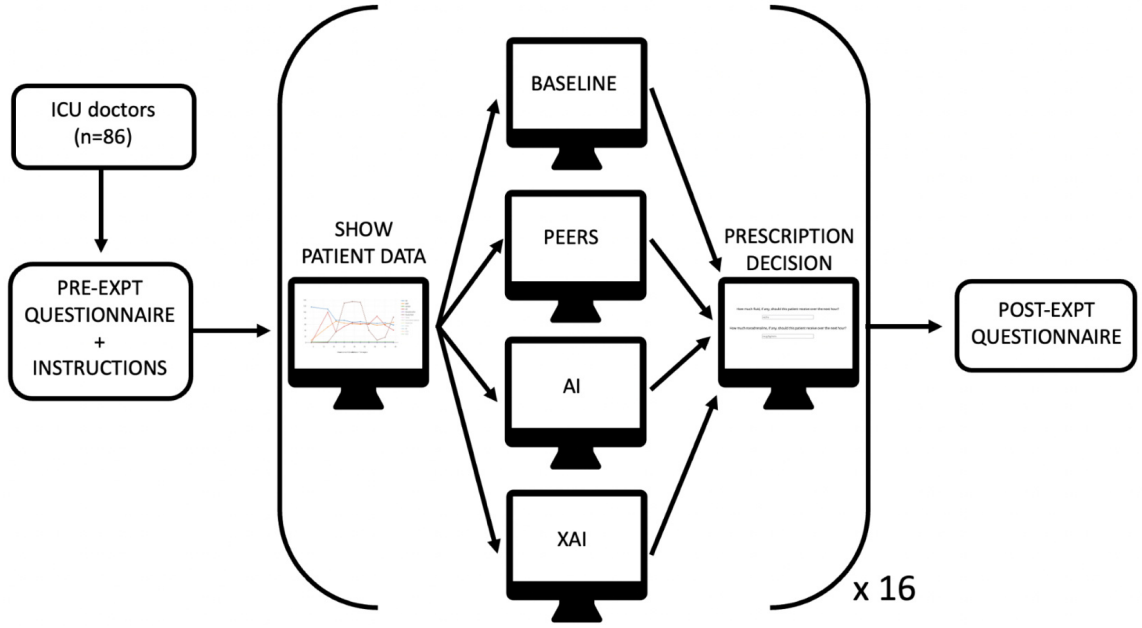


Fig. 2.7 **Experiment protocol.**

2.3 Results

Eighty-six ICU doctors were recruited, comprising 31 seniors (consultant/attending), 42 intermediate grade (registrar/fellow), and 13 juniors (senior house officer/resident). The median age of the participants was 37 years (interquartile range (IQR) 34–43), and the median years of clinical experience was 11 years (IQR 9–19). All subjects completed the task successfully with no significant difference in per-trial completion

time between the arms (baseline 83 seconds, peers 83 seconds, AI 79 seconds, XAI 83 seconds; p=0.574 by Kruskal–Wallis test). Practice variation was defined as the dose distance of an individual physician from the average of all physicians in the baseline/control arm for any given scenario.

2.3.1 Impact of arms on dosing shift

An example of prescription shift for an individual patient scenario is shown in Figure 2.8a. For the same patients in different arms, providing subjects with additional information from their respective arm led to an absolute prescription shift for fluid of 70 mL/hr (peers, standard deviation (SD) 86 mL/hr), 90 mL/hr (AI, SD 83 mL/hr), and 85 mL/hr (XAI, SD 60 mL/hr) relative to the baseline arm (p=0.872 for peers, p=0.002 for AI, p=0.007 for XAI, all by independent T-test). For vasopressor, the prescription shift was 0.04 mcg/kg/min (peers, SD 0.06 mcg/kg/min), 0.05 mcg/kg/min (AI, SD 0.09 mcg/kg/min), and 0.05 mcg/kg/min (XAI, SD 0.09 mcg/kg/min) relative to the baseline arm (p=0.201 for peers, p=0.010 for AI, p=0.002 for XAI, all by independent T-test). The aggregate prescription shifts are displayed in Figure 2.9. The individual patient scenario dosing shift figures for all 24 patients are shown in Figures 2.15, 2.16, 2.17.

2.3.2 Impact of arms on practice variation

Providing doctors with a recommendation (whether it be peer, AI, or XAI) had a consistent effect: inter-clinician dose variability was influenced by whether the recommendation was higher or lower than what subjects in the baseline arm did. Specifically, when the recommendation was higher than baseline, the prescriptions of doctors in the peer/AI/XAI arms exhibited greater variability across doctors; when the recommendation was lower than baseline, prescriptions were less variable across doctors. This effect is illustrated in Figure 2.8b.

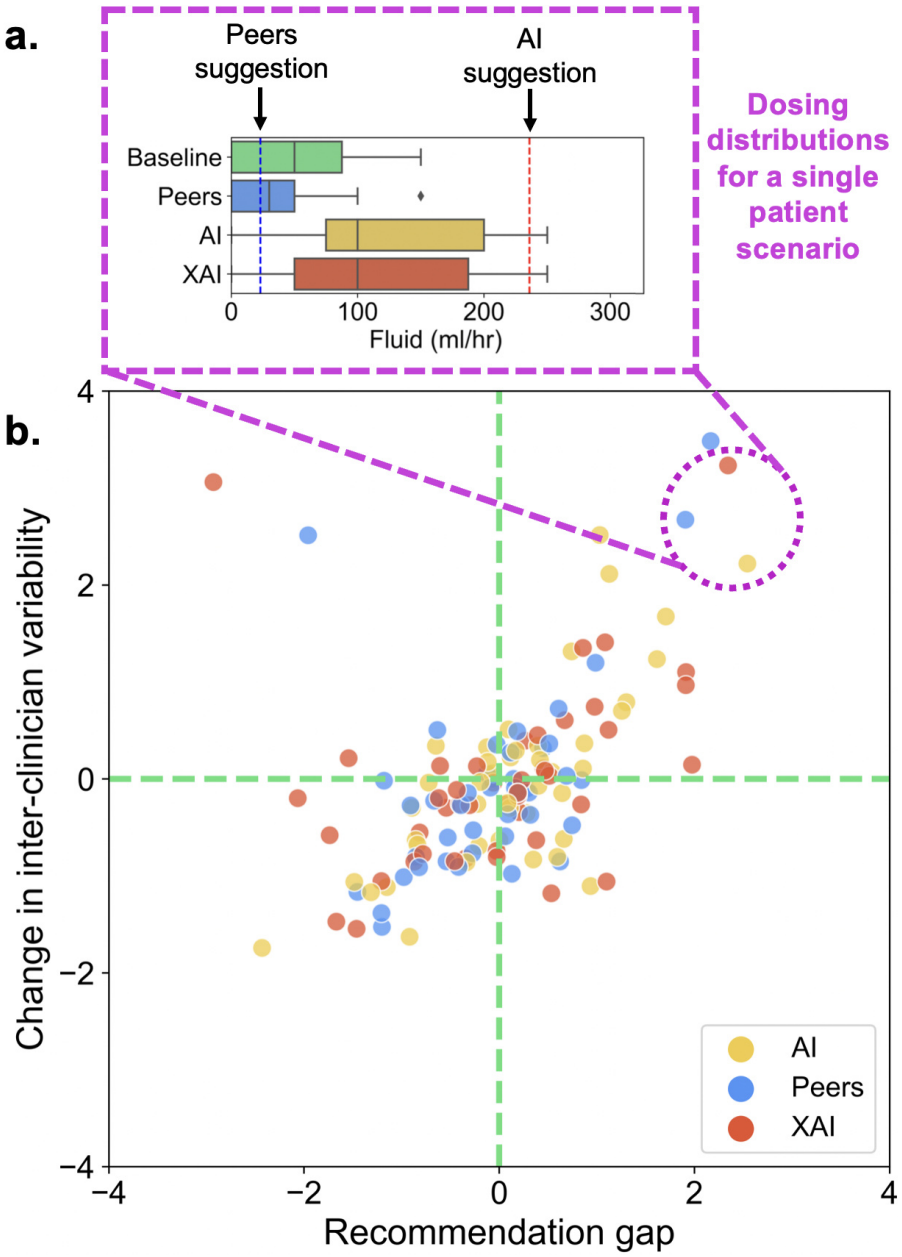


Fig. 2.8 **Dose variability by intervention arm.** In (a), the prescription distributions for a single patient scenario (and, for illustrative purposes only, mapped onto (b). For each boxplot, the centre line represents the median, box edges represent upper and lower quartiles, whiskers represent 1.5x inter-quartile range and diamonds are outliers. Blue dashed line represents the median of the peer distribution data (only available to those in the 'Peers' arm). Red dashed line represents the AI suggested dose (only available to those in the 'AI' or 'XAI' arms). In (b), change in inter-clinician variability by size of recommendation difference for peers/AI/XAI is shown (i.e. was the recommended dose higher (positive recommendation gap) or lower (negative recommendation gap) than the baseline average (dashed green line) and how does this affect variability of clinicians (x- and y-axes scales are arbitrary units, normalised to allow fluid and vasopressor to be plotted together.

2.3.3 Association of clinician factors with adherence to AI suggestions

Clinician attitude towards AI was derived as a principal component from the four pre-experiment AI enthusiasm questions posed to subjects (Figure 2.10a). The first component accounted for 69% of the variance (Figure 2.10b). Attitude towards AI did not show a significant linear association with the difference between physician-selected dose and AI-recommended dose for either fluid (see Figure 2.11a, r=-0.078, p=0.075 by LLSR (linear least-squares regression)) or vasopressor (see Figure 2.11b, r=-0.074, p=0.092 by LLSR). Similarly, years of clinical experience did not show a significant association with the difference between physician-selected dose and AI-recommended dose for either fluid (r=0.001, p=0.862 by LLSR) or vasopressor (r=-0.086, p=0.047 by LLSR), as illustrated in Figure 2.12a and Figure 2.12b.

2.3.4 Clinician opinions on AI and the explanations

Post experiment, physicians rated their likelihood of using an AI system for sepsis prescriptions on a scale from 1 to 5 with higher values indicating a greater likelihood of use. The mean score was 2.55 for training doctors (which encompasses both junior and intermediate doctors, SD 0.96) versus 2.16 for non-training doctors (senior/consultants, SD 1.07), p=0.091 by independent T-test (Figure 2.13a). Physicians were asked to rate the usefulness of the explanations on a scale from 1 to 5 (higher values indicating greater usefulness) with mean 2.22 for training doctors (SD 1.03) versus 1.97 for non-training doctors (SD 1.11), p=0.296 by independent T-test (Figure 2.13b). Subjects were also asked to rate the utility of displaying peer and AI suggestions together on a scale from 1 to 5 (higher values indicating greater utility) with mean 2.98 for training doctors (SD 0.73) versus 2.39 for non-training doctors (SD 1.09), p=0.003 by independent T-test (Figure 2.13c). Finally, subjects were asked to rate importance of evidence on a 1 to 5 scale (higher values indicating greater importance) for their use of an AI system with a mean rating of 3.01 (SD 0.85) for observational evidence versus mean 3.33 (SD 0.76) for

randomised clinical trial evidence, p=0.011 by independent T-test. The self-reported utility of explanations did not correlate with adherence to XAI suggestions (Figure 2.14).

2.4 Discussion

2.4.1 Summary of results

This study presents several key findings that enhance our understanding of how AI-driven decision support recommendations can influence prescription decisions. First, additional information (peer, AI, or XAI) strongly impacts prescription decisions as indicated by dose shift (statistically significantly for AI, not so for peers). However, the form of the recommendation - whether plain (AI alone) or accompanied by an explanation (XAI, here using simple feature importance) - did not substantially alter its impact. Second, inter-clinician dose variability was affected differentially depending on whether the recommendation (peer, AI, or XAI) was higher or lower than what the physicians in the baseline arm prescribed, suggesting that decision support systems might have a mixed impact on practice variation once deployed to a live setting. Third, there was no correlation between either attitude towards AI or clinical experience on the AI-supported decisions suggesting a certain unvarying degree of AI acceptance in clinical experts, or one moderated more by variability in patient scenario than the physician themselves. Fourth, there was no correlation between physician self-reported usefulness of XAI and whether XAI actually influenced their decisions. This calls into question the value of using XAI self-reports as an outcome metric in clinical XAI studies.

2.4.2 Limitations

These findings should be considered in the context of several limitations. First, the XAI condition used feature importance as the explanatory modality. While feature importance is commonly employed in XAI, both within and outside of healthcare [45, 143, 144, 42, 145, 146], it is also considered to be on the lower end of the spectrum regarding how comprehensive an explanation is according to cognitive science [57]. Indeed, the variability in XAI modality can impact medical decisions differently depending on the complexity or ambiguity of the context and the physician's experience with AI systems [147, 148]. Second, although I took steps to ensure a broad and representative range of patient scenarios, there may still be gaps in the patient state space. My two-pronged selection strategy was intended to make scenario choices as generalisable as possible (see Figure 2.1 for graphical representation of the MIMIC state space, based on three principal components for visualisation purposes).

Third, the patient scenarios were low fidelity due to the experimental vignette format. While this format allowed me to standardise scenarios, it lacks the dynamic aspect of evaluating a real patient over time and observing the dynamic effects of treatments. Fourth, although my sample of doctors was large compared to similar clinical studies, it relied on convenience sampling and may not be representative of the medical population as a whole. Fifth, the AI suggestion itself was presented as an isolated message containing the recommended doses for fluid and vasopressor without confidence bounds or ranges. Adherence to AI suggestions might be higher when presented with estimates of certainty, when presented graphically or using other algorithmic approaches [149].

Sixth, a detailed insight into the physicians' internal decision-making processes was lacking. The experiment produced data in the form of self-reports (which can be biased from many causes and in many directions) and final prescriptions (the end of the line in the decision-making process). Therefore, these conclusions are based on

inferring that the input conditions (baseline, peer, AI, XAI) had a direct impact on the final prescription decisions via some internal effect on the cognitive act of decision-making. To further explore this decision-making process, alternate methodologies such as qualitative interviews and 'think-aloud' studies, where verbalisation of decision-making is encouraged, could be employed. Consequently, the interpretation of the study findings and conclusions here are based on a data-driven framework of values rather than a more cognitive theoretical framework about how exactly the decision-making is influenced.

2.4.3 Implications and comparison with other studies

Despite the limitations acknowledged above, reviewing my findings alongside existing literature provides important insights into how we can improve the design and deployment of AI-driven clinical decision support tools. Attempts to quantitatively evaluate AI recommendations in general (non-clinical) settings have demonstrated the often counterintuitive nature of how performance is impacted by explainability. In one experiment, the response shift (a marker for AI influence in a manner similar to my experiment) was larger when an explanation accompanied the AI suggestion [150]. However, explanation quality (good vs. poor) did not affect this shift. The authors suggest that explanation presence might have reassured subjects when the AI performance itself was good rather than them actually assessing explanation fidelity or quality [150], a type of automation bias. In a medical environment, the downstream consequences of erroneously acting on poor advice can be significant with a study of 50 general practitioners (GPs) showing that only 10% were able to correctly recognise and disagree with incorrect AI advice on a dermatology problem [151] while in a radiology setting even task experts were not immune to the effects of poor AI advice (though they were considerably more likely to reject it than non-task experts) [87].

While explanations have traditionally been offered as a way to rescue users from poor AI advice, this has not clearly been borne out in clinical studies. For example, in a study investigating a psychiatric medication decision support tool, the presence of explanations did not mitigate the impact of intentionally poor AI recommendations, suggesting a level of automation bias that could be problematic in a real-world clinical environment [68]. In this study, the explanation did not significantly increase adherence above and beyond the AI suggestion alone, which could be due to several potential causes. It is possible that adherence and trust were maximised in most subjects by the AI suggestion alone leading to a ceiling effect. Alternatively, the very basic nature of the XAI might not have been persuasive. Or perhaps the nature of needing to make repetitive and cognitively burdensome decisions led physicians to rapidly adopt a heuristic, one way or the other, as to whether they used the explanation in their decision-making process or not.

Furthermore, there were several users who commented that the variables provided by the AI as part of the feature importance explanation did not seem physiologically plausible (see appendix C for a selection of post-experiment subject comments). This presents a challenge for XAI system designers. On one hand, a more complex explanation might satisfy users and lead to higher adherence. However, the strength of AI in identifying patterns in large datasets that are imperceptible to human clinicians can also be a weakness in developing XAI. If some clinicians equate the quality of an AI explanation with known physiological plausibility, then an AI explanation based on patterns that clinicians do not usually see in their everyday practice is unlikely to be persuasive. Ultimately, a mixed qualitative-quantitative approach might be crucial for gaining a better insight into the decision-making process that physicians use as well as shedding light on where specifically in this process AI can be of most benefit and least harm.

Other pertinent topics within medical AI-driven decision support research include the source of the advice and the experience level of the clinical audience receiving it. Gaube and colleagues studied a chest radiograph diagnostic challenge (with XAI support) among both clinical experts and non-task experts. They found that experts rated their confidence higher when advice was labelled as originating from an AI (although their accuracy was unchanged), while non-task experts showed improved accuracy with the provision of explanations (compared to experts who did not) [92]. I am not able to compare the peer impact directly between studies because in my experiment the peer information was genuine as opposed to synthetic as in the experiment by Gaube and colleagues (and so differed in magnitude from the AI suggestions). However, I nonetheless also found that AI suggestions were more influencing than peers as a source of suggested advice. I did not, though, find any association between adherence to AI and level of clinical experience.

2.4.4 Conclusion

Taken together, these findings on a comparatively large clinical expert population raise important questions for the design of clinical XAI systems. Specifically, I show that the marginal impact of XAI was low in this experimental setting. The exact form, presentation, and feedback loops for clinical XAI systems that effectively influence physicians remain unclear. It is likely that a more comprehensive consideration of the social and cognitive aspects of decision-making alongside the technical deployment of AI systems is required. Additionally, this study casts doubt on the utility of self-reports as a valid metric for assessing XAI among medical experts. What is instead required is a higher fidelity and more granular marker with which to assess the natural behaviour of physicians as they interact with AI-driven decision support tools.

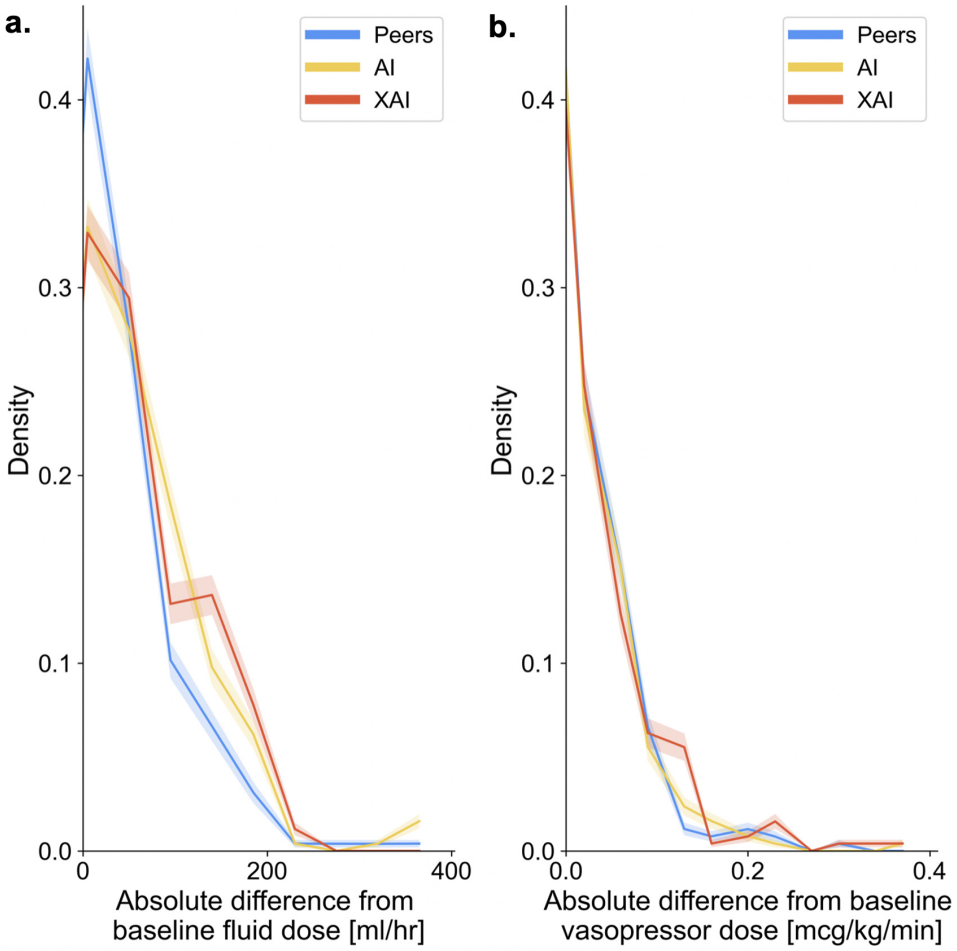


Fig. 2.9 **Aggregate dose shift by intervention arm.** Absolute difference (i.e. 50ml in either direction treated as +50ml discordance) from dose in the baseline group is shown, aggregated for all 24 patient scenarios for (a) fluid and (b) vasopressor. The error bars are formed by randomly taking 1,000 bootstraps of the data (80% subset with replacement) and estimating a distribution for prescription doses (error bar is distribution mean +/- standard deviation).

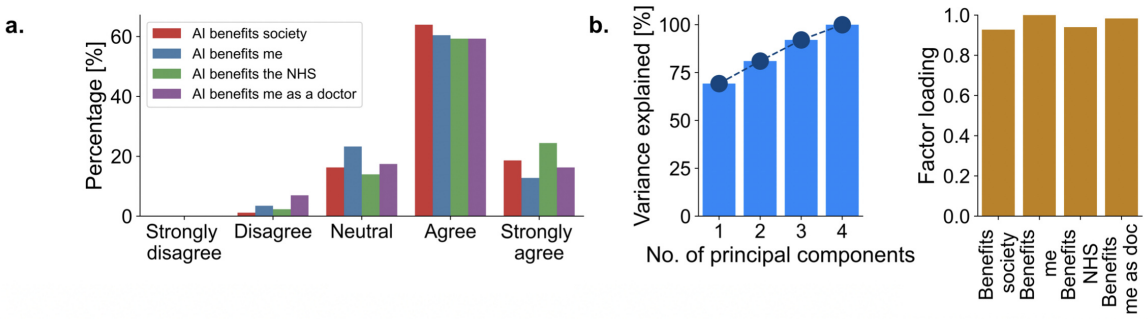


Fig. 2.10 **Principal component analysis of attitude to AI.** Four AI statements were presented to subjects pre-experiment who were asked for their agreement. In (a), principal component analysis (PCA) was applied to the results of these four questions with 69% of variance explained by a single component, see (b).

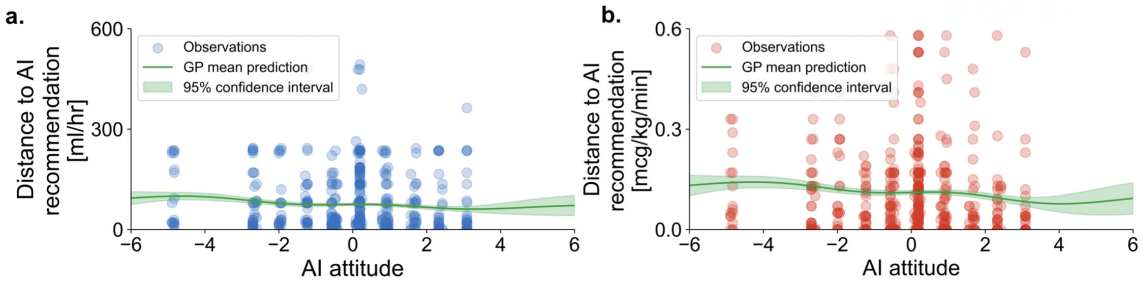


Fig. 2.11 **Impact of AI attitude on adherence to AI suggestions.** The single component composite for attitude to AI (see 2.10, higher value implies more positive AI attitude) was compared to absolute difference from the AI suggested dose, i.e. a proxy for AI adherence with lower value indicating greater adherence and vice versa in (c) for both fluid (blue) and vasopressor (red). Dot transparency in (c) represents density of points at any given location. 'GP mean' refers to a predicted Gaussian Process regression fit of the data with accompanying 95% confidence interval.

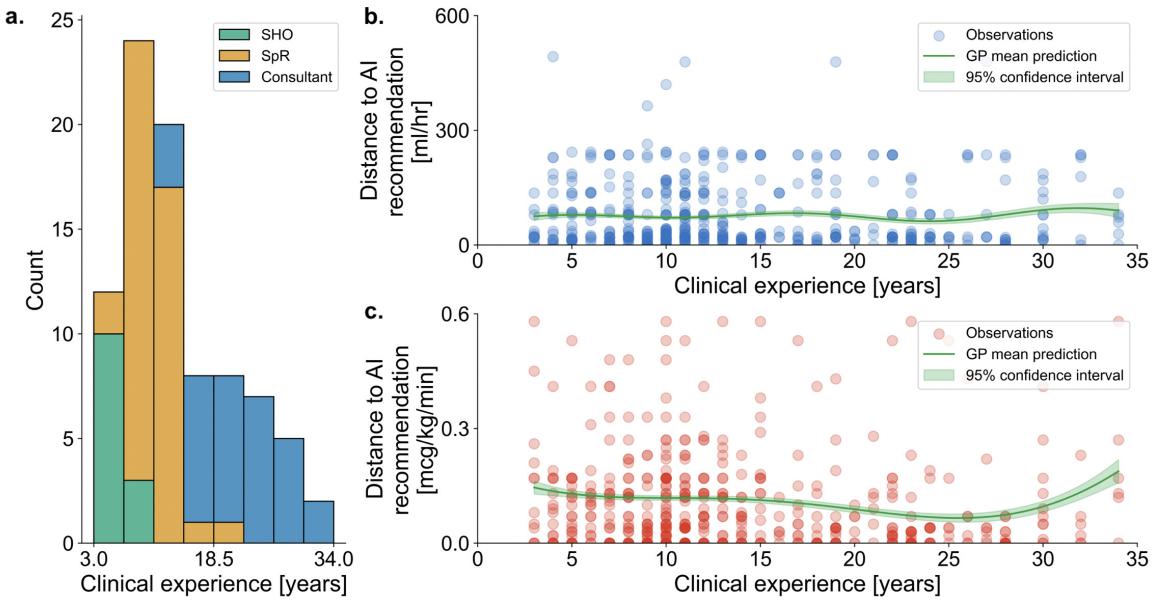


Fig. 2.12 **Impact of duration of clinical experience on adherence to AI suggestions.** In (a), the distribution of experience levels among the 3 categories of seniority is shown (Consultant, most senior and equivalent to attending in the United States (US); SpR, specialist registrar and equivalent to fellow in the US; SHO, senior house officer and equivalent to resident in the US). Experience level was compared to AI adherence (in the form of absolute difference from the AI suggested dose) for both fluid in (b) [blue] and vasopressor in (c) [red]. Lower value indicates higher adherence. Dot transparency in (b) and (c) represents density of points at any given location. 'GP mean' refers to a predicted Gaussian Process regression fit of the data with accompanying 95% confidence interval.

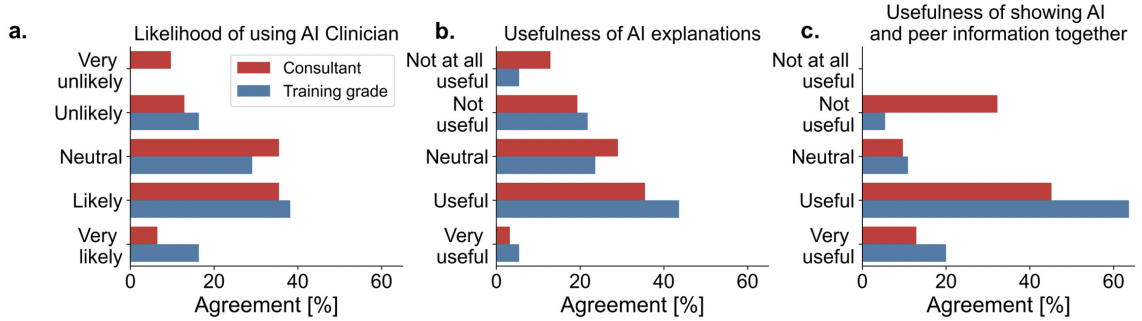


Fig. 2.13 **General post-experiment questions.** In (a) to (c), responses to post-experiment questions are shown down by training status.

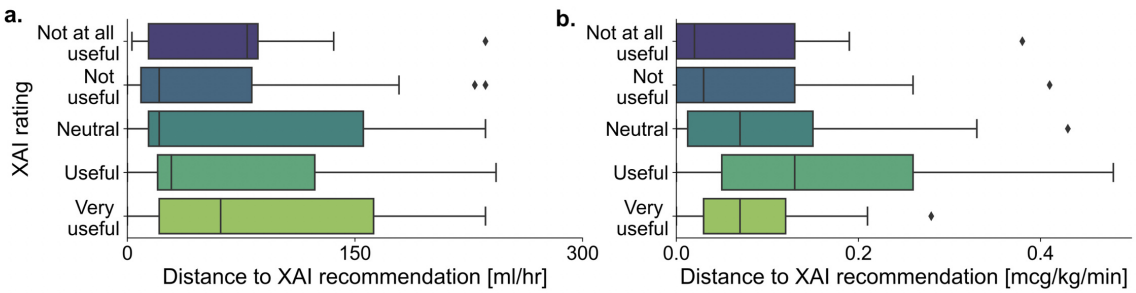


Fig. 2.14 **Post-experiment questions related to XAI utility.** Self-reported usefulness of explanations is plotted against adherence to XAI recommendation for both fluid in (a) and vasopressor in (b). Lower values indicate greater adherence and vice versa. For each boxplot, the centre line represents the median, box edges represent upper and lower quartiles, whiskers represent 1.5x inter-quartile range and diamonds are outliers. Colour spectrum gradient over boxplots relates to how useful XAI was rated.

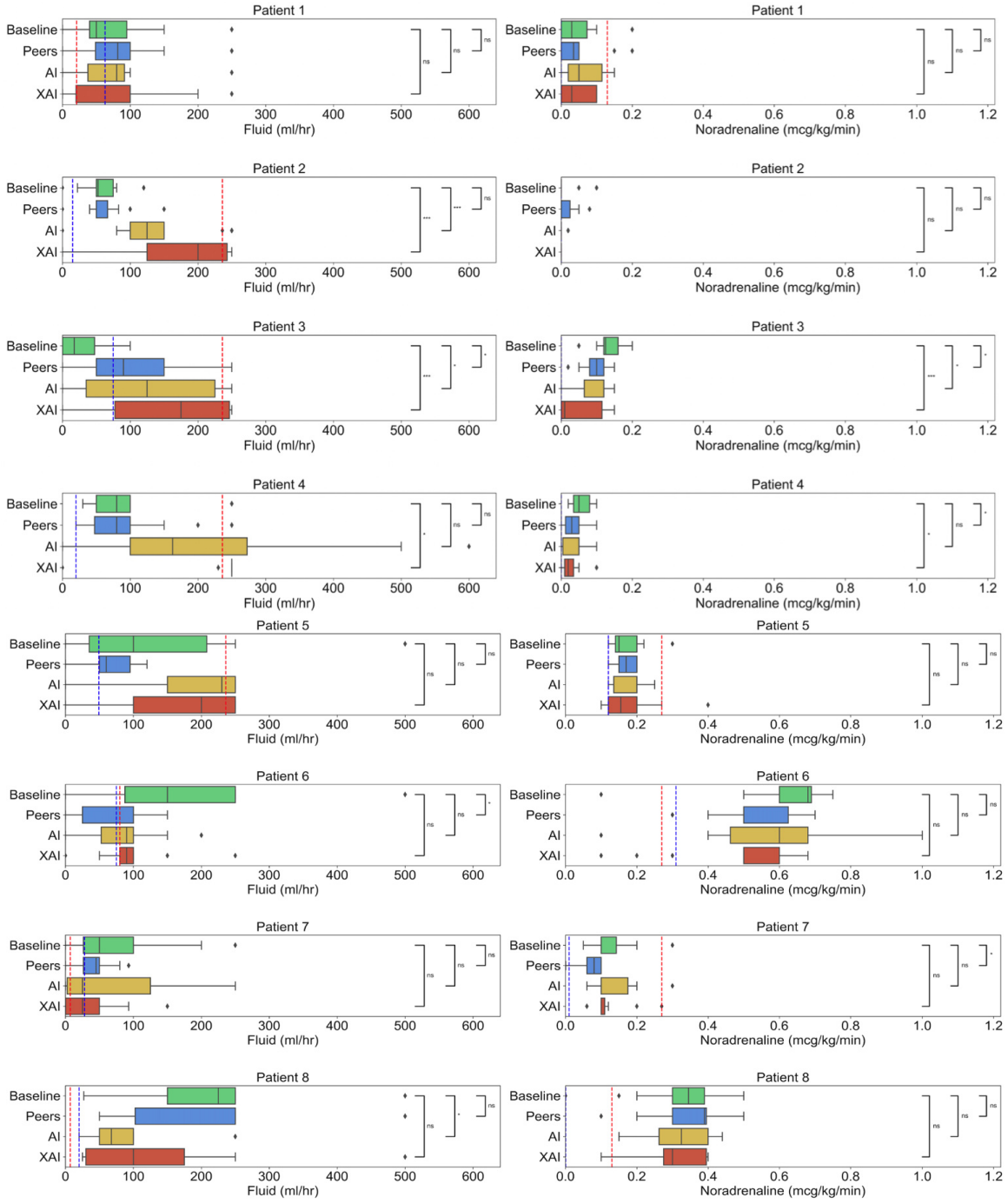


Fig. 2.15 **Boxplots of prescription dose distributions by trial arm 1/3.** A different box colour for each arm, same colour scheme as 2.9) for either fluid or vasopressor and for each patient scenario. For each boxplot, the centre line represents the median, box edges represent upper and lower quartiles, whiskers represent 1.5x inter-quartile range and diamonds are outliers. Blue dashed line represents the median of the peer distribution data (only available to those in the 'Peers' arm). Red dashed line represents the AI suggested dose (only available to those in the 'AI' or 'XAI' arms). Significance stars are obtained from independent T-tests for the means of the distribution ('Baseline' vs. intervention arm) with 'ns' referring to non-significant, a single star referring to a p-value <0.05 and three stars referring to a Bonferroni corrected p-value <0.000694 (for 72 comparisons in this figure).

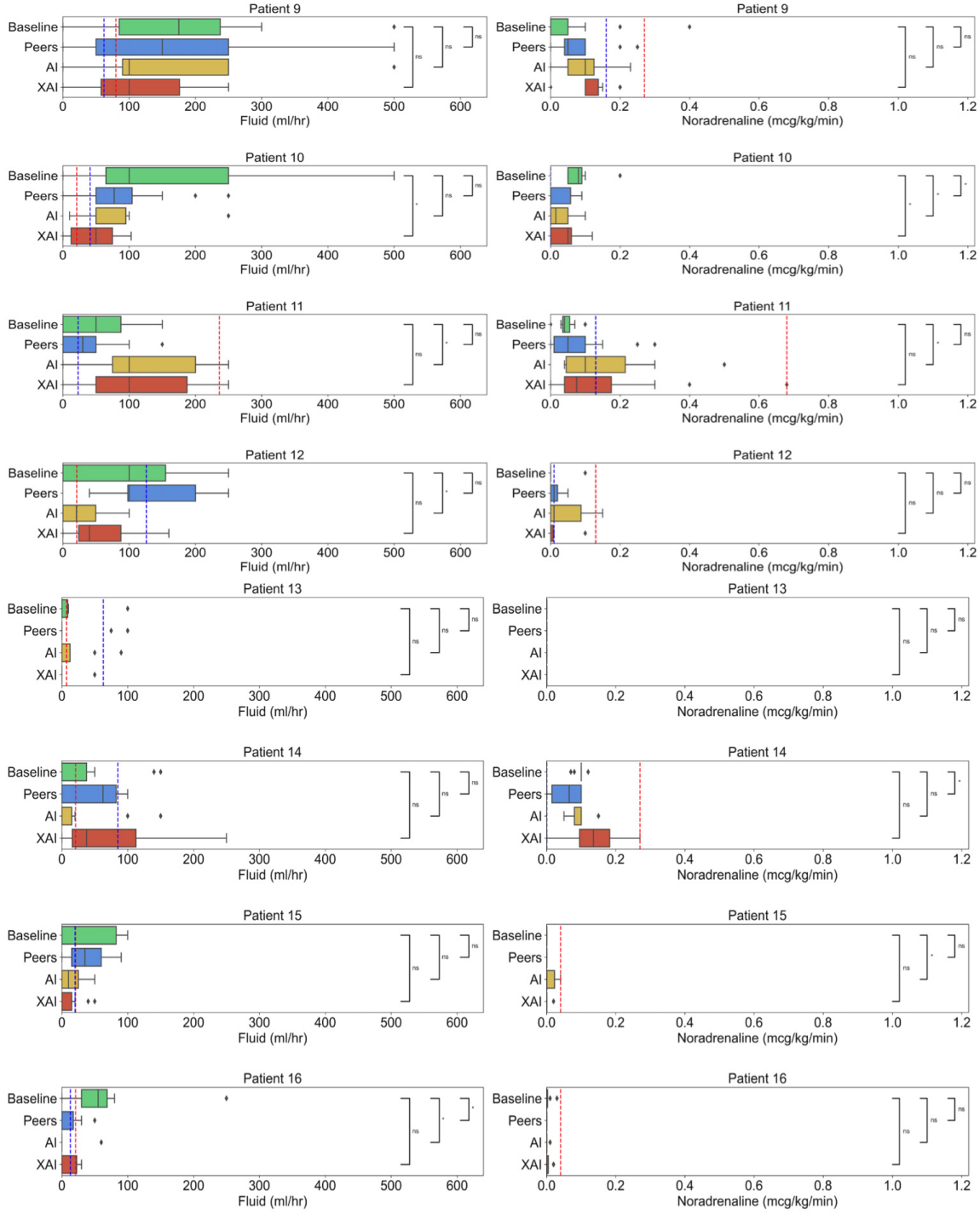


Fig. 2.16 **Boxplots of prescription dose distributions by trial arm 2/3.** Identical legend to Figure 2.15

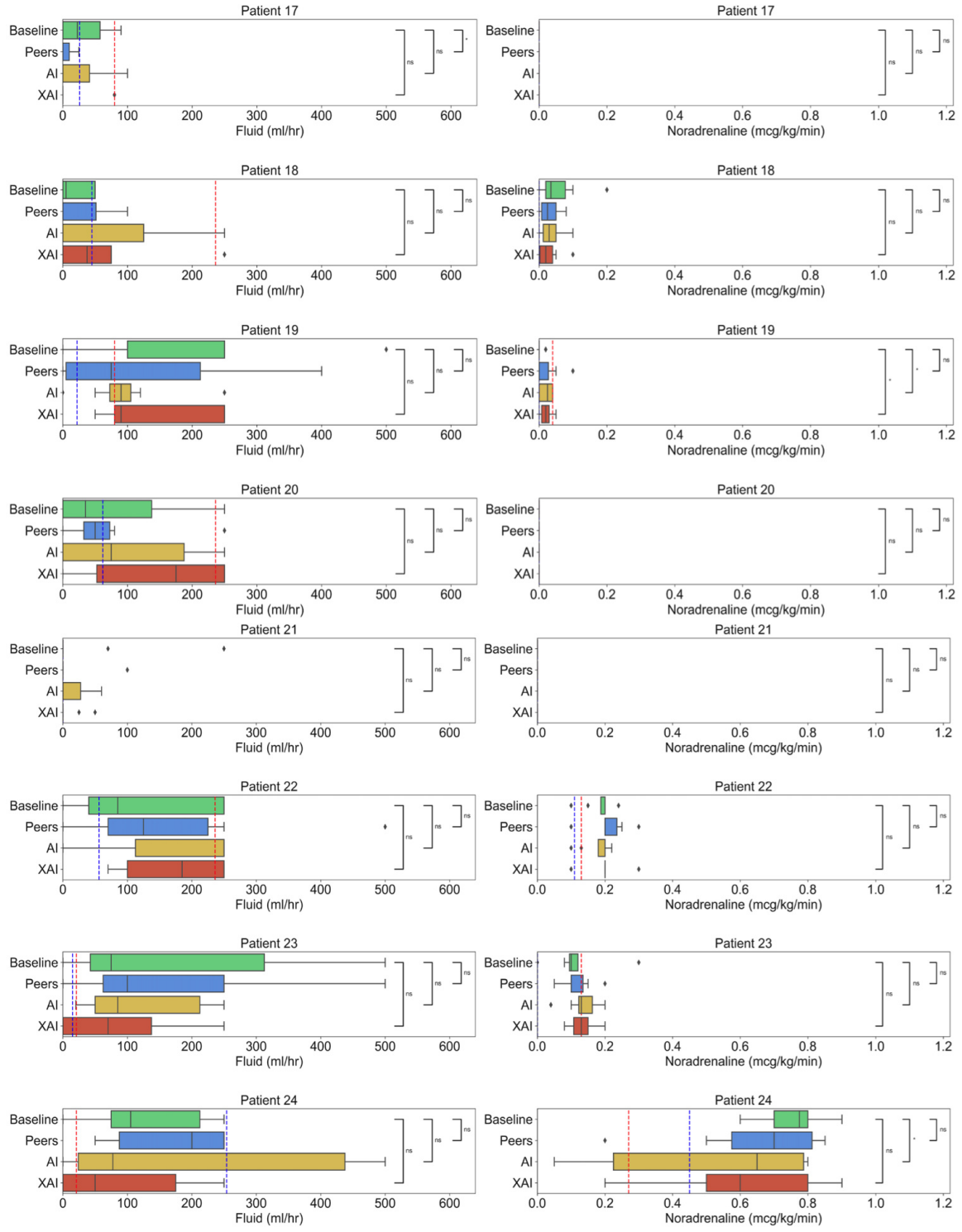


Fig. 2.17 **Boxplots of prescription dose distributions by trial arm 3/3.** Identical legend to Figure 2.15

Chapter 3

Development of an eye-tracking pipeline to evaluate user interaction with XAI

This chapter is closely based on a manuscript under preparation on which I am the first author and my supervisors are co-authors:

Enhancing physician-AI interaction studies with free-movement eye tracking. Submission to a peer-reviewed journal specialising in digital health is planned in due course.

3.1 Introduction

3.1.1 Alternatives to self-reports

The evaluation of XAI within AI-driven clinical decision support tools covers several behavioural aspects of human-AI interaction as discussed in Chapter 1 (see section 1.4). Self-reports seek to determine the volitional and conscious perceptions of clinical end-users regarding the utility of XAI but are subject to numerous biases that can distort the true link between actual and perceived XAI utility. They have been found to be unreliable even in other non-AI settings [152]. The previous experiment in

Chapter 2 demonstrated a clear lack of correlation between self-reported utility of XAI and the influence that XAI had on actual prescription decision-making [138]. This counter-intuitive finding calls into question the use of self-reports as a core evaluation metric and requires a search for alternatives.

One such alternative involves collecting directly observed behavioural data and doing so in real-time (as opposed to self-reports collected after the fact). This could, for example, be done using a think-aloud protocol (TAP) where participants verbalise their decision-making process as it occurs and this is recorded for downstream analysis [153, 154]. However, TAP has several drawbacks [155], including altering participants' approach to the task through the mere act of verbalising, increased cognitive load (simultaneously thinking and speaking will not come naturally to all participants), and resulting in qualitative data that typically then require thematic analysis. Eye-tracking is an alternative behavioural proxy that focuses on where a participant's attention is focused via their gaze [156, 157]. It satisfies many of the criteria we might want for this purpose by providing data that is objective, quantifiable, free of reporting bias and obtained in a non-invasive manner. Indeed, eye-tracking has been employed in many other non-clinical settings for quantifying aspects of human behaviour through analysing where a participant's attention is focused [158–161].

Eye-tracking can be performed in a variety of settings ranging from vignettes to simulation suites to real-world clinical environments. The granularity of data collected is proportional to the fidelity of the setting. If used in a vignette experiment then only gaze data on the computer screen will be collected. In contrast, a free-movement eye-tracking experiment set in a simulation suite can instead capture gaze patterns over a range of surfaces of interest (and in a manner that is more natural and representative of true clinical practice).

A key advantage of real-time data collection is its ability to reduce feedback loop latency, enabling quicker learning from user behaviour [162]. For instance, in a system

that personalises explanations for clinicians, this could be approached as a reinforcement learning problem. Here, the state is the user's observed behaviour when interacting with the AI, the reward is +1 if the user follows the AI's advice and -1 if they don't, and the actions available to the reinforcement learning agent involve displaying one of several explanation types. The agent learns with each time-step. Self-reported data might be sporadic, but real-time eye-tracking data provides continuous feedback after every interaction, facilitating faster learning and a more efficient system overall. Of note, the eye-tracking in this experiment was not, strictly speaking, available in real-time as the aim was to develop a proof of concept regarding the use of eye-tracking as a modality more generally (this is discussed more in the limitations section at the end of this chapter (see section 3.4).

3.1.2 How eye-tracking works

Eye-tracking provides quantitative data on eye position and characteristics relative to the external world [163]. This is used to infer information about a subject's behaviour (and by extension their cognition) as they interact with their environment. There are two main types of eye-tracking devices: screen-based and wearable. Screen-based devices, usually mounted to a computer screen, are best for tasks where the subject reviews information primarily on the screen without the need for ambulation. They offer quick calibration and no need for external equipment but are unsuitable for experiments involving free movement, such as in simulation suites or real-world clinical settings.

In contrast, wearable devices address this weakness as they are lightweight and have no ambulation restrictions. Such devices usually take the form of glasses with three cameras [164]. An external camera captures the subject's (ego-centric) perspective (referred to hereafter as the 'world-camera') [165]. Another pair of cameras (referred to hereafter as 'pupil cameras') focus on the subject's eyes. Eye-tracking is achieved

using pupil-corneal reflection. The pupil cameras continuously record images of both eyes while an infrared illuminator on the glasses shines light onto the eyes, invisible to the subject but detectable by the cameras [166]. The reflections are used to distinguish the pupil from the rest of the eye. Gaze angle is calculated using image processing algorithms that can then map the angle onto the subject's external world-view captured by the world camera to determine where in the world-view the subject's gaze is directed [166].

The raw data captured by the cameras allow both gaze-dependent and gaze-independent eye-tracking metrics to be calculated. The core components of eye-tracking are fixations and saccades [163]. A fixation is defined as a period when the eyes remain fairly immobile, focusing on a single point. Fixations provide time for encoding and processing of visual information by the brain. In contrast, saccades are rapid, often jerky movements of the eyes as they shift focus from one point to another. These quick movements allow repositioning of the fovea (small central pit in the retina responsible for sharp central vision) onto a new area of interest. Encoding and processing of information is suppressed during a saccade (although this intermittent functional blindness is not perceptible by a subject due to the brain constructing a continuous and seamless subjective experience of the vision).

Gaze-dependent metrics include gaze duration, gaze speed, fixation count, fixation duration, saccade count, saccade duration, saccade amplitude and saccade velocity. Gaze-independent metrics include pupil diameter and blink rate which are both highly responsive to external stimuli (respectively to lighting levels or lubrication levels/external eye irritants such as grit or dust). However, pupil dilation and blink rate can also reflect noradrenaline and dopamine activity respectively, which are essential neurotransmitters for cognitive functions concerned with promoting goal-directed behaviour [163, 167]. For example, dilated pupils also occur during conditions of mental effort, extreme task focus and cognitive arousal thereby acting as a potential

proxy for cognitive load [168]. In contrast, blink rate is typically reduced during reading and other times of high cognitive load [169].

3.1.3 Eye-tracking in clinical settings

Eye-tracking has been used in clinical settings for both the direct diagnosis of medical conditions in patients as well as the assessment of medical professionals. Examples of the former include diagnosis of conditions such as attention-deficit/hyperactivity disorder (ADHD), mild cognitive impairment, Alzheimer's disease, amyotrophic lateral sclerosis (ALS) and dementia [170, 171]. Eye-tracking studies of clinicians typically fall into one of two categories: (1) assessment of performance in a task, often looking for differences between novices and experts [172–174] and (2) assessment of clinician interaction with some aspect of healthcare. This aspect might be, for example, a surgical instrument, the electronic health record itself [175, 176] or medication labels (looking to see if different styles of label are associated with different patterns of gaze behaviour and thereby affect the likelihood of medication errors [177]).

Where eye-tracking studies have had a focus on AI, some studies have concentrated on using the eye-tracking data as additional input into the AI with which to improve diagnostic performance of the AI rather than interaction with the AI system per se. For example, combining radiologist fixation data with chest radiographs to predict areas of pathology on the image [178]. This has become a popular enough approach that specific datasets now exist to facilitate reproducibility and encourage further development of this topic [179]. In other cases, machine learning techniques have been used to facilitate the analysis of eye-tracking data such as machine learning based classification of radiologists into expert or novice experience categories [180]. Few studies have focused specifically on using eye-tracking to assess human-AI interaction. One study found that higher levels of AI support led to lower user cognitive workload [181] while another found strong positive correlation between both user-AI agreement and

self-reported perceived reliance on AI with gaze duration [182]. A study by Cioffi and colleagues concluded that higher viewing efficiency was achieved when an AI-enabled optical coherence tomography image support device was used by cardiologists while performing coronary intervention (specifically reduced time to first fixation, reduced total task duration and increased gaze duration within the region of interest [183]). A rare example of eye-tracking used to assess human-AI interaction with a clinical AI-driven decision support tool (as opposed to image support alone) is a randomised trial by Arsiwala-Scheppach and colleagues [184]. They found that AI-supported dentists, when compared to their non-AI supported colleagues, fixated sooner and for longer on relevant image areas while also examining the images less systematically than those without AI [184]. In other words, their viewing behaviour was more efficient (which in non-AI circumstances is usually a characteristic discriminating experts from novices).

In this study, I developed and tested a pipeline for recording and analysing free-movement eye-tracking data of clinical experts interacting with XAI in a simulation suite.

3.2 Methods

3.2.1 Simulation suite and XAI experiment

The details of the setting and conduct of the XAI aspects of the experiment are provided more comprehensively in Chapter 4 (see section 4.2) as the emphasis in this chapter is instead on the eye-tracking pipeline itself. In brief, physicians in the simulation suite (see Figure 3.1) could encounter any of six patient scenarios, each paired with either a safe or unsafe AI recommendation (independent variable) accompanied by four kinds of explanations. For each scenario, physicians had to conduct an assessment, including reviewing patient data and examination of the patient mannequin, before being queried

by a nurse about intravenous fluid and vasopressor (norepinephrine/noradrenaline) prescriptions for the next hour of the patient's admission. After viewing the AI recommendations and simultaneously presented explanations on a nearby large display, physicians were then asked to affirm or alter their prescription doses (dependent variable). Gaze detection was employed throughout.

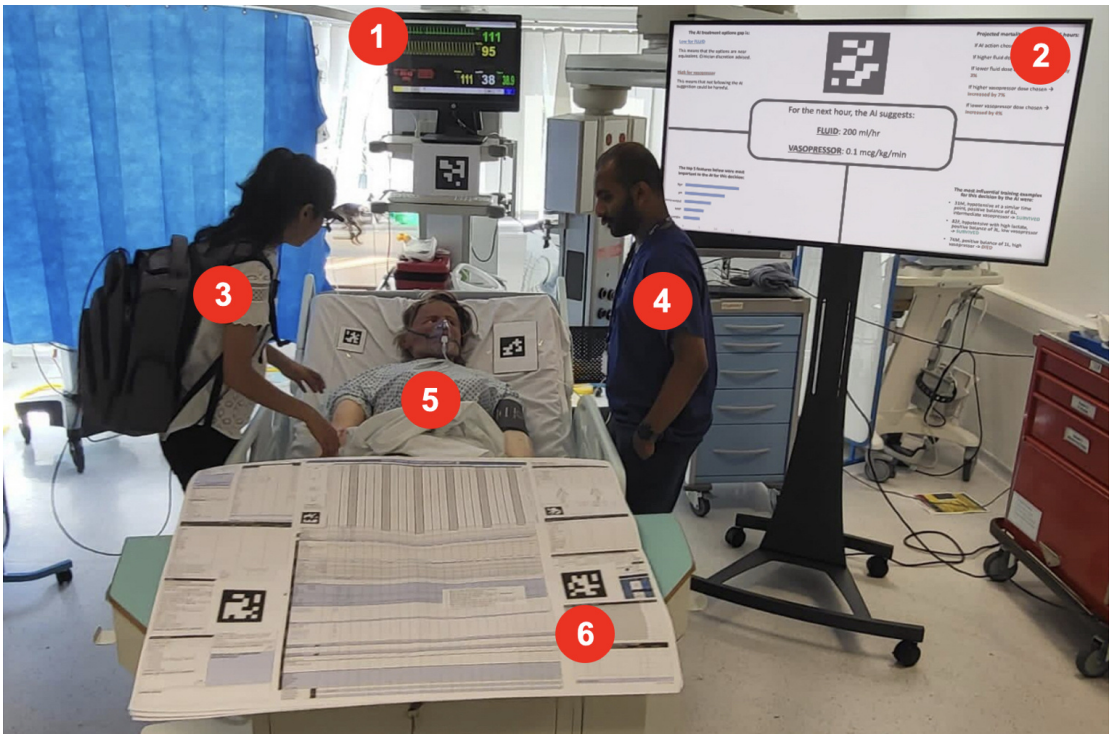


Fig. 3.1 **Simulation suite.** (1) vital signs monitor, (2) AI screen, (3) physician, (4) bedside nurse (played by experimenter i.e. me), (5) high-fidelity patient mannequin, (6) bedside ICU data chart.

3.2.2 Equipment and software

I used non-invasive commercially available eye-tracking glasses (Pupil Core headset and Pupil Labs software, Core, version 3.3), see Figure 3.2a. The Pupil Labs software (Pupil Capture, version 3.5.7) was used to demarcate the pupil from raw video input (see Figure 3.2b) and use this data to calculate where in the world-view the subject's gaze was directed. During each experiment, the glasses were connected to a laptop

(Lenovo Thinkpad) which was placed into a lightweight backpack worn by subjects to allow free movement around the simulation suite (and without trip hazard risk from wired connections to a fixed computer).

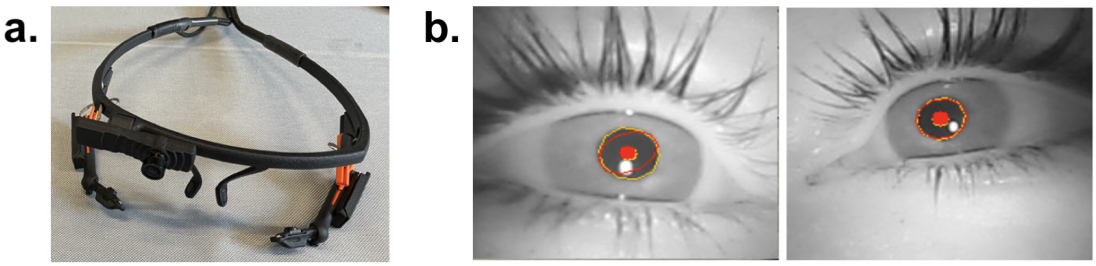


Fig. 3.2 **Eye-tracking glasses and pupils.** Eye-tracking glasses in (a) and example of pupil identification in (b).

3.2.3 Tagging regions of interest

Regions of interest (ROIs) were defined as surfaces within the simulation suite for which recording gaze based attention data might shed light on the behaviour of a subject interacting with XAI. This consisted of two categories: AI and non-AI (or traditional clinical) surfaces (as one particular question that arises is whether physicians have a different attention profile with AI versus traditional clinical surfaces).

The AI ROI consisted of the AI display screen itself (containing an AI treatment suggestion in the centre with four explanations in each corner of the screen). This surface was further subdivided into five regions (corresponding to the AI suggestion alone and each of the four explanations). The non-AI (traditional clinical) surfaces included a patient mannequin (Simman 3G, Laerdal Medical, Stavanger, Norway), a clinical vital signs monitor and a paper ICU chart containing daily patient data including blood results, drug infusions and ventilator settings among many other data features.

Each ROI was connected to at least one (but in some cases multiple) QR codes known as April tags (see Figure 3.3) that could be recognised by the Pupil Labs

software. These tags allowed for post-processing identification of surface data and delineation of surface boundaries (i.e. they facilitated the drawing of ROI boundary boxes during post-processing).

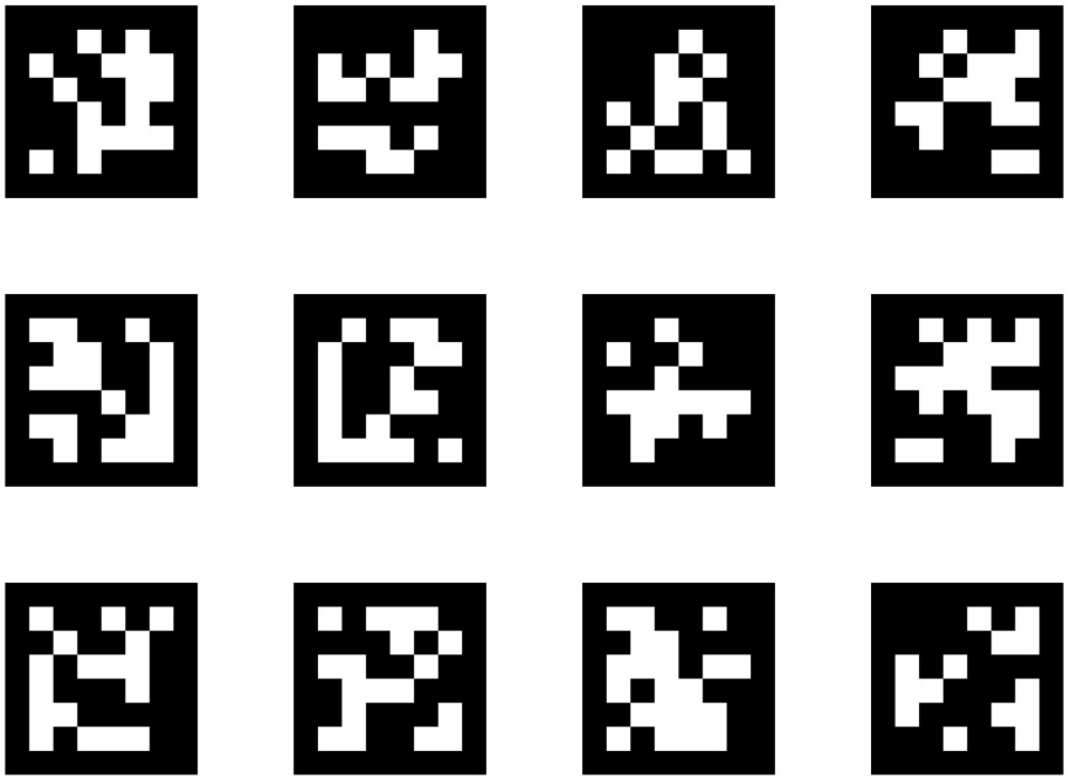


Fig. 3.3 **April tags**. April tags are QR codes used to identify regions of interest during post-processing.

April tags came in three sizes (30, 60 and 100mm squares) for use depending on how far away the surface was likely to be from the world-view of the subject. A smaller tag was appropriate for close surfaces so as not to cover too much of the ROI while a larger tag was required for more distant objects. The white margin around each tag was also essential for correct identification by the Pupil Labs software. Given the free-movement nature of the task (i.e. a surface might be both near and far from the subject's world-view perspective at different points in the task) I opted to use a range of sizes for some surfaces such as the patient mannequin and ICU paper chart.

Another complicating consequence of the free-movement environment was that subjects could view ROIs from different angles too. This was less likely for certain ROIs (those that were intrinsically flat such as the AI screen and the paper ICU chart) and more likely for the patient mannequin which could be approached from either side. This necessitated multiple April tags to cover all potential viewing angles.

3.2.4 Calibrating eye-tracking glasses to the subject

Calibration of the eye-tracking glasses to each individual subject was necessary for (i) optimal identification of pupil position and (ii) for correctly calculating gaze position based on the video from the world-view camera. For the first of these two steps, each of the pupil cameras on the glasses could be adjusted in two planes of movement. This allowed me to ensure that the image of the eye was in the centre of each camera's video output and that the software had correctly delineated the pupil with a red circle (see Figure 3.2b).

For the second of the two calibration steps, two gaze-to-target exercises were conducted. The first exercise was a static gaze-to-target task in which the subject was asked to stand with their eyes 60cm from a blank laptop screen. Five circular markers sequentially appeared on the screen, one after another, with each in a different corner of the screen and the final marker in the centre of the screen. The subject was instructed to keep their head still and move only their eyes to ensure that their gaze fell upon each marker as it sequentially appeared. Once the Pupil Labs software had confirmed gaze upon the relevant marker, it would disappear and reveal the next marker. This was the 'screen marker' mode on the Pupil Labs calibration system.

For the second exercise, another gaze-to-target exercise was conducted but this one was depth-based incorporating standing at 1 metre distance initially and then 2 metre distance from a 60-inch display screen. At each distance, nine simultaneous circular markers were displayed on the screen in a 3x3 geographical configuration. The subject

was instructed to again keep their head still and move their eyes to focus gaze on the marker in the top left of the screen. Sequentially, they were instructed to move their gaze to the next marker along from left to right in each row and then from the first row to the last row. The change in depth ensured that calibration was optimised across a range of potential distances from subject to target ROI (which facilitated eye-tracking in the free-movement environment of the simulation suite). This exercise used the 'natural features' mode on the Pupil Labs calibration system with each marker being manually selected as a natural feature in turn. Total length of calibration ranged from 10-15 minutes depending on whether any repetition was required.

3.2.5 Post-processing of videos using proprietary Pupil Labs software

Raw video recorded from all three cameras underwent real-time processing by the Pupil Labs software to calculate gaze position. However, post-processing (defined here as processing performed after the completion of the real-time recording) was still required. The initial post-processing took place within the Pupil Labs software suite. First, each video underwent a manual review (typically at between 2x and 3x watching speed with pauses or slow-down to real-time at any points of interest). The primary purpose of this review was to act as an initial quality assurance step with poor quality videos excluded for further analysis. There were two main reasons for video exclusion at this stage. First, a key ROI (e.g. AI screen or ICU chart) was not appropriately captured in the world view. This occurred because the world-view camera recorded in ultra-wide aspect ratio and therefore the vertical height of the recording was not always large enough in scope to capture the full panoramic view of the person. Second, videos with extremely poor detection of one or both pupils to the extent that less than 50% of the recording had a gaze position calculated.

The timings for the start and end of each scenario within the XAI experiment as well as the time of AI suggestion reveal were also extracted during manual review of

each video and recorded separately. The start of a scenario was defined as the first frame with the simulation suite door open while the end of a scenario was defined as the first frame with no visible simulation suite door.

The next step in the post-processing pipeline involved manually drawing boundary boxes for ROIs of interest. Each boundary box defined the ROI surface and was linked to one or more April tags. Manual review of each recorded video was performed to ensure that boundary boxes were consistently positioned over the actual ROI throughout the video. Finally, offsetting corrections (if required) were made. Where there had been an issue with calibration resulting in a fixed offset of gaze that was identical throughout the video, this offset was corrected. Alternatively, where there had been a new offset in gaze at some point during the recording, this offset was also corrected from the point it occurred. This second scenario typically occurred if, for example, a subject inadvertently moved the eye-tracking glasses slightly while touching their head. Despite subjects being reminded not to touch the glasses or their head after calibration, this nonetheless occurred in a few cases.

3.2.6 Post-processing of gaze data using Python

Tables D.1, D.2, D.3, D.4 and D.5 in Appendix D list the raw eye-tracking data exported from the Pupil Labs software for each subject's video. These data were then further processed by a customised Python pipeline I wrote. The final product is a cleaned master dictionary with each entry in the dictionary corresponding to a subject and containing a dataframe of processed eye-tracking data for the subject.

The key aspects of this pipeline were as follows. First, for each subject an empty dataframe was created with a single row per frame of video. Each row was then classified according to whether it was part of one of the six experiment trials (in which case the trial number was recorded) or whether it was not (i.e. the subject was in between trials).

Next, each frame was populated with raw eye-tracking data in a manner that facilitated analysis. Minimum and maximum gaze timestamps were used to compute the duration of each frame. Custom functions were applied to annotate each frame with case, AI reveal, and safety status. The area and position of ROIs within the world-view was handled by parsing raw transformation matrices of position and applying these to calculate corner coordinates. These corners provided the 2D world-view positions of the ROI per frame of video. Area of the ROI was also calculated per frame.

Gaze distance was computed by calculating the Euclidian distance between consecutive gaze points (where the recorded data confidence of both X and Y values was greater than 0.6 (0 to 1 scale). Blink data were processed to remove blinks where the recorded data confidence was less than 0.15 or where the duration of a blink was greater than 1 second. All confidence values and cut-off points for data were designed to maximise the trade-off between dropping low-quality data and not excessively dropping data that could meaningfully contribute to the analysis. At the end of this process, the following eye-tracking metrics were available for analysis: (i) gaze duration, (ii) gaze speed, (iii) fixation count, (iv) fixation duration, (v) saccade duration, (vi) saccade amplitude, (vii) saccade velocity, (viii) pupil diameter, (ix) blink rate.

Gaze per ROI was also analysed in an alternate manner that adjusted for 'visual real-estate' (i.e. the proportion of a subject's world-view that the ROI occupied). This was done by comparing the proportion of gaze on the ROI that would be expected by chance alone and comparing this to the proportion of actual gaze. So if an ROI occupied 25% of the visual field for 50% of the time, then we would expect the gaze to fall within the ROI, by chance alone, 12.5% of the time during the experiment. By then comparing this 12.5% 'random gaze' figure against the actual gaze proportion, a ratio could be calculated for the relative importance of ROIs that is adjusted for the proportion of world-view occupied by an ROI (thereby allowing a fairer comparison

between ROIs that might be of different sizes and positions as a subject moved around the simulation suite).

3.3 Results

Of 38 doctors recruited to the experiment, 19 had data of sufficient quality to include for eye-tracking analysis (13 male, 6 female). Mean age was 33 years (standard deviation (SD) 6 years). Reasons for exclusion of the other 19 were as follows: 13 were too low quality (4 had part of a key ROI cut off in the world-view camera, 2 had an issue with pickup of at least one pupil while 7 had an issues with pickup of both pupils), 2 had corrupted software files due to use of an alternative laptop, 2 had hardware issues (laptop failed to boot or recording failed), 1 failure of the sim suite control centre meant that the vital signs monitor was not operational and 1 subject declined to wear the eye-tracking glasses due to the glasses physically feeling uncomfortable to them. In general, contact lenses and even physical glasses worn underneath the eye-tracking glasses were not an absolute contra-indication (2 of 5 subjects wearing glasses had eye-tracking data of sufficient quality to include). Of 19 subjects included for the eye-tracking analysis, 13 required minor offset adjustment and 4 required offset adjustment to only part of the video (i.e. the subjects likely touched the eye-tracking glasses inadvertently after the calibration had been performed). Gaze heatmaps over ROIs are shown as an example for one subject in Figure 3.4. An example video of how the eye-tracking was performed is available on YouTube: https://www.youtube.com/watch?v=V3tM7nelvS8

3.3.1 Gaze

Mean gaze distance for all six patient scenarios subjects undertook in normalised units is displayed in Figure 3.5a stratified by whether the subject was junior or senior (defined as greater than 1 year experience in ICU). Gaze distance correlated with time

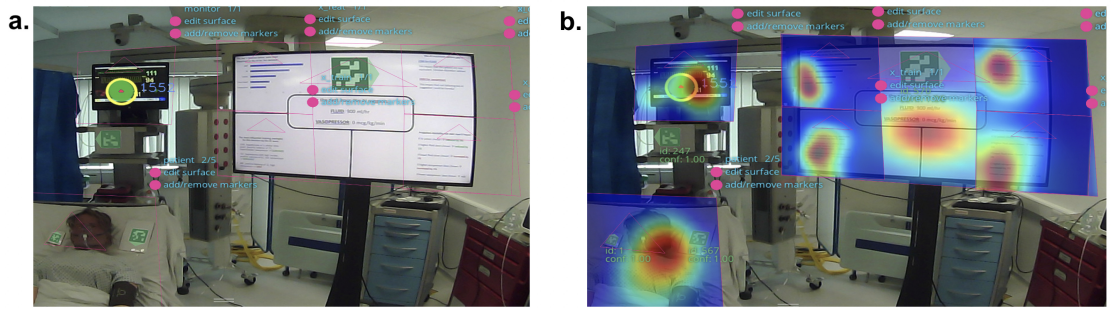


Fig. 3.4 **Video post-processing.** In (a), April tags (highlighted in green) are auto detected and allow regions of interest (ROIs) to have manual bounding boxes drawn. In (b), heatmaps are shown for gaze proportion within each of the annotated ROIs (where red represents high duration of gaze and blue represents low duration of gaze).

taken (i.e. distance was longer for subjects who took longer to complete a trial) so I adjusted for time by instead also assessing gaze speed (see Figure 3.5b). Mean speed by senior participants was 1.4 normalised distance units per second (SD 0.9, 95% CI 0.7 to 2.2) and mean speed for junior participants was 1.2 normalised distance units per second (SD 0.7, 95% CI 0.7 to 1.6). The vast majority (98%) of the variance in gaze speed was attributable to inter- rather than intra-subject differences as expected.

3.3.2 Fixations and saccades

Mean fixation duration overall was 149 ms (SD 7.7 ms) with no major difference seen due to experience level or whether the patient scenario was coupled to a safe or unsafe AI suggestion (see Figure 3.8a). Mean number of fixations was 636 overall (SD 540, 95% CI 549 to 722). Mean saccade duration overall was 130 ms (SD 53 ms, 95% CI 104 to 155) with no major difference seen due to experience level or whether the patient scenario was coupled to a safe or unsafe AI suggestion (see Figure 3.8b). Similar results were seen for saccade amplitude and saccade velocity (see Figure 3.8c and Figure 3.8d respectively).

The ratio of fixation duration to saccade duration is sometimes used as a proxy for cognitive load (ratios greater than 1 indicating greater cognitive load and vice versa),

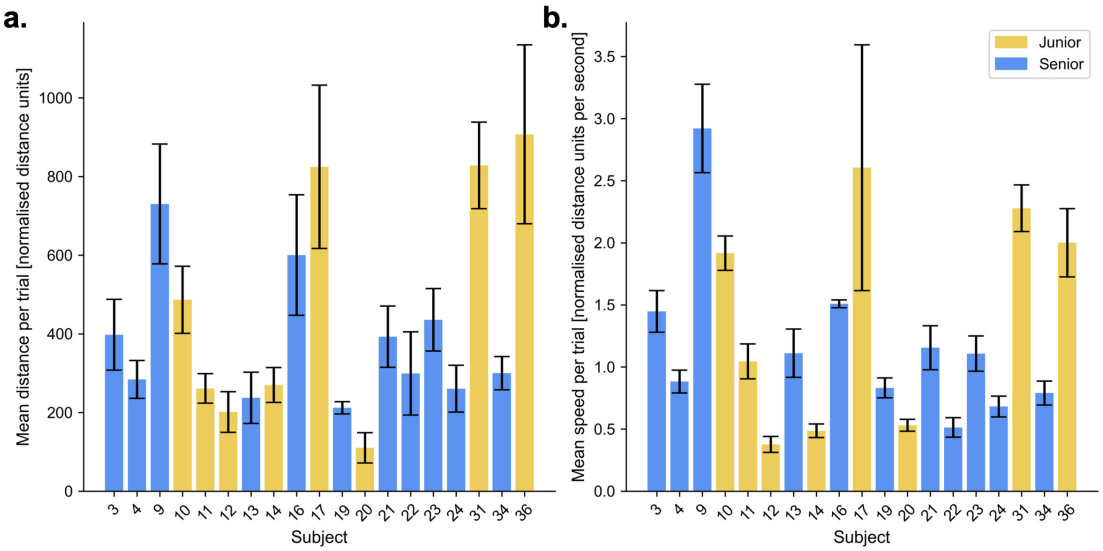


Fig. 3.5 **Gaze distance and speed.** Mean gaze distance by subject in (a) and mean speed by subject in (b).

hereafter referred to as the FS ratio. The FS ratio is displayed in Figure 3.9 stratified by safety status of the scenario and experience level of the subject. The main findings are that cognitive load is lowest on the patient mannequin and similar for the ICU data chart, vital screen monitor and the AI screen. Within the ICU data chart, the FS ratio is significantly larger for senior doctors than junior doctors (senior FS ratio 2.45 versus 1.27 for juniors, $p<0.001$). This was the most data rich of the four major surfaces.

3.3.3 Visual real-estate adjustment

Despite physicians moving around the simulation suite and the ROIs being different sizes within the world-view video, the adjustment performed for visual real-estate enabled a fairer like-for-like comparison. The corner coordinates calculated by parsing raw transformation matrices of position can be seen in Figure 3.10 which shows an example of the delineation of the AI screen ROI over a period of 500 frames (limited to 500 for illustrative purposes so that the lines can be made out easily). Each red

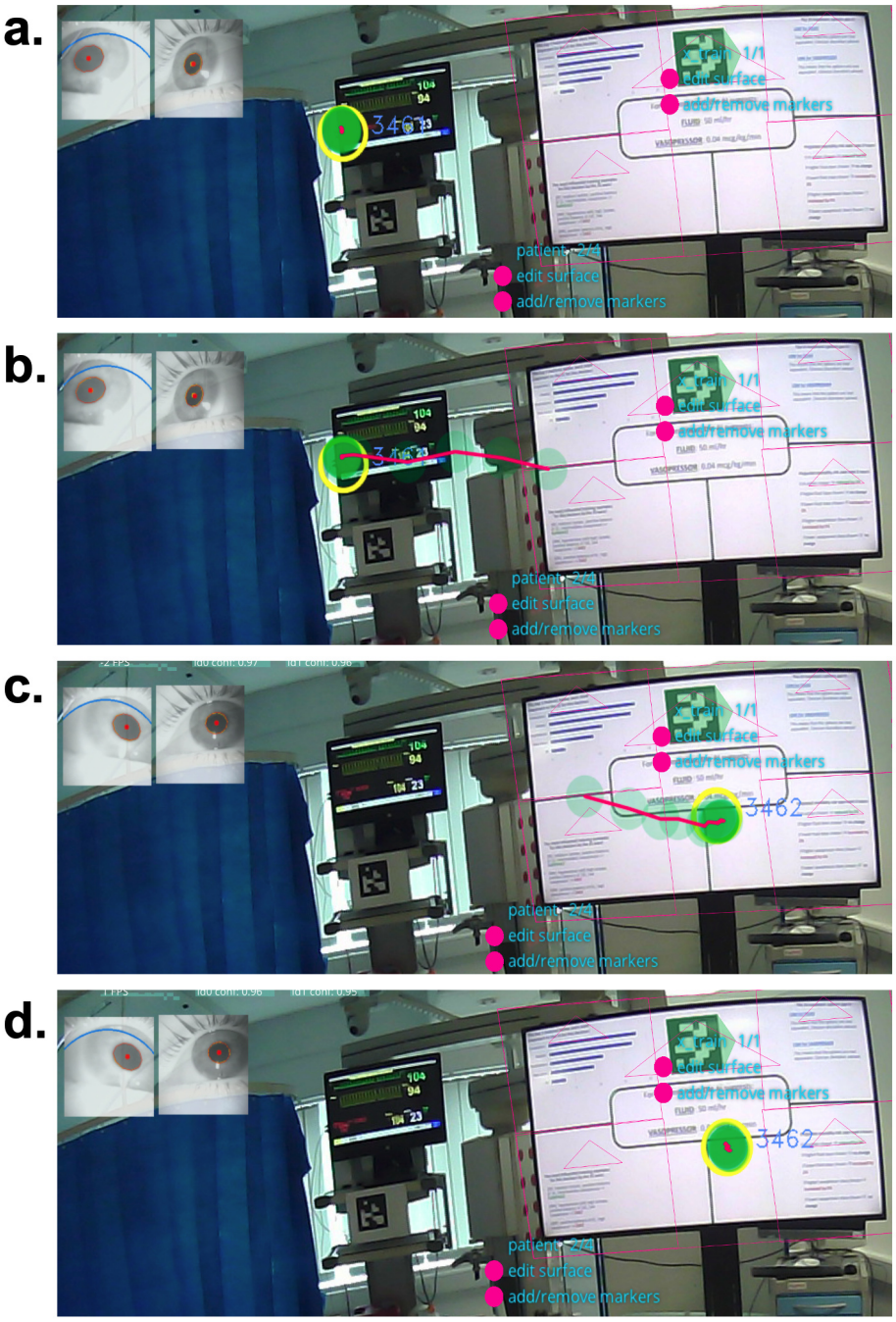


Fig. 3.6 **Demonstration of fixations, saccades and scanpath.** The four images from (a) to (d) cover 4 frames of video. In (a), the red dot within a green circle and yellow bounding box represents a fixation on the lower left corner of the vital signs monitor. In (d), the subsequent fixation is on the lower edge of the AI suggestion near the centre of the AI screen. In between these 2 fixations, the scanpath (i.e. path of saccadic movement between fixations) can be seen as the subject directs their vision from left to right. In all 4 images, the raw pupil position is visualised in the top left corner with red dots marking detection of the pupil centre.

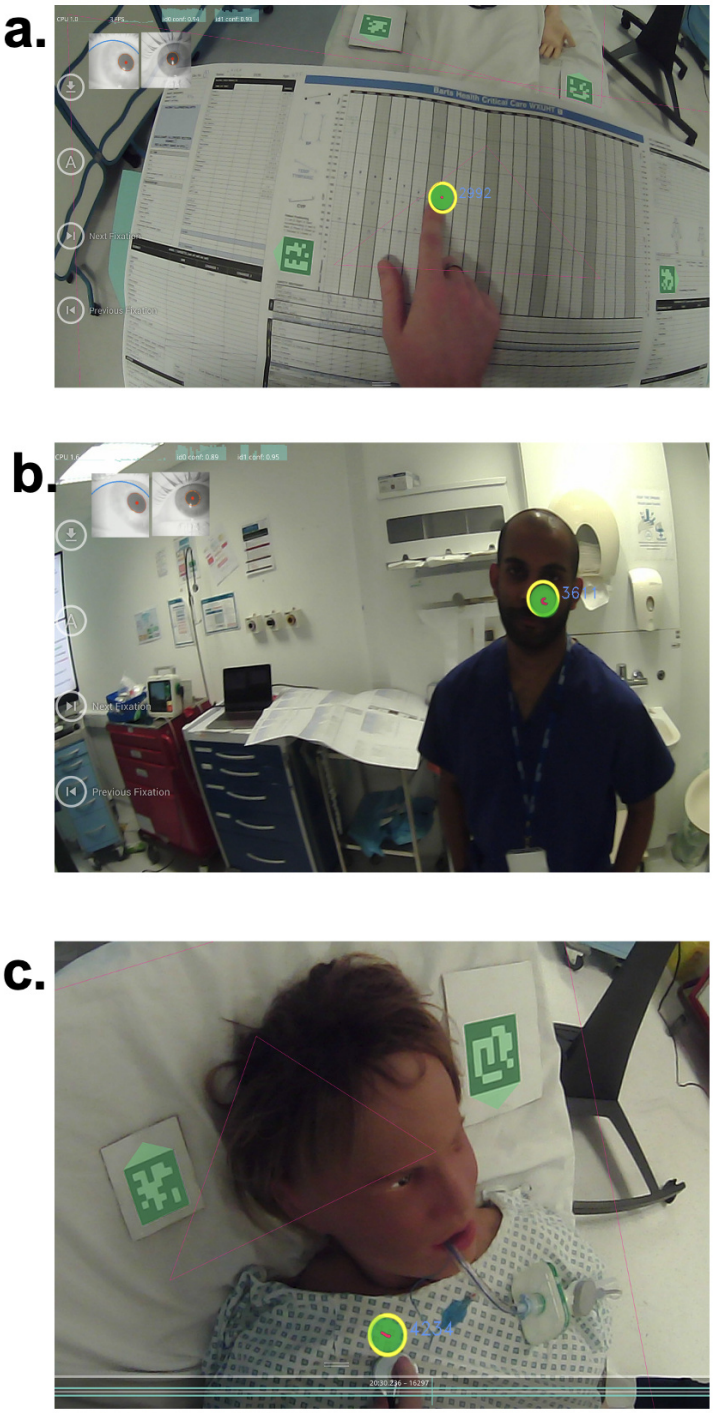


Fig. 3.7 **Fixations on various surfaces.** In (a), a fixation is shown over the data where the blood pressure is recorded on the ICU chart which happens to line up with the subject's finger. In (b), the subject is speaking to the bedside ICU nurse (experimenter role played by me according to a standardised script. In (c), the subject is examining the patient mannequin and a stethoscope can be seen at the bottom of the image with a fixation close to the point at which the subject is auscultating the mannequin's chest.

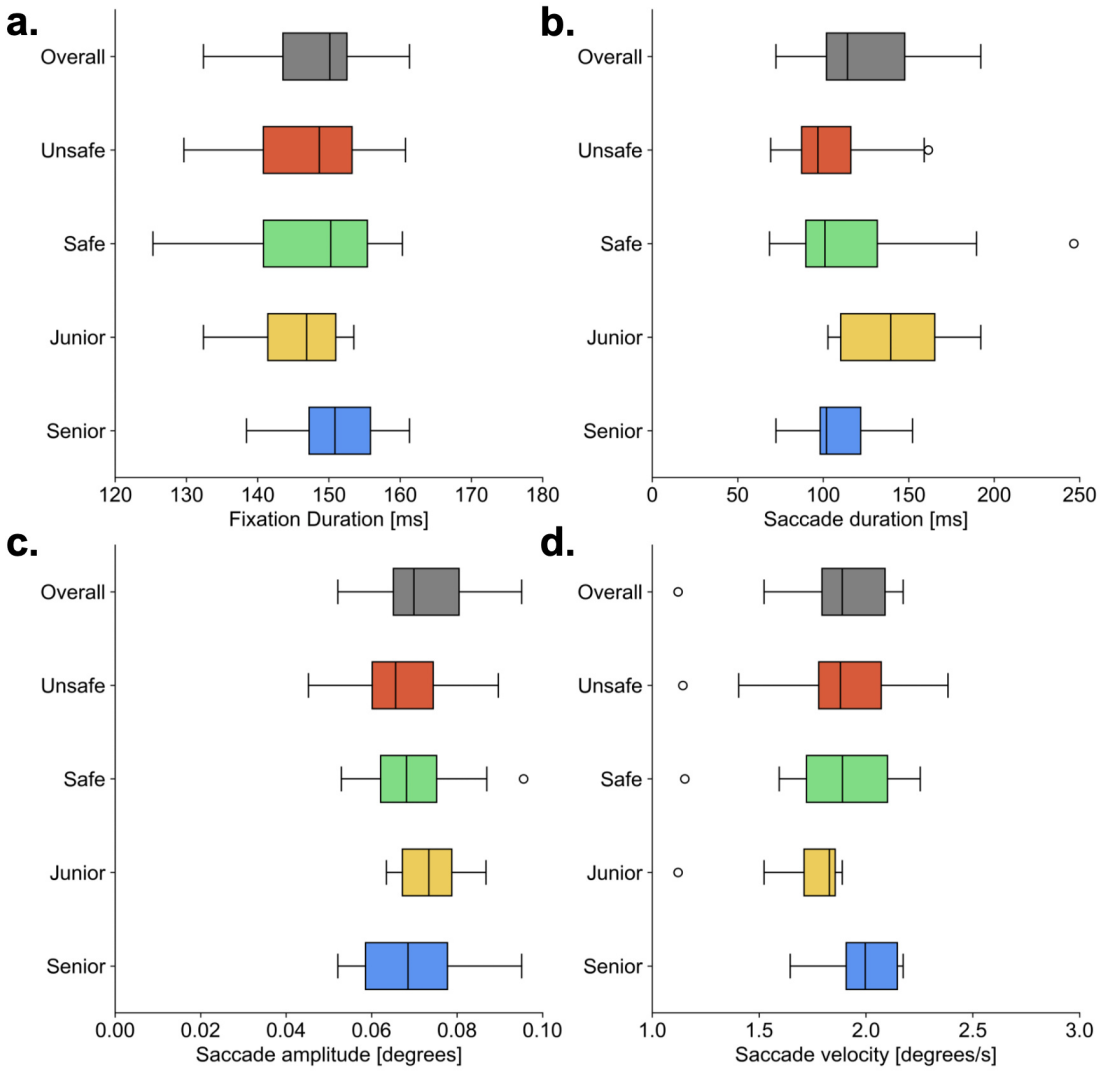


Fig. 3.8 **Fixation and saccade metrics stratified by trial safety status and experience level.**

outline represents the area taken by the ROI as a proportion of the total world-view over a single frame of video.

For all ROIs except the patient mannequin, there was a significantly higher actual than chance gaze proportion ($p<0.001$ for all comparisons except patient mannequin, see Figure 3.11). The ratio of actual to chance gaze was 6.5, 1.6, 12.5 and 1.3 respectively for the major ROIs (AI screen, ICU chart, vital signs monitor, patient mannequin). The ratio of actual to random gaze was 6.1, 4.2, 5.3 and 3.3 respectively for the XAI

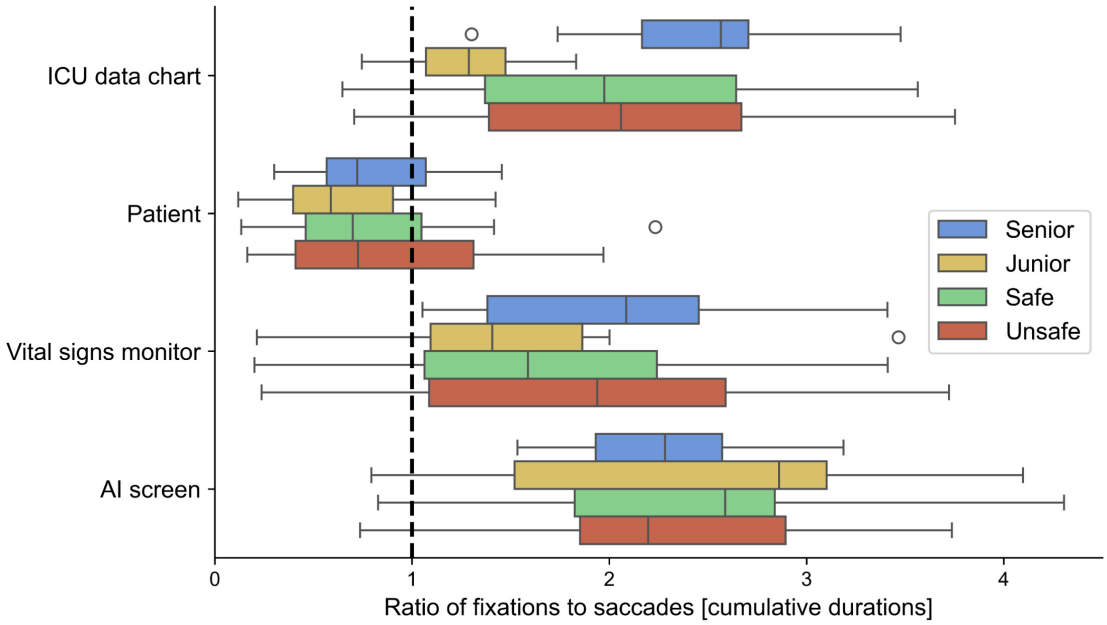


Fig. 3.9 **Fixation to saccade duration ratios stratified by trial safety status and experience level.**

ROIs (training examples, Q-value difference, mortality, feature importance) (see Figure 3.12).

3.3.4 Blink rate

Mean blink rate per minute is displayed in Figure 3.13 stratified by safety status of the scenario and experience level of the subject. Mean blink rate was lowest for the ICU data chart, highest for the AI screen and similar for patient mannequin and vital signs monitor. There was a significantly lower blink rate by junior clinicians on the patient mannequin (mean 9.1 junior versus 18.8 senior, p=0.017).

3.3.5 Pupil diameter

Pupil diameter was heavily confounded by ambient lighting and was therefore not used for formal analysis. An exemplar plot of pupil diameter over time for one subject is shown in Figure 3.14 with green lines indicating the start of a patient trial and pink

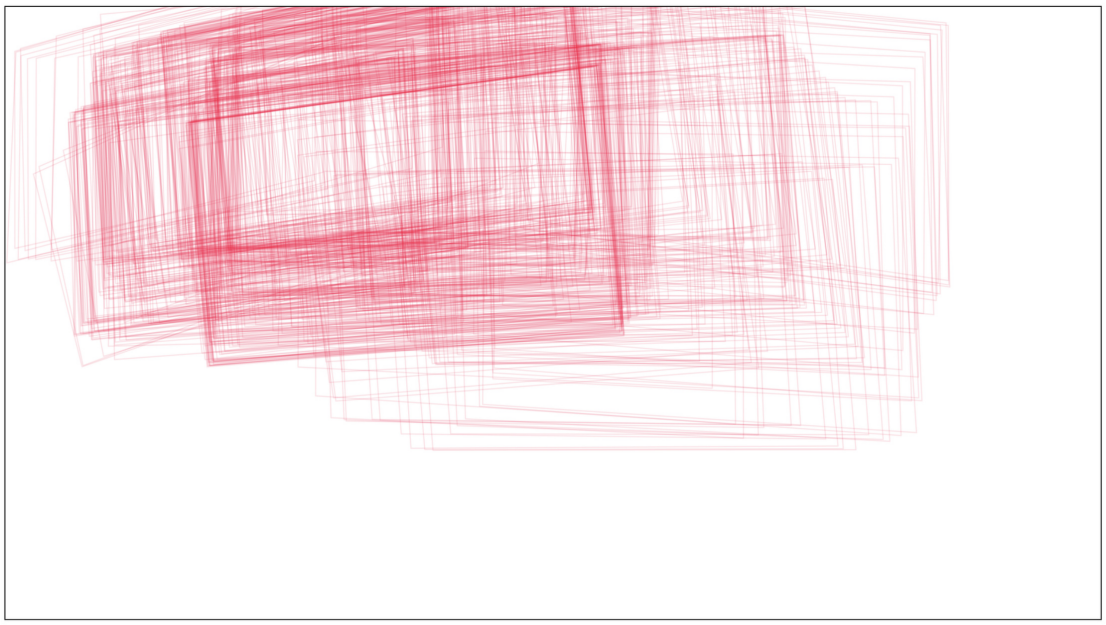


Fig. 3.10 **Position of AI screen ROI over 500 frames.** The corner coordinates (calculated by parsing raw transformation matrices of position) are joined to form a rectangle representing the ROI (in this case the AI screen). Are per ROI can then be calculated both in absolute terms and relative to the total world-view area. This example has been limited to 500 frames for illustrative purposes so that the lines can be made out easily.

lines marking the end. In between trials, the subject waited outside the simulation suite to be called back in. This area was more dimly lit and the association with increased pupil diameter can be clearly seen in the intervals between trials (i.e. the short regions between pink and green lines). The consistent region of low pupil diameter corresponds most strongly to the time spent assessing the patient mannequin which was quite well lit (nearly directly under one of the ceiling lights in the simulation suite).

3.4 Discussion

3.4.1 Summary of results

This chapter demonstrates the feasibility of performing free-movement eye-tracking on doctors within a high fidelity simulation suite as a means of evaluating their interaction

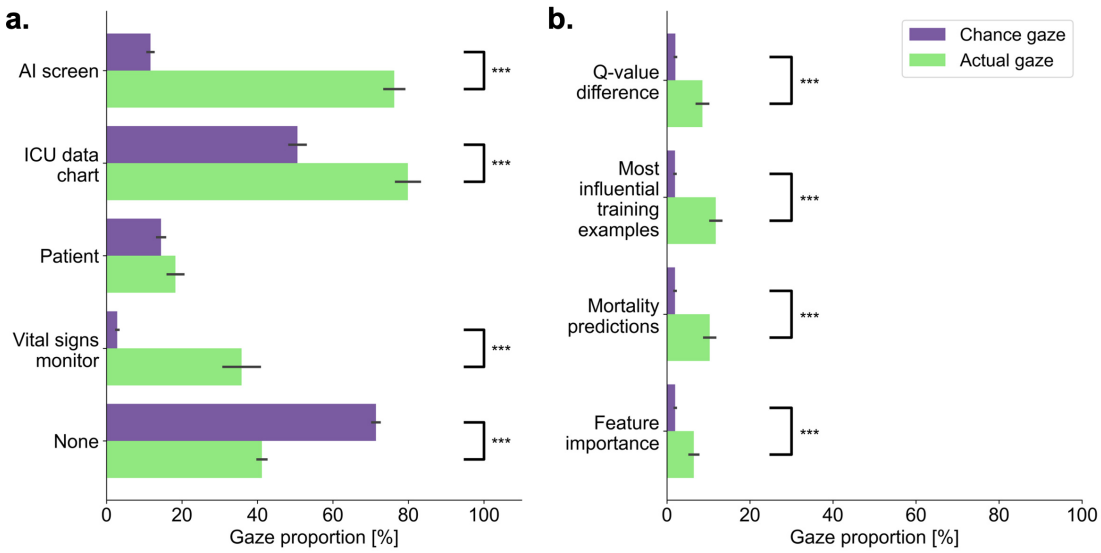


Fig. 3.11 **Gaze proportion per ROI.** Mean and SEM error bars. Chance gaze is the proportion expected based on the area occupied by an ROI within the visual field (i.e. if ROI takes up 50% of the screen, 50% of the time, we would expect gaze to fall within it 25% of the time by chance alone). Actual gaze is the proportion observed during the experiment. Significance levels are one star for $p<0.05$ and three stars for $p<0.001$ (based on Student's t-test).

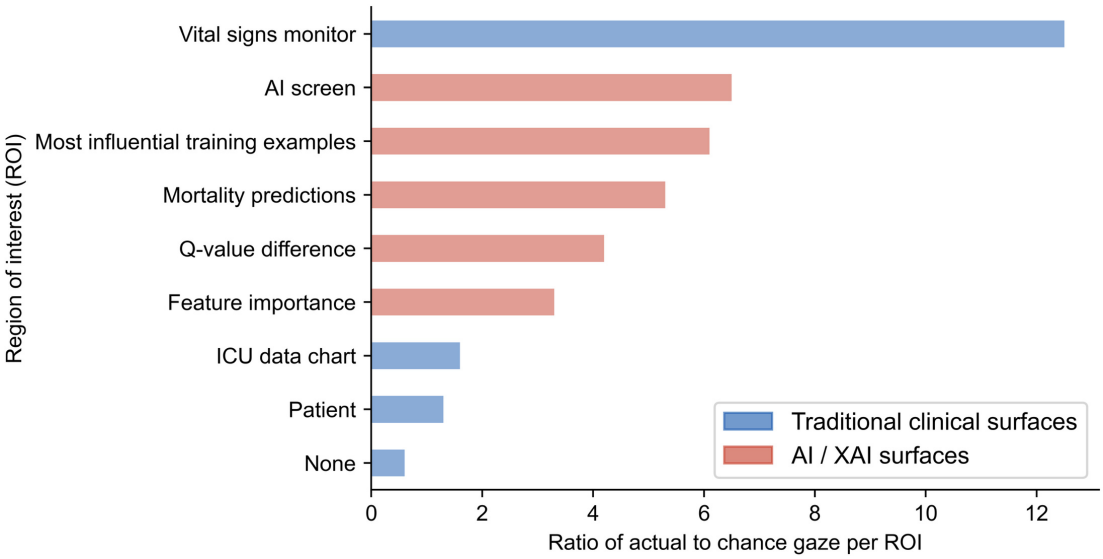


Fig. 3.12 **Ratio of actual to chance gaze per ROI.** Chance gaze is the proportion expected based on the area occupied by an ROI within the visual field. Actual gaze is the proportion observed during the experiment. Both are displayed separately in Figure S1, here the ratio of the two is displayed.

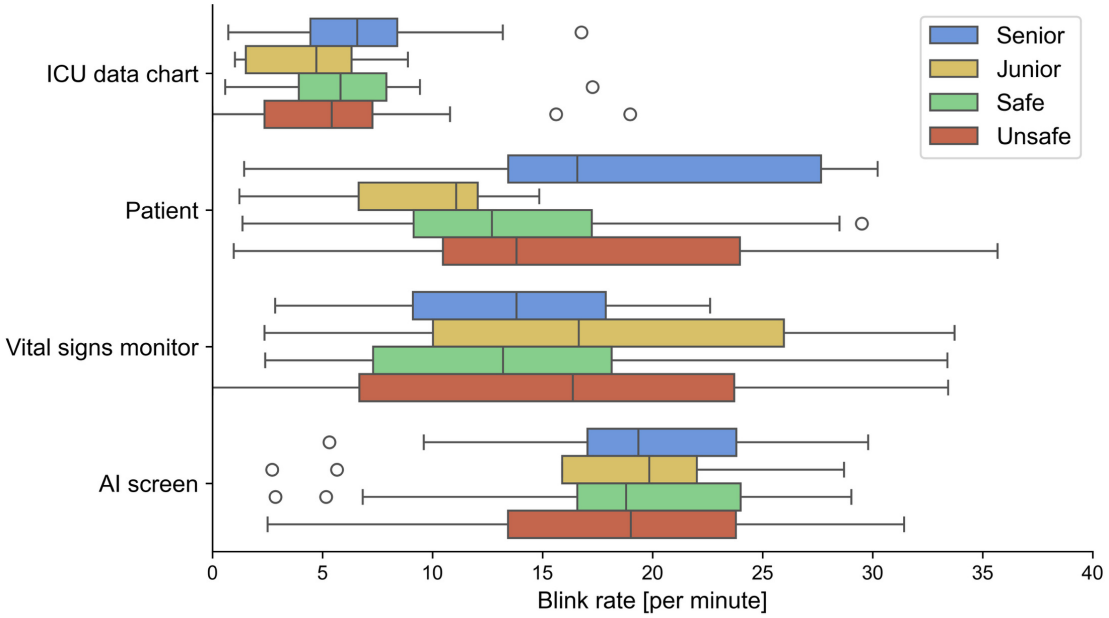


Fig. 3.13 **Blink rate stratified by trial safety status and experience level.**

with an XAI-driven decision support tool. I was able to successfully construct a pipeline for (i) acquiring eye-tracking data, (ii) conducting post-processing and (iii) automating scripts for analysis of common eye-tracking metrics. I also assessed a novel metric for visual real-estate that took into account size of the ROI. In summary, the main findings (non-XAI related in this chapter, see Chapter 4 for XAI findings, section 4.3) included that senior doctors had a slightly higher gaze speed than juniors while gaze distance expectedly correlated with task duration. Fixation and saccade metrics showed no significant differences based on experience level or safety status of the AI suggestion. Cognitive load (via the fixation to saccade ratio) was lowest on the patient mannequin (expectedly as it is not a real patient and therefore has less clinical information available), with senior doctors exhibiting higher load ratios on the ICU data chart, the most data rich of the ROIs.

Adjusted visual real-estate analysis revealed significant differences between actual and chance gaze proportions across all ROIs except for the patient mannequin. Again, this result for the patient mannequin is not necessarily surprising given the artificial

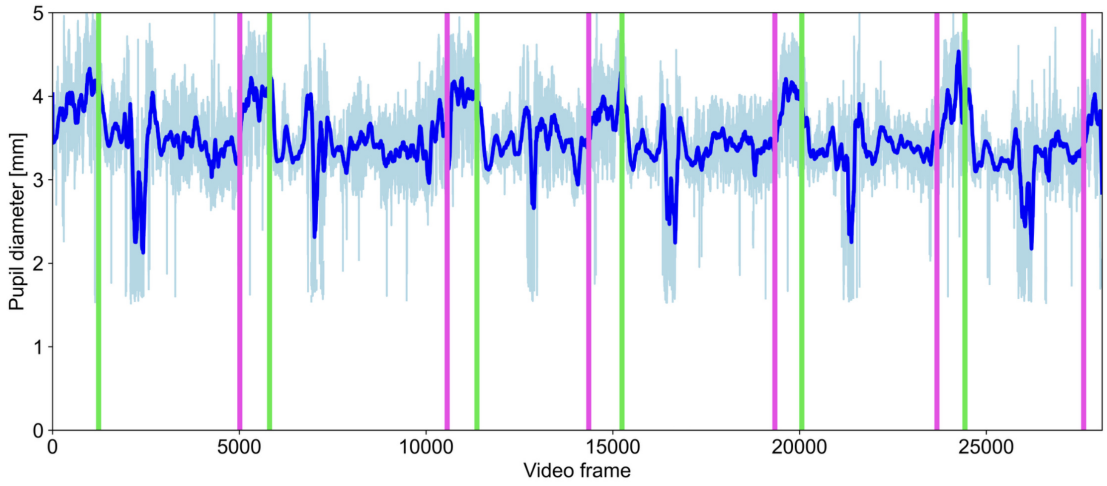


Fig. 3.14 **Pupil diameter over time for a single subject.** Green lines represent the start of each of the 6 trials. Purple lines represent the end of each of the 6 trials. The light blue shading shows the raw pupil diameter data with the dark blue line representing a 100 moving average of the same data.

nature of the simulation suite and might be different in a true clinical environment with a real patient.

3.4.2 Limitations

There are several practical limitations to eye-tracking as performed in this experiment. First, while the simulation suite is superior in fidelity to a vignette format, it still does not compare to the granularity of a real medical environment (the extensive clinical implications of this point are discussed in more detail in Chapter 4, see section 4.4). A simulation suite evaluation would be only an initial step in a multi-study pathway for assessing an AI-driven decision support tool that would also require real-world clinical studies. In this experiment, I was most concerned with the practical implications of performing eye-tracking in a simulation environment. The main limitations of my approach include issues that can be broadly categorised to hardware, software and the approach to analysis of the eye-tracking data.

From a hardware perspective, the eye-tracking glasses I used had to be continuously connected to the laptop for recording. While subjects were still able to walk around the simulation suite freely, this was clearly more uncomfortable for them than had they been wearing glasses alone and was directly responsible for technical failures in two cases (one recording failure due to actual disconnection of the wire between glasses and laptop and one failure of recording due to battery drainage in the laptop). One participant also declined to wear the eye-tracking glasses. Furthermore, subjects had to wear a backpack containing the laptop that the eye-tracking glasses were connected to which may also have contributed to discomfort. A recording was only actually saved to disk at the conclusion of each experiment which exposed data to significant risk if any problems (such as wire disconnection) occurred mid-experiment. Finally, the calibration step added at least 10 minutes to each experiment which had to be factored in when deciding how many subjects could be scheduled to undertake the experiment on a single day.

From a software perspective, poor pick-up of pupils in many subjects meant that the number contributing eye-tracking data to the analysis was far below the total number recruited into the experiment. This had implications for statistical power (see Chapter 4, section 4.4) which limited the conclusions I could draw from some of the analyses. It is also possible that bias could have been introduced if there was a systematic cause underlying the exclusion of participants. This did not seem to fall along gender or experience criteria - 13/19 male doctors had gaze data (68% male compared to 66% male in the overall sample), 6/19 senior doctors had gaze data (32% senior compared to 55% senior in the overall sample). Anecdotally, it was not easy to predict in advance which participants were likely to have poor pupil pickup. The restricted vertical dimensions of the world-view camera (ultra wide-screen) risked some ROIs being missed if, for example, the subject was especially tall and required extra time during the pre-calibration steps to ensure optimal positioning of the world-camera.

Pupil diameter was not used for further analysis owing to obvious confounding by ambient light. However, there are potential software solutions that can correct for ambient light and thereby potentially isolate pupil changes related to cognitive load alone. One such example is a proprietary algorithm known as the Index of Cognitive Activity [185].

From the analysis perspective, some time-consuming steps in the pipeline were not automated. For example, performing quality control by manually reviewing videos and the requirement for annotation of ROI boundary boxes. Given recent advances in AI-driven object recognition/segmentation from videos [186], it is certainly possible that these steps could be automated in future (or at least semi-automated and thereby considerably reduce the resource cost of performing eye-tracking within XAI evaluation studies). Lastly, audio recordings were not collected in this experiment but would have been a valuable source of additional data (potentially in conjunction with a think-aloud study protocol where subjects are encouraged to verbalise their thinking process). This audio data could have been auto-transcribed to text and used for a thematic analysis. There are examples in the literature that combine eye-tracking with a think-aloud methodology [187].

To apply eye-tracking in a real-world clinical environment would likely also require a more detailed consideration of ethical aspects of the study to account for inadvertent capture of other patients and staff into videos who might not necessarily have been directly consented. This issue would not be insurmountable as there are, for example, TV shows that have filmed in UK emergency departments and which used an opt out and blurring system for people who did not provide explicit consent to be filmed. The impact of the Hawthorne effect (i.e. a change in behaviour attributable to a subject's knowledge of being observed) might affect ecological validity and would need mitigation too.

3.4.3 Conclusion

Measuring a neuro-behavioural marker of human interaction with XAI addresses a key weakness of the self-reported data from the previous vignette experiment. However, two significant challenges remain before this approach can be scaled up and applied routinely for evaluating AI-driven decision support tools. First the fact that only 50% of subjects recruited into the simulation experiment were able to contribute eye-tracking data of sufficient quality places a major constraint on sample size recruitment and statistical power. The other challenge is the high resource cost of performing such studies which may be mitigated if more of the pipeline can be automated in future.

Chapter 4

Eye tracking of physicians interacting with safe and unsafe XAI

This chapter is closely based on two published manuscripts. I am first author on the first of these manuscripts and joint first author on the second manuscript (my supervisors are co-authors on both). All analyses presented in this chapter are my own.

Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations, published in npj Digital Medicine [188].

Safety of human-AI cooperative decision-making within intensive care: a physical simulation study, published in PLOS Digital Health [189].

4.1 Introduction

As discussed in Chapter 1, the medical setting is an example of a high risk decision environment with similarities to other high-stakes settings (e.g. nuclear power plants, military, airlines) [190, 191]. This contrasts with many XAI evaluations in the computer science literature focused on low-stakes decision environments (e.g. quizzes, Atari games) [192]. There are unfortunately only a limited number of XAI evaluations with clinical end users (i.e. the target audience) [193], few of which have assessed physician

behaviour within a high-fidelity (i.e. more realistic or life-like) setting. The previous chapter (see Chapter 3) established the feasibility of eye-tracking as a useful neuro-behavioural marker of human interaction with XAI. For eye-tracking to be effective, it should be deployed in a setting that closely resembles real-world clinical practice (i.e. as high fidelity as possible while still permitting experimental standardisation [194, 195]). Combining eye-tracking with such a setting allows for a more granular and detailed assessment of the physician-XAI interaction dynamic compared to existing medical XAI studies.

In this chapter, I tested four types of simultaneously presented AI explanations on clinicians in a high-fidelity simulation suite aiming to quantify how XAI influences prescription decision behaviour in a high-stakes decision environment. The primary research question was whether the attention placed on XAI is greater when clinicians encountered an unsafe versus safe AI suggestion (which might have suggested that the XAI is within the causal pathway used to disregard the unsafe AI advice). Two secondary questions were: (i) whether self-report data correlated with attention as proxied by eye-tracking (a lack of correlation would support the idea that self-reports are unreliable for evaluating XAI interaction), and (ii) whether physician attention profiles (i.e. their patterns of attention in specific situations or tasks) correlated with practice variation or the tendency to follow AI advice. Investigating these correlations is important because identifying individuals more or less likely to deviate in prescribing behaviour or follow AI advice could aid in developing personalised AI decision-support tools. For example, if eye-tracking data could identify physicians who are consistently two standard deviations away from their peers on prescribing decisions, some hospitals might want to target decision support tools that reduce unwarranted practice variation specifically to such users. This is especially true given that a common aim for clinical decision support tools is to reduce unwarranted practice variation [196]. Identification of physicians who are outliers might allow better individualisation of AI decision-support

tools. Or if eye tracking data were predictive of which physicians were more or less likely to heed AI advice, then AI decision support tools could be individualised to different explanation types on such users to learn if their adherence could be increased.

4.2 Methods

4.2.1 Piloting of simulated sepsis cases

This experiment was pitched at evaluating technology readiness level (TRL) 5-6 (basic validation of technology model or prototype in a relevant environment) [197]. Six simulated ICU patient scenarios were created. The rationale behind selecting these cases was to use typical clinical scenarios with sepsis common to the UK ICU experience (urosepsis, pneumonia, intra-abdominal sepsis secondary to GI perforation, COVID-19, necrotising fasciitis, endocarditis). The physicians would then perform an abbreviated virtual ward round within the simulation suite. Prior to the experiment, the six simulated scenarios (designed by the three doctors, of which I was one, in our investigator team) were piloted on seven doctors from the Critical Care Research Group of Imperial College London. Each doctor reviewed a subset (4 of the 6) cases. For each case, there was a safe or unsafe AI suggestion. Doctors in this pilot were asked to rate whether they would intervene to stop the treatment suggestion if it were to be automatically administered. Agreement was calculated by the total proportion of doctors that would correctly stop an unsafe AI suggestion and correctly not stop a safe AI suggestion. With seven doctors reviewing four cases each and evaluating two treatment suggestions per case, there were 56 data points in total, of which 52 were 'correct' resulting in 93% overall agreement with the safe/unsafe ratings. This was deemed satisfactory to proceed with the scenarios for the main experiment.

4.2.2 Simulation experiment

Physicians who participated in the experiment first completed a standardised experiment briefing (see appendix E) and a pre-experiment questionnaire on demographics and attitudes towards AI (see appendix E). They were then taken into the simulation suite for orientation before conducting the eye-tracking calibration exercises (see Chapter 3 Methods, section 3.2). Eye-tracking was conducted throughout the experiment as detailed in the Chapter 3 Methods, section 3.2). I played the role of the bedside ICU nurse throughout all scenarios. Within each of the six scenarios, physicians were tasked with performing an assessment, which included reviewing the available patient data and examining the patient mannequin. Afterward, they were asked by the bedside nurse for a fluid and vasopressor prescription for the next hour of the patient's admission. Furthermore, they were asked to rate the following: (i) their confidence in the prescription decision being optimal for the patient on a 10-point Likert scale and (ii) whether (or not) they would like to seek a second opinion from a senior physician (or, in the case of senior physicians, another senior physician). Responses from the physician were recorded by a member of the research team not taking part in the experiment.

Physicians were then shown the AI recommendation along with four simultaneously presented explanations on a large display adjacent to the patient bed. They were subsequently asked to confirm or change their prescription doses. To mitigate the potential negative impact of unsafe AI suggestions on a doctor's confidence in future scenarios, I (i) ensured that the first scenario any physician encountered was always a safe scenario and (ii) used a trial matrix that had different physicians encounter the unsafe scenarios in different orders (see Figure E.1).

They were then asked for the following (i) their agreement with the AI suggestion (on a 5-point Likert scale), (ii) their confidence in the updated prescription decision they had made (on a 10-point Likert scale), (iii) whether or not they would like to seek

a second opinion from a senior physician (or in the case of senior physicians, another senior physician) and (iv) whether (or not) they would stop the AI recommendation if, hypothetically, it were to be automatically administered to the patient. The aim of this final question was to draw out the nuance between merely disagreeing with an AI suggestion but not necessarily considering it a danger to patient safety. The nuance between these two questions was very clearly introduced and explained to physicians during the pre-experiment briefing.

Influence of AI was calculated on a continuous scale from 0 (completely ignoring advice) to 1 (completely relying on advice) using the formula [(final estimate—initial estimate) / (advice – initial estimate)] per work by Yaniv et al. [141]. The analysis was also repeated using an alternative metric: the distance between a physician's final prescription (after viewing the AI recommendation) and the value of the AI recommendation for any given trial/scenario. A higher distance would suggest that the physician was less adherent to the AI advice, and vice versa.

4.2.3 Experiment conditions

Each of the six patient scenarios could be encountered by a physician under one of two conditions (pseudo-randomised order per trial matrix, see appendix E): safe AI recommendation (four cases) or unsafe AI recommendation (two cases). The categorisation of recommendations to safe or unsafe was based on extreme over- or under-dosing of fluid and vasopressor as per previous work from our group [198]. The unsafe arm was further divided into two types: standard unsafe and incentivised unsafe. In the latter, after the physician had reported whether or not they would stop the AI recommendation if it were automatically administered, I (in my role as the bedside nurse) was permitted three attempts (all following a standardised script, see Table 4.1) to verbally try and convince them to change their mind. The rationale for the incentivised condition was to assess how human interactions affect AI-supported

decision-making. After all, the ICU is a multi-disciplinary team environment where physicians would very likely encounter AI-driven recommendations during ward rounds or other settings where clinicians interact and have the potential to influence (i.e. bias) each other.

Theme	**If participant *WOULD* override unsafe AI suggestion**	**If participant *WOULD NOT* override unsafe AI suggestion**
Volume of patient data	"But hasn't this AI seen thousands more patients than we have? Are you really sure you'd override it?"	"But how do we know that the AI has seen a patient like the one we've got here? Are you really sure you'd allow this suggestion to be acted on?"
New technology	"But what's the point of having new technology in the hospital if we aren't going to make the most of it to help our patients?"	"But do we really understand how this AI was designed or made though?"
Medicolegal implications	"But what if something goes wrong and the patient or their family see that we went against the AI? How protected or vulnerable would we be?"	"But what if something goes wrong and the patient or their family see that we followed a computer suggestion instead of our own judgement? How protected or vulnerable would we be?"

Table 4.1 **Nurse challenge verbal prompts.**

4.2.4 Explanation modalities

The AI recommendations themselves were synthetic given that the purpose of this experiment was to test interaction dynamics between physician and AI (rather than a specific AI system per se). The explanations for both safe and unsafe conditions in the same patient scenario were identical. If I had instead varied the explanations between safe and unsafe conditions, this would introduce another factor (explanation quality) which would have inadvertently created more arms in the experiment (e.g. 12 arms in the experiment by Shafti and colleagues; AI performance good/bad (2) x AI explanation good/poor/none (3) x explanation in training (2) = 2x3x2 = 12 arms) [150]. I felt the simpler and more intuitive choice for this early exploratory study was therefore to not vary the explanation between safe and unsafe scenarios but this is acknowledged as a limitation (covered in greater detail in the discussion, see section 4.4).

Four types of explanations were constructed (all based on realistic explanation types that have been applied in reinforcement learning). As described in Chapter 1 (see section 1.2), offline reinforcement learning is a variant of reinforcement learning where an agent learns from a fixed dataset of previous interactions, consisting of states, actions, and rewards, without real-time environmental feedback. This approach aims at maximising cumulative rewards based on historical data, enabling decision-making in situations where direct interaction is impractical or risky.

The first explanation type was the Q-value difference. This approach leveraged the fact that once training is complete, the reinforcement learning (RL) Q table contains Q values for any given state-action pair. The optimal action is the one with the highest Q value for any given state. However, the difference between this highest Q value and the alternatives might be small or large. If large, then there is much higher anticipated value from following the recommended action compared to an alternative. This was

dichotomised arbitrarily from a continuous value to make interpretation more simple for clinical users not familiar with the nuances of reinforcement learning and Q-values.

The second explanation type used mortality predictions. This approach leveraged the fact that mortality can be predicted for any given state in the RL state space. Therefore the impact of different dosing strategies that might result in transition to alternative states with different predicted mortalities can be displayed to the subject to highlight how alternate strategies might change the risk of death.

The third explanation type used was feature importance. This approach used the fact that the state space for RL based sepsis algorithms is commonly constructed using a k-means clustering algorithm to enable dimensionality reduction. Per the description in the Chapter 2 Methods (see section 2.2), this was exploited to rank features in terms of their proximity to the cluster centroid (or average state feature values) given that the archetypal patient for whom an RL agent policy action would most apply is a patient who is most typical of that state. Subjects could then be shown the top five ranked features contributing to state assignment.

The fourth explanation type concerned the most influential training examples. This approach used the fact that the difference in Q values for any given state-action pair between one iteration of Q-learning and the previous iteration reflects how valuable the currently seen training episode is for learning the optimal action (for any given state). This is similar to how instance-based explanation used in deep learning for imaging XAI where the explanation consists of showing similar image instances to the training instance to explain why a particular image classification has been made, see Figure 4.1 [199]. The algorithm I created is shown below in Algorithm 1.

A natural criticism that arises is whether the episodes with the largest Q-differences are heavily impacted by the order in which training episodes are seen. I evaluated this by adjusting the code to utilise parallelised multi-core processing, aggregating the results of multiple runs to compute combined logs and summaries of Q-learning

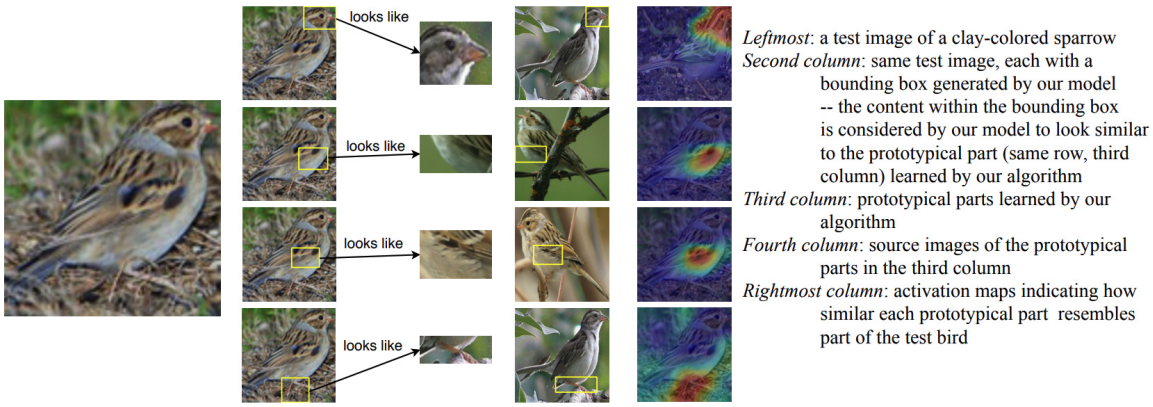


Fig. 4.1 **Example of instance-based XAI in imaging.** Image of a clay colored sparrow and how parts of it look like some learned prototypical parts of a clay colored sparrow used to classify the bird's species. Reproduced with kind permission from Professor Cynthia Rudin's paper on explainable deep learning for imaging using instance-based explanations [199].

outcomes. Two metrics were assessed: (i) the ***convergence metric*** represents the average proportion of times that the most frequent episode appears for each state-action pair across all runs up to that point. It measures the global dominance of specific episodes over the entire history of the algorithm's execution. And (ii) the ***Jaccard similarity*** measures the average overlap of episodes for each state-action pair between two consecutive runs. It quantifies the local consistency of the algorithm's

Algorithm 1 Q-Learning with Influential Episode Tracking.

```
create empty Q table and Q-difference table (both indexed by Q(S,A) tuple)
for 500,000 episodes do
    select random training episode
    for each time-step in episode do
        perform Q-learning update
        if Q difference is among top 3 for this state-action tuple then
            Update Q-difference table with this episode's ID
        end if
    end for
end for
return a dictionary of top 3 influential episodes per Q(S,A) tuple
```

behaviour from one iteration to the next, independent of the overall history. The formal definitions for both are available in appendix E.

The data for both metrics are shown in Figure 4.2 and suggests that the algorithm is repeatedly identifying a core set of important episodes for each state-action pair. In Figure 4.2a, the episode dominance decreases from just over 70% to approximately 58% but stabilises at this level. In Figure 4.2b, the Jaccard similarity remains constant between comparisons of consecutive runs (i.e. there is around 58% similarity of most influential episode between any two consecutive runs). The implication of this is that the most influential episodes are not identified at random but instead converge to a core set encompassing just over half that are not affected by the order in which training occurs. For our purposes, this is reasonable enough to use as an explanation modality as we are not necessarily interested in identifying the definitively most influential episodes but rather proving that there is consistency over time.

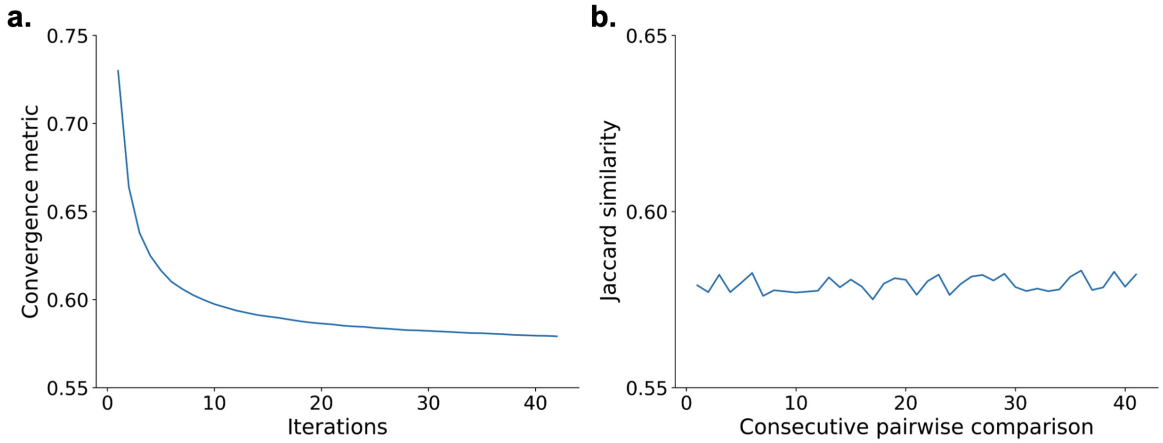


Fig. 4.2 **Consistency assessment for instance-based XAI.** An assessment of consistency between episodes identified as most influential over 42 runs of the algorithm listed in 1. In (a), the convergence metric is displayed which represents the average proportion of times the most frequent episode appears for each state-action pair across all runs up to that point. In (b), Jaccard similarity is shown which measures the average overlap of episodes for each state-action pair between two consecutive runs.

4.2.5 Participant recruitment and experiment conduct

ICU physicians were recruited as participants using both convenience sampling and targeted advertising to a local NHS trust (Imperial College Healthcare NHS Trust). The following inclusion criteria were used: (i) practising physician, (ii) has worked for at least 2 months in an adult ICU, (iii) currently works in ICU or has worked in ICU within the last 6 months. Electronically recorded informed consent was obtained from all physicians. Each experiment lasted approximately 60 minutes in total and participants were compensated for their time. The study was approved by the Research Governance and Integrity Team (RGIT) at Imperial College London and the Health Research Authority (Ref: 22/HRA/1610). All physicians were debriefed at the end of the study on the synthetic nature of the AI suggestions so as not to bias their opinions of future interactions with AI-driven decision support tools.

4.3 Results

Nineteen ICU physicians with eye-tracking data available were included (13 male, 6 female). Mean physician age was 33 years (standard deviation (SD) 6 years). Mean ICU experience was 3.6 years (SD 4 years, range 2 months to 14 years), see Figure 4.3a. The attitudes to AI from the pre-experiment questionnaire (see Figure 4.3b) followed a similar pattern to the study population from the vignette experiment (see section 2.3). All physicians completed the task successfully with a mean scenario completion time of 5.4 minutes (SD 1.2 minutes). The mean practice variation for fluid was 217 ml/hr (SD 205 ml/hr) and for vasopressor was 0.04 mcg/kg/min (SD 0.04 mcg/kg/min).

4.3.1 Eye-tracking metrics

There were significantly more gaze fixations for the AI screen during unsafe versus safe scenarios (mean 962 [95% confidence interval (CI) 861 to 1063] vs. 704 [95% CI

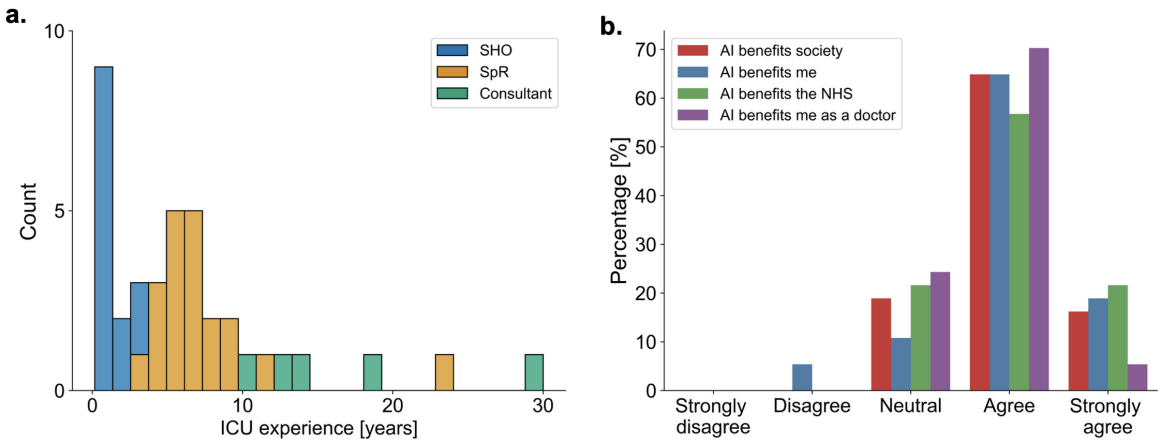


Fig. 4.3 **Physician experience and attitudes.** Physician experience levels in (a) and answers to AI attitude questions in (b).

593 to 814] respectively, p=0.002, Figure 4.4b). However, there were no appreciable significant differences in the number of gaze fixations between the simultaneously presented different XAI modalities for either safe (mean 75 [95% CI 59 to 90]) or unsafe (mean 94 [95% CI 77 to 111]) scenario (Figure 4.4c) as the difference for the AI screen was driven almost entirely by the AI recommendation. However, a post-hoc power calculation suggested insufficient power to detect a significant difference with the recruited sample size. Gaze fixations stratified by *both* safety status and seniority of ICU physician are shown in Figure 4.5. The difference between number of gaze fixations on the AI screen between safe and unsafe scenarios was only significant for senior ICU physicians.

When comparing the number of gaze fixations pre- and post-AI reveal, there was no significant difference on non-AI ROIs prior to the AI reveal (see Figure 4.6a). This was expected as the safety status of each case was, by definition, based on the nature of the revealed AI suggestion (i.e. physicians would not be able to tell that a case was unsafe until the AI-reveal). However, of more interest was the fact that no significant difference in gaze fixations was observed on these ROIs even post AI-reveal (see Figure 4.6b). In other words, having been confronted with an unsafe AI suggestion, physicians

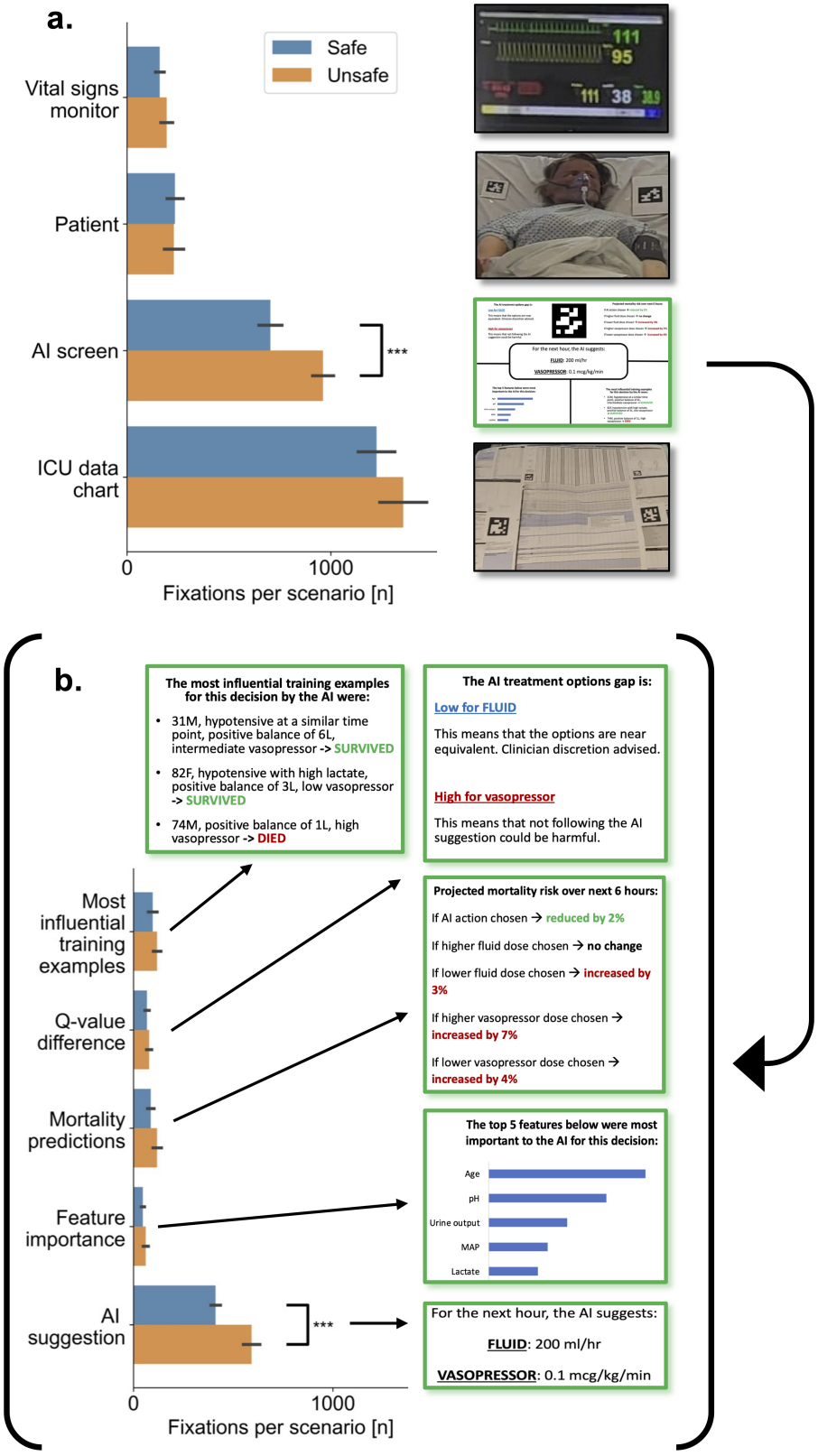


Fig. 4.4 **Gaze fixations per scenario by safety status of scenario.** Fixation values are mean with SEM error bars. The AI screen in (a) is a super-set encompassing the AI recommendation and all four XAI ROIs (green boxes) shown in (b). Significance levels are one star for $p<0.05$ and three stars for $p<0.001$ (based on Student's t-test).

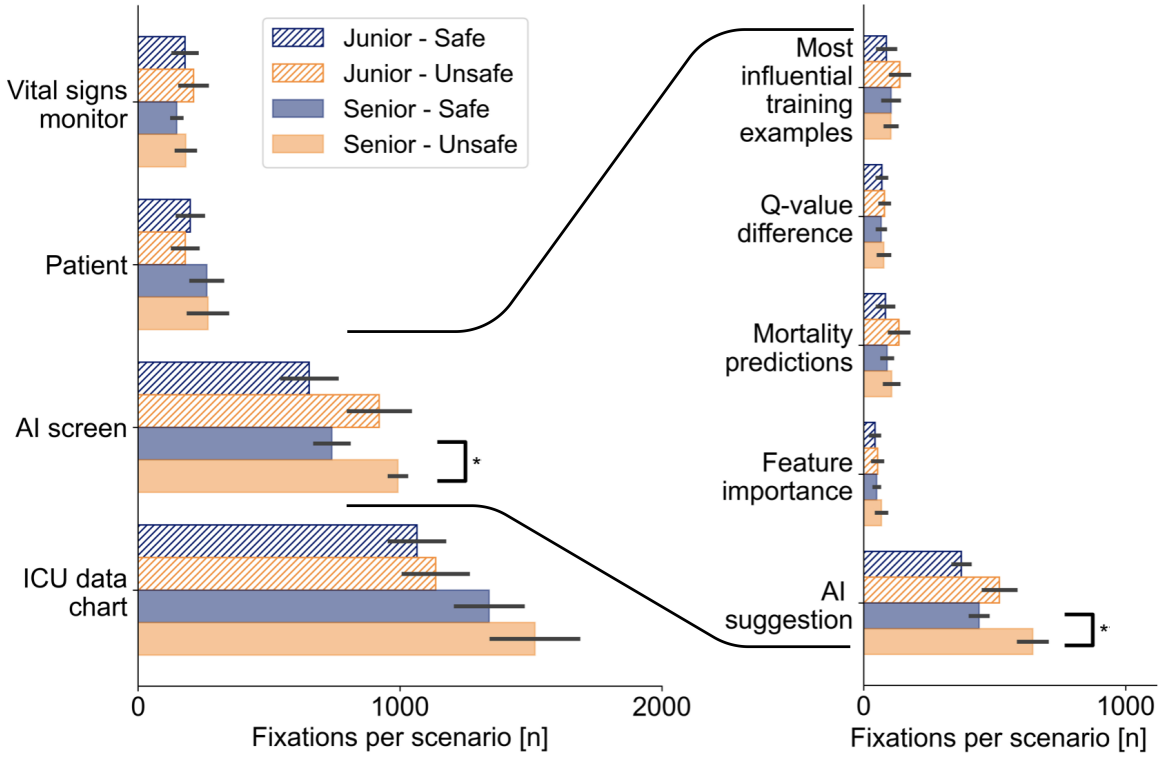


Fig. 4.5 **Fixations per scenario by safety status of scenario and experience level of doctor.** Mean and SEM error bars. The AI screen is a super-set encompassing the AI suggestion and all four XAI ROIs. Significance levels are one star for $p<0.05$ and three stars for $p<0.001$ (based on Student's t-test).

did not look back to the ICU data chart, patient mannequin or vital signs monitor to try and understand why the (unsafe) AI suggestion was so extreme.

Mean fixation duration was lowest for the patient mannequin (135 ms, SD 8 ms) and similar for all other surfaces (see Figure 4.7a). Mean blink rate was lowest for the ICU chart (6.1 blinks per minute (bpm), SD 4.1), similar for both vital signs monitor and patient mannequin (mean 15.2 bpm and 14.7 bpm, SD 8.7 and 9.2 respectively) and notably higher for the AI screen (mean 19.9 bpm, SD 10.7), see Figure 4.7b. When comparing all conventional clinical ROIs (chart, patient mannequin, monitor; blue bars in Figure 2b) to all AI ROIs (including XAIs; red bars in Figure 2b), there was a significantly lower mean blink rate on the conventional clinical ROIs than the AI ROIs (12.0 bpm vs. 23.7 bpm, $p=0.002$ by independent T-test).

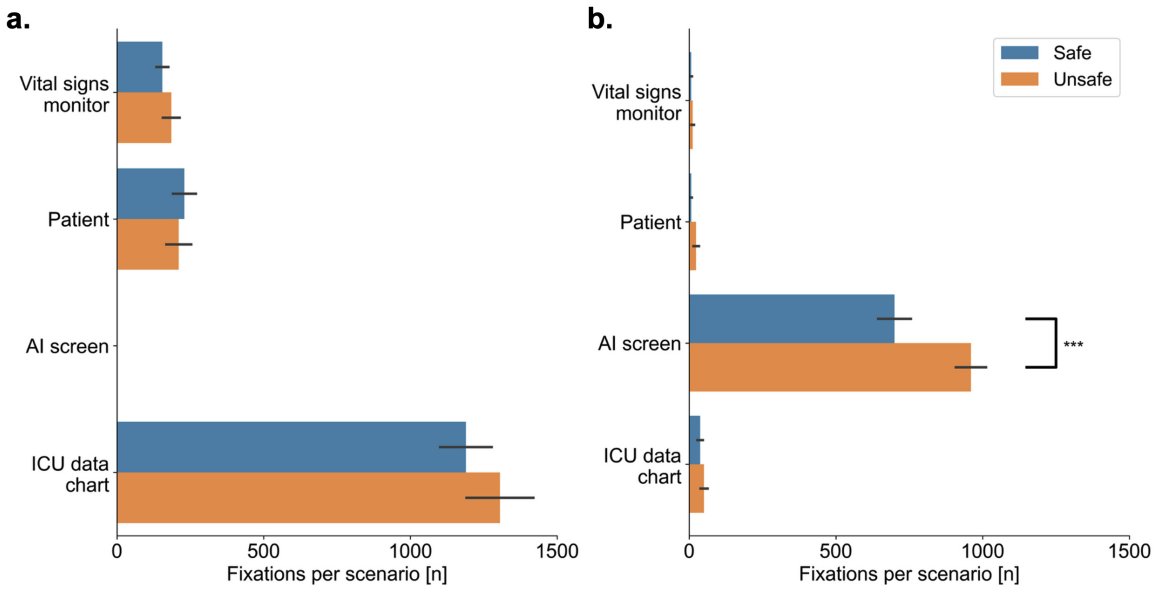


Fig. 4.6 **Gaze fixations per scenario.** Gaze fixations per scenario are shown both (a) pre-AI reveal and (b) post-AI reveal stratified by safety status of the AI recommendation. Significance levels are one star for $p<0.05$ and three stars for $p<0.001$.

4.3.2 Clinical practice variation among physicians

Practice variation was defined as the dose distance of an individual physician from the average of all pre-AI reveal physician prescriptions in any given trial/scenario (higher dose distance suggesting that the physician was more of an outlier in that particular trial and vice versa). When assessing blink rate on the clinical ROIs, there was weak positive correlation for fluid (r=0.20, p=0.04, Figure 4.8a) and weak negative correlation for vasopressor (r=-0.07, p=0.44, Figure 4.8b). When assessing the number of gaze fixations on the clinical ROIs, there was weak negative correlation for fluid (r=-0.19, p=0.05, Figure 4.8c) and weak negative correlation for vasopressor (r=-0.08, p=0.42, Figure 4.8d). Overall, there was no strong pattern between the eye-tracking metrics (number of gaze fixations and blink rate) and the degree of clinical practice variation (Figure 4.8).

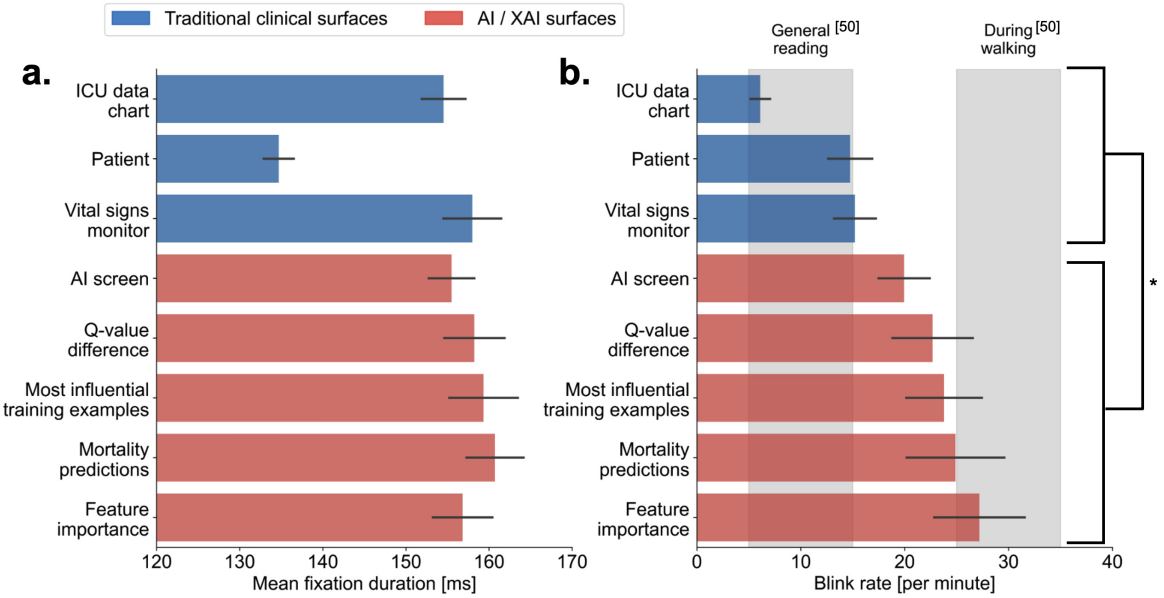


Fig. 4.7 **Mean fixation duration and blink rate by region of interest.** Mean and SEM error bars. Blue bars are traditional clinical surfaces while red bars are AI / XAI surfaces. In (a), fixation duration is shown while in (b), blink rate is shown. The grey shaded regions show typical blink rate estimates associated with either general reading (requiring more concentration) and walking (less concentration) as per Chidi-Egboka et al. [169]. Significance levels are one star for $p<0.05$ and three stars for $p<0.001$ (based on Student's t-test).

4.3.3 Influence of AI and simultaneously provided XAI

There was no strong evidence of correlation between blink rate and influence of AI regardless of safety status or drug (see Figure 4.9). Nor was there strong evidence of correlation between number of gaze fixations and influence of AI regardless of safety status or drug (see Figure 4.10). There was no change when instead looking at the distance between a physician's final prescription (having had the opportunity to view the AI recommendation) and the value of the AI recommendation rather than influence of AI per se (see Figures 4.11 and 4.12). The absolute distance between final prescription and AI recommendation for fluid was 171 mL/hr and 155 mL/h respectively for safe and unsafe conditions, $p=0.68$. For vasopressor, the distance between a physician's final prescription and the value of the AI recommendation was

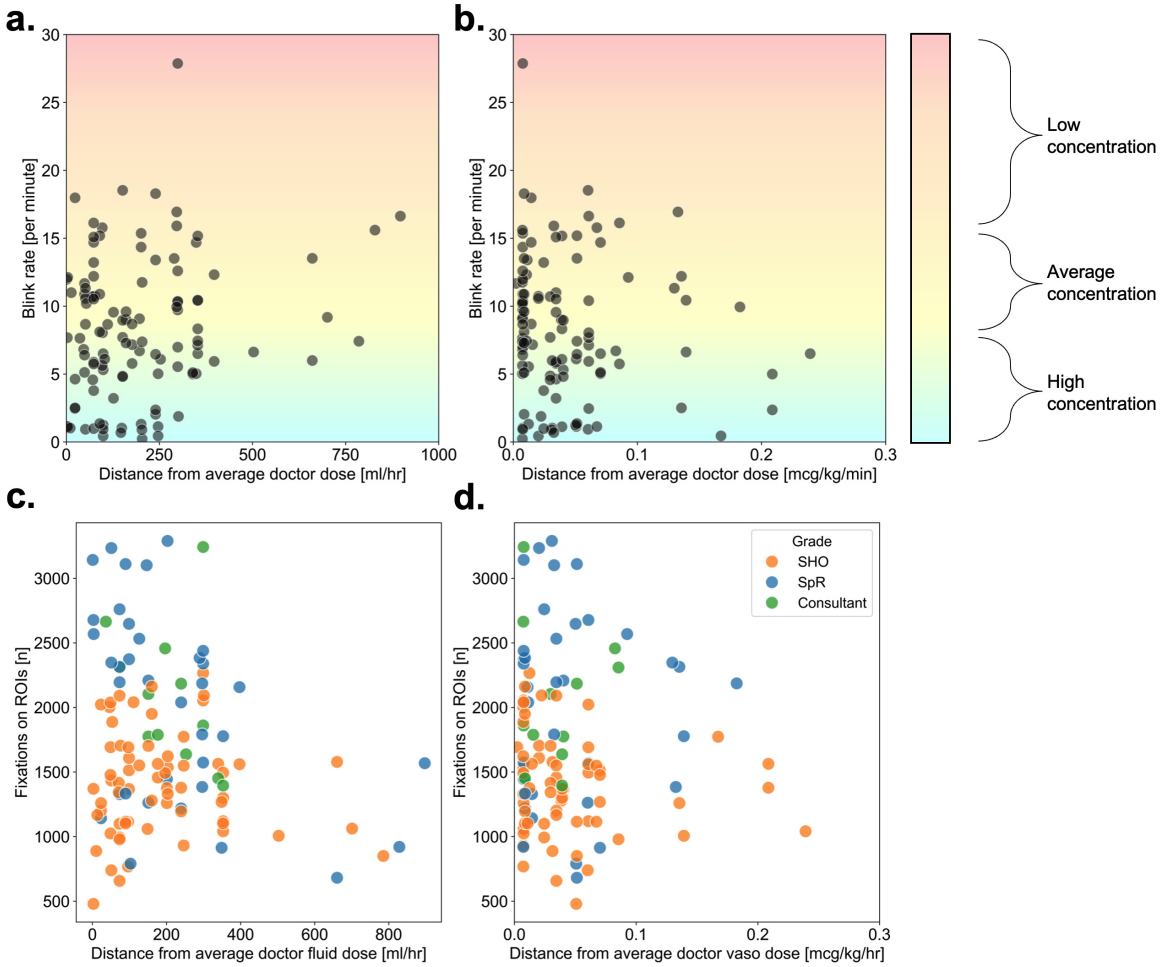


Fig. 4.8 **Correlation between practice variation and eye tracking metrics, for both fluid and vasopressor.** In (a), blink rate versus practice variation for fluid. In (b), blink rate versus practice variation for vasopressor. In (c), number of gaze fixations versus practice variation for fluid. In (d), number of gaze fixations versus practice variation for vasopressor. The degree of practice variation is defined by the per trial distance between a physician's prescription and the average prescription dose of all physicians. The background shading of plots (a) and (b) gives an indicator for expected blink rate depending on level of focus/concentration. Typical spontaneous blink rate is around 8-15 with lower levels suggesting higher focus/concentration and vice versa.

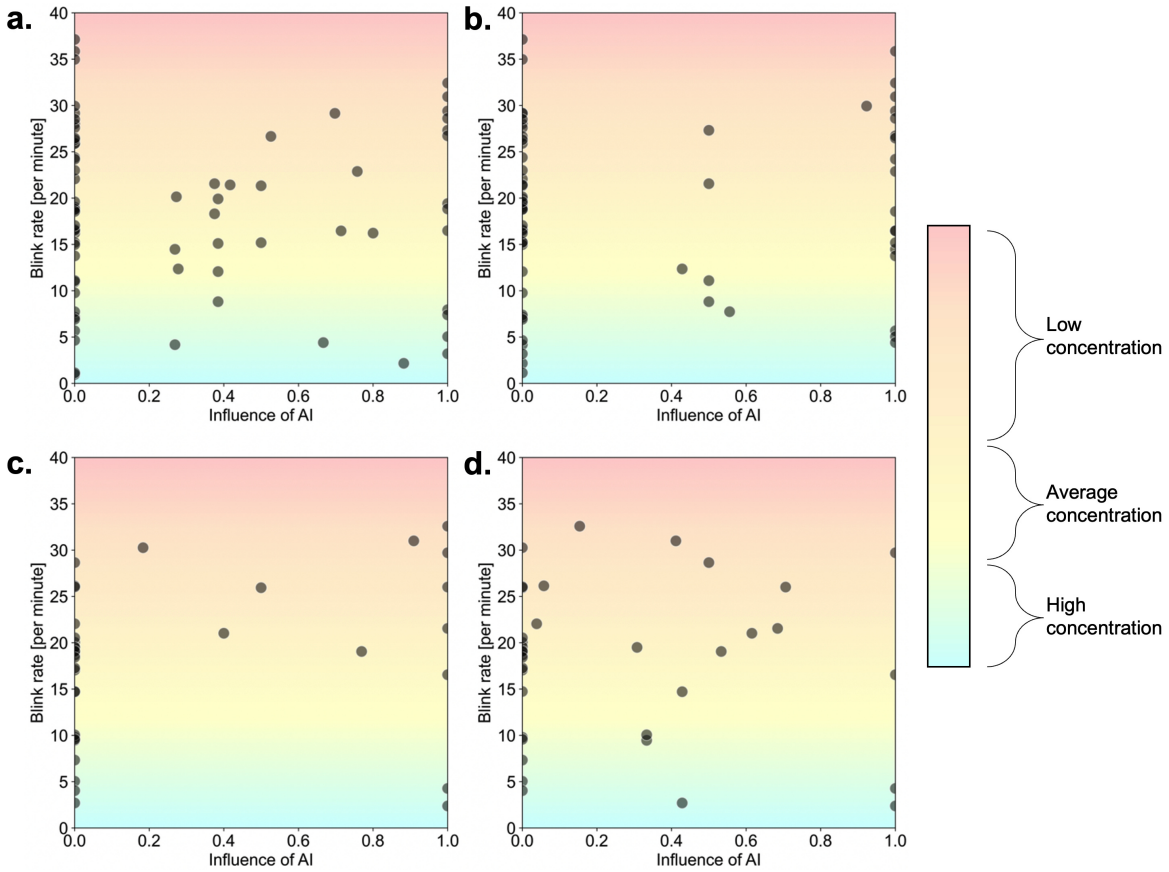


Fig. 4.9 **Blink rate by influence of AI for both fluid and vasopressor and for both safe and unsafe AI suggestions.** In (a), blink rate versus influence of AI for fluid with safe AI recommendation. In (b), blink rate versus influence of AI for vasopressor with safe AI recommendation. In (c), blink rate versus influence of AI for fluid with unsafe AI recommendation. In (d), blink rate versus influence of AI for vasopressor with unsafe AI recommendation. Typical spontaneous blink rate is around 8-15 with lower levels suggesting higher focus/concentration and vice versa.

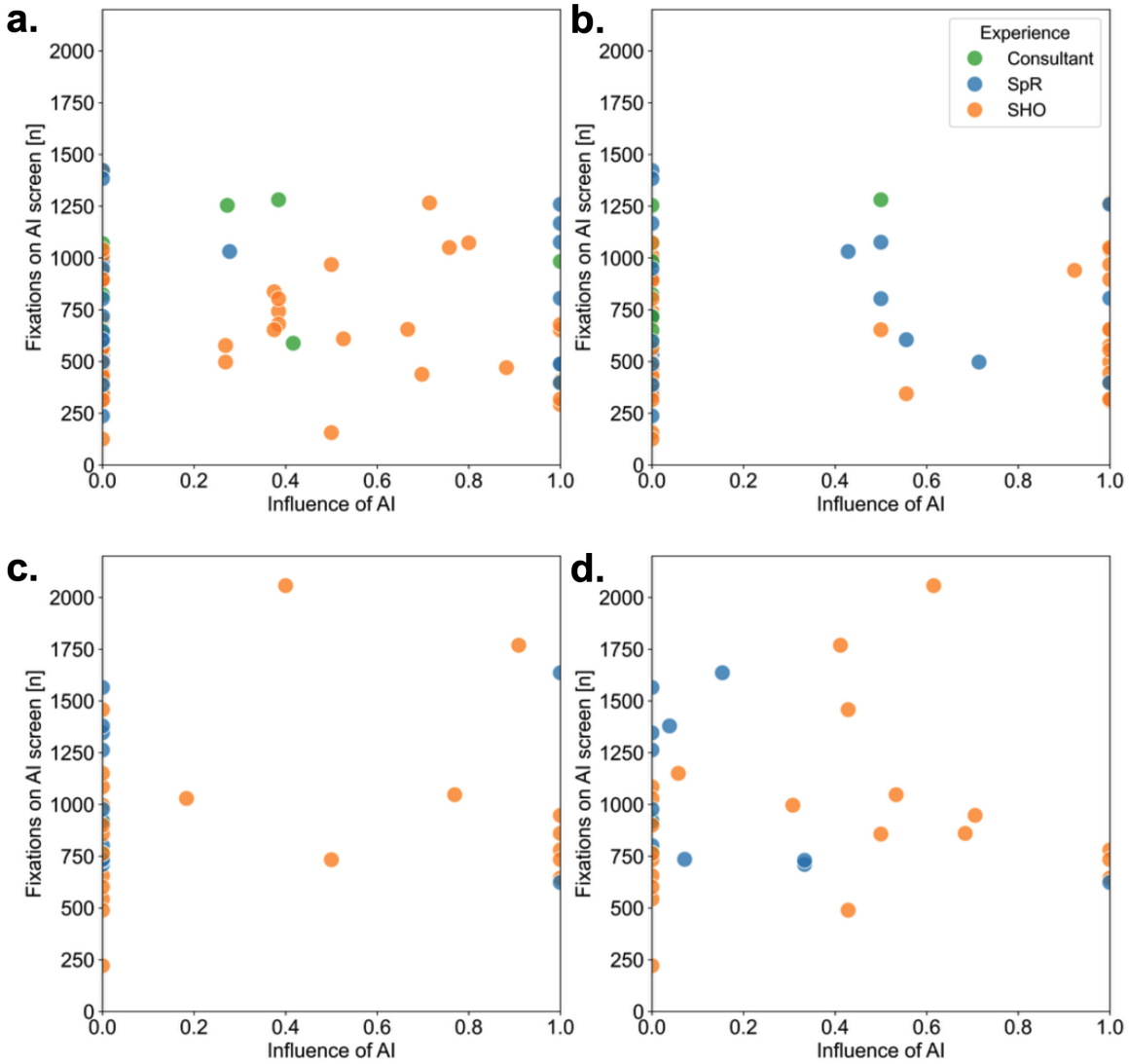


Fig. 4.10 **Number of gaze fixations by influence of AI for both fluid and vasopressor and for both safe and unsafe AI suggestions.** In (a), gaze fixations versus influence of AI for fluid with safe AI recommendation. In (b), gaze fixations versus influence of AI for vasopressor with safe AI recommendation. In (c), gaze fixations versus influence of AI for fluid with unsafe AI recommendation. In (d), gaze fixations versus influence of AI for vasopressor with unsafe AI recommendation. Points are also categorised by grade of experience from lowest (SHO) to highest (consultant).

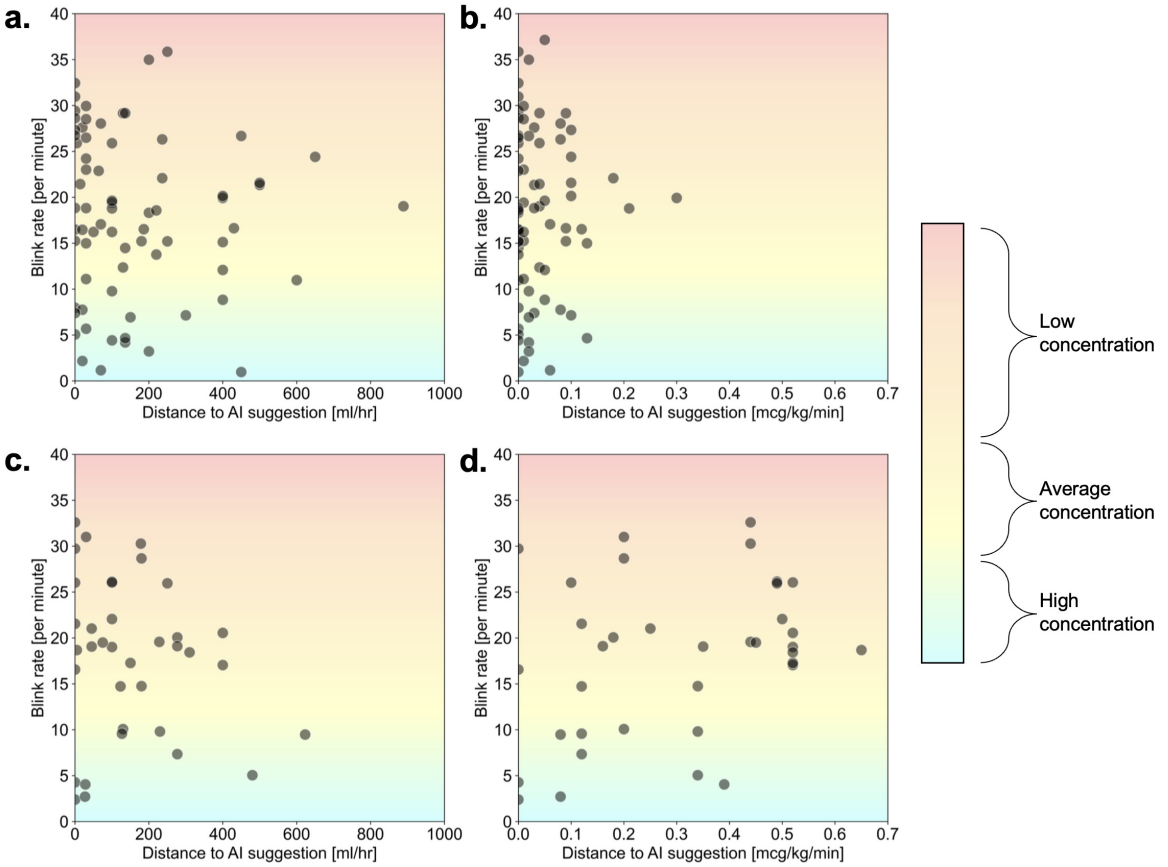


Fig. 4.11 **Blink rate by degree of adherence to AI for both fluid and vasopressor and for both safe and unsafe AI suggestions.** The degree of adherence to AI is defined by the per trial distance between a doctor's prescription and the AI suggestion. The larger the gap/distance between them, the lower the adherence to AI and vice versa. In (a), blink rate versus safe AI adherence for fluid. In (b), blink rate versus safe AI adherence for vasopressor. In (c), blink rate versus unsafe AI adherence for fluid. In (d), blink rate versus unsafe AI adherence for vasopressor. The background shading of the plots gives an indicator for expected blink rate depending on level of focus/concentration. Typical spontaneous blink rate is around 8-15 with lower levels suggesting higher focus/concentration and vice versa. Correlation measures: (safe fluid: r=0.003, p=0.981), (safe vasopressor: r=-0.000, p=0.998), (unsafe fluid: r=-0.076, p=0.649), (unsafe vasopressor: r=0.335, p=0.040).

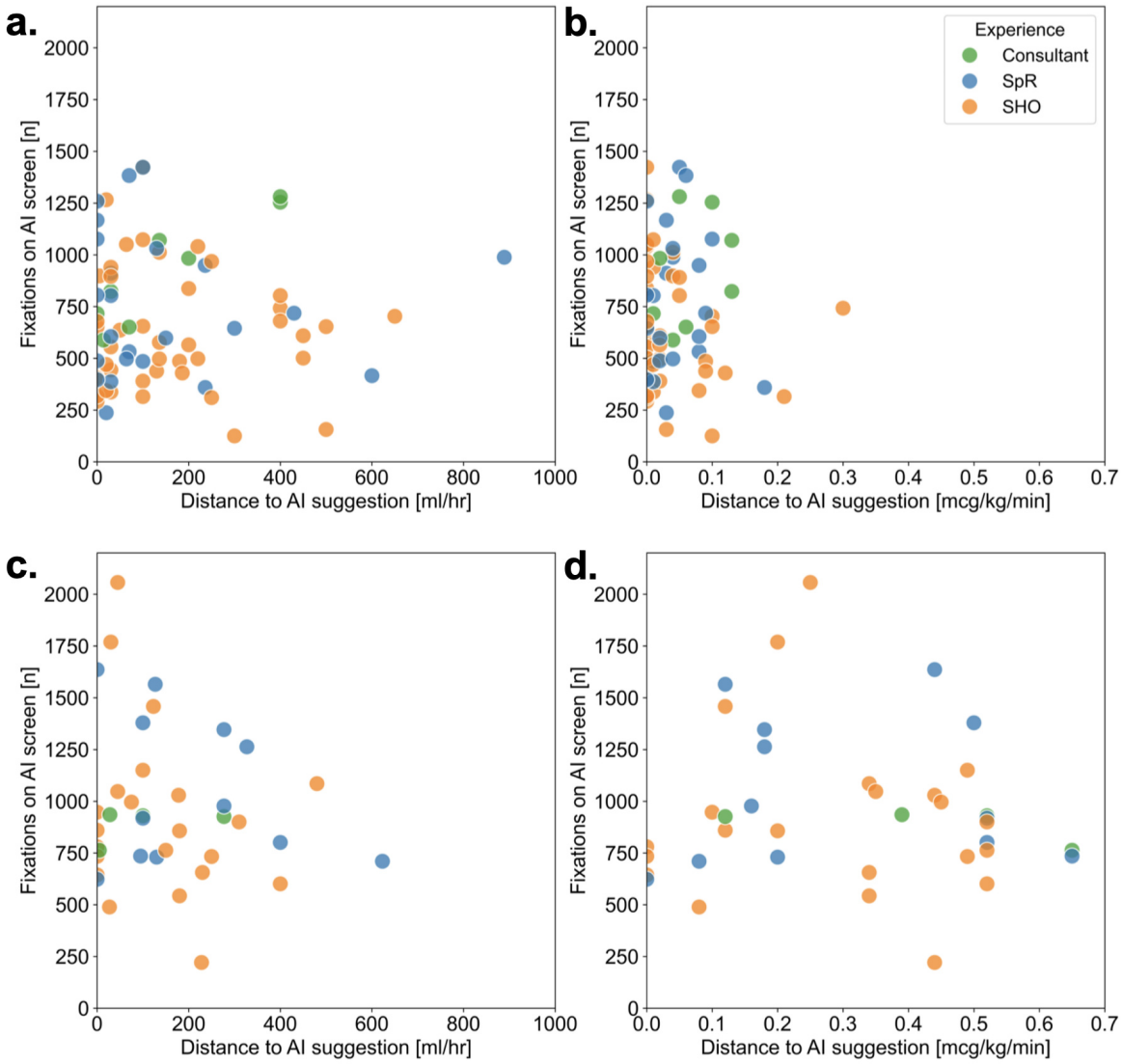


Fig. 4.12 **Number of gaze fixations by degree of adherence to AI for both fluid and vasopressor and for both safe and unsafe AI suggestions.** The degree of adherence to AI is defined by the per trial distance between a doctor's prescription and the AI suggestion. The larger the gap/distance between them, the lower the adherence to AI and vice versa. In (a), number of gaze fixations versus safe AI adherence for fluid. In (b), number of gaze fixations versus safe AI adherence for vasopressor. In (c), number of gaze fixations versus unsafe AI adherence for fluid. In (d), number of gaze fixations versus unsafe AI adherence for vasopressor. Points are also categorised by grade of experience from lowest (SHO) to highest (consultant). Correlation measures: (safe fluid: r=0.048, p=0.678), (safe vasopressor: r=-0.021, p=0.856), (unsafe fluid: r=-0.157, p=0.346), (unsafe vasopressor: r=-0.057, p=0.735).

0.04 mcg/kg/min and 0.30 mcg/kg/min respectively for safe and unsafe conditions, p<0.0001.

4.3.4 Rejection of unsafe AI suggestions and second opinions

When asked if they would act to stop an AI suggestion were it to be automatically administered to a patient, physicians rejected the unsafe AI suggestion 82% of the time compared to rejecting a safe AI suggestion only 29% of the time, p<0.001 (see Figure 4.13a). Physicians were also asked if they would like a second opinion (i.e. to escalate their uncertainty and seek consensus, a process designed especially for the more junior and less experienced clinicians). The proportion either rejecting an unsafe AI suggestion *or* escalating to a second doctor was 92% (i.e. only 8% of unsafe doses would have potentially been actioned in this experimental setting, see Figure 4.13a).

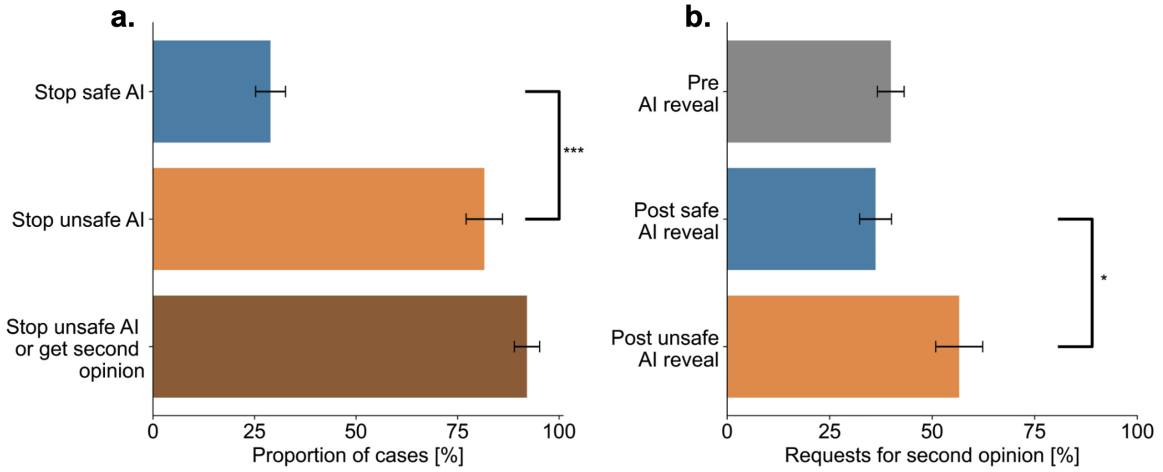


Fig. 4.13 **Physician responses regarding stopping AI suggestions and seeking second opinions.** In (a), a significantly larger proportion of AI recommendations were stopped in unsafe versus safe scenarios (and the proportion being stopped rose even higher when also taking into account seeking second opinions which would have likely led to rejection of unsafe AI). In (b), there are (appropriately) significantly higher second opinion requests after viewing unsafe versus safe AI suggestions. Significance levels are one star for p<0.05 and three stars for p<0.001.

Prior to seeing an AI suggestion, 40% of physicians asked for a second opinion. After seeing the AI suggestion, this proportion fell to 36% if they saw a safe AI

recommendation or rose to 56% if they saw an unsafe AI recommendation (i.e. unsafe AI suggestions led to a greater propensity to seek a second opinion than safe AI suggestions, p=0.003, see Figure 4.13b). Overall, second opinions were more commonly sought by junior (two years or less ICU experience) than senior physicians (more than two years experience), 79% versus 28%.

4.3.5 Effect of nurse challenge

In the incentivised unsafe case, 95% of physicians (36/38 as this comparison considers all physicians rather than just those with eye-tracking) did not change their decision despite challenge by the bedside nurse using the three verbal attempts described in Table 4.1. The two physicians who were persuaded had both initially said they would stop the AI suggestion if it were to be automatically administered but changed their mind to allow it after persuasion by the bedside nurse.

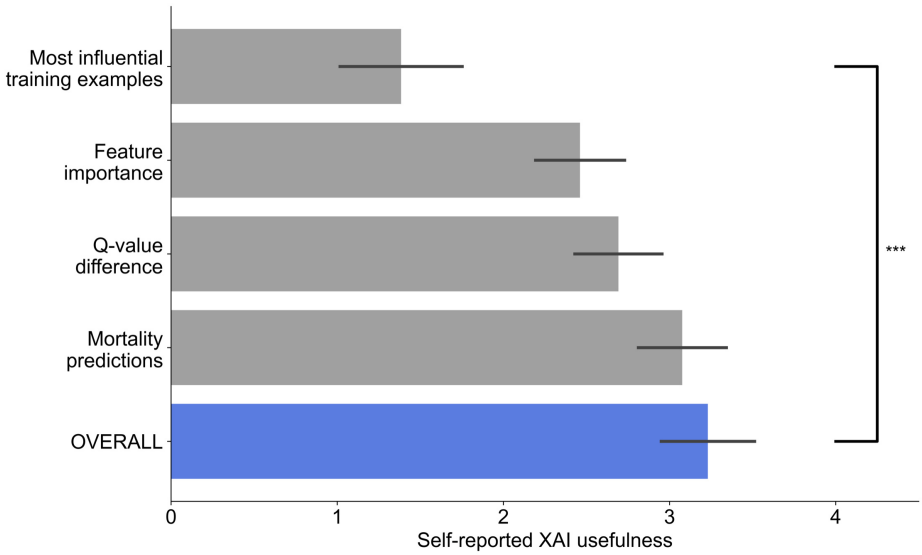


Fig. 4.14 **Usefulness rating for each type of explanation as well as overall rating.** Mean and SEM error bars. The only significantly different explanation type compared to overall was the 'most influential training examples'. Significance levels are one star for p<0.05 and three stars for p<0.001 (based on Student's t-test).

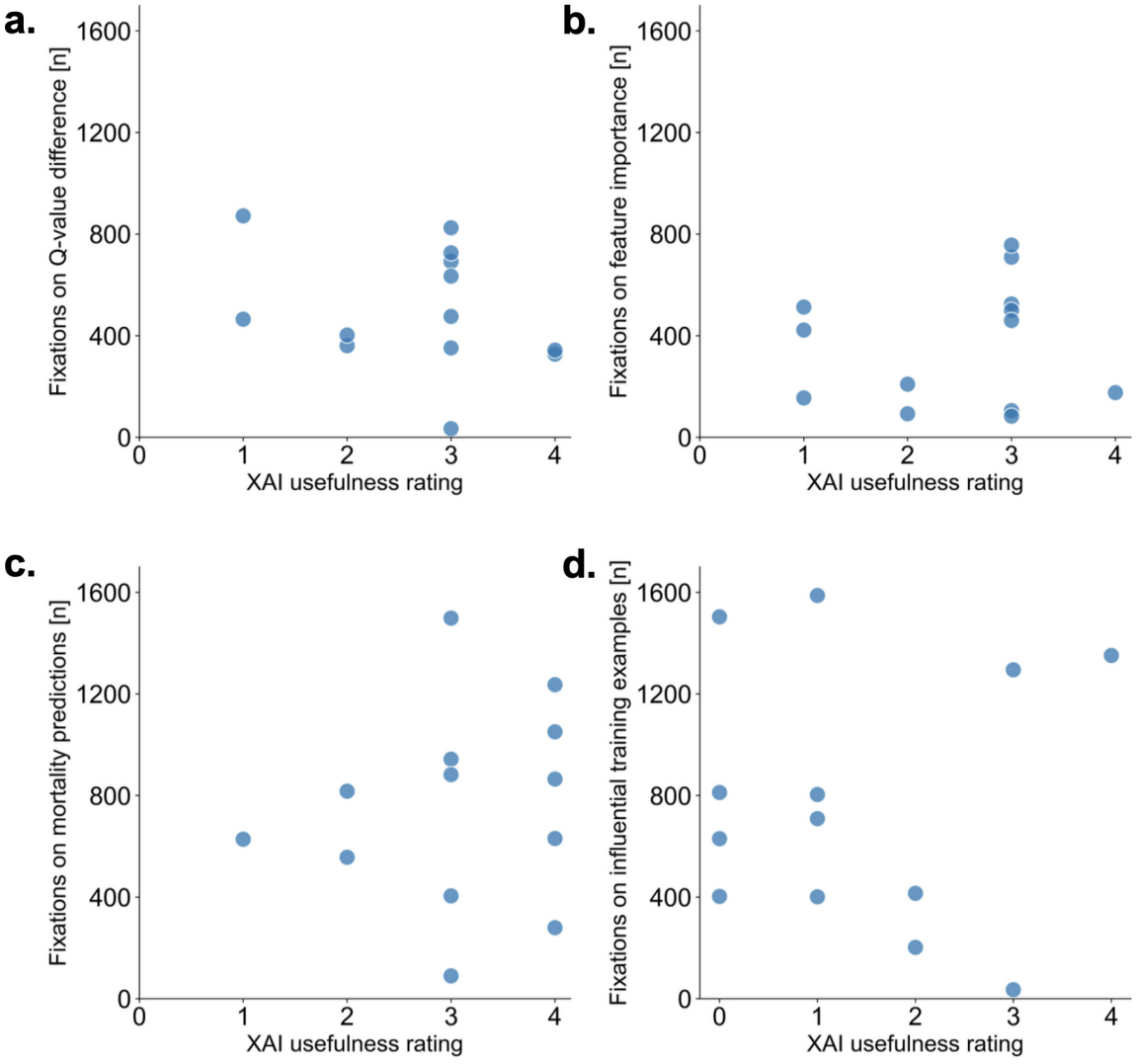


Fig. 4.15 **Fixations on XAI vs self-reported usefulness.** There was no significant correlation between fixations and usefulness rating for any of the four types of XAI (Q-value difference: r=-0.28, p=0.35, feature importance: r=0.08, p=0.79, mortality predictions: r=0.15, p=0.62, most influential training examples: r=0.01, p=0.97).

4.3.6 Self-reported XAI utility

The overall mean post-experiment utility rating for the XAI was 3.2 (SD 1.0) on a 0-4 scale with higher value implying the XAI was more useful. The training examples explanation was the only one of the four to be rated significantly lower than the overall rating for explanations in general (mean 1.4 versus 3.2, SD 1.3 versus 1.0, $p<0.001$, see Figure 4.14). When comparing the 'objective' marker of how many fixations there were on the four different types of XAI to the 'subjective' marker of how clinicians rated the usefulness of the four XAIs, no correlation for any XAI was found (see Figure 4.15).

4.3.7 Comparison to a low-fidelity non-simulation setting

Practice variation and adherence to AI from this simulation-based experiment were compared to the low-fidelity vignette experiment from Chapter 2). There were significantly different practice variation patterns for all grades of physician with fluid prescriptions but only for registrars and consultants with vasopressor (see Figure 4.16a, b). When assessing adherence to XAI recommendations, significant differences were again found for both fluid and vasopressor for all grades except for consultant vasopressor prescriptions (see Figure 4.16c, d).

4.4 Discussion

4.4.1 Summary of results

This study provides key insights into physician behaviour when interacting with AI-driven decision support tools and their explanations. First, measuring gaze fixations and blink rate as indicators of attention to an AI support tool was feasible in a high-fidelity simulation. Extending this to real-world clinical settings would require less intrusive eye-tracking hardware and careful handling of privacy concerns related to

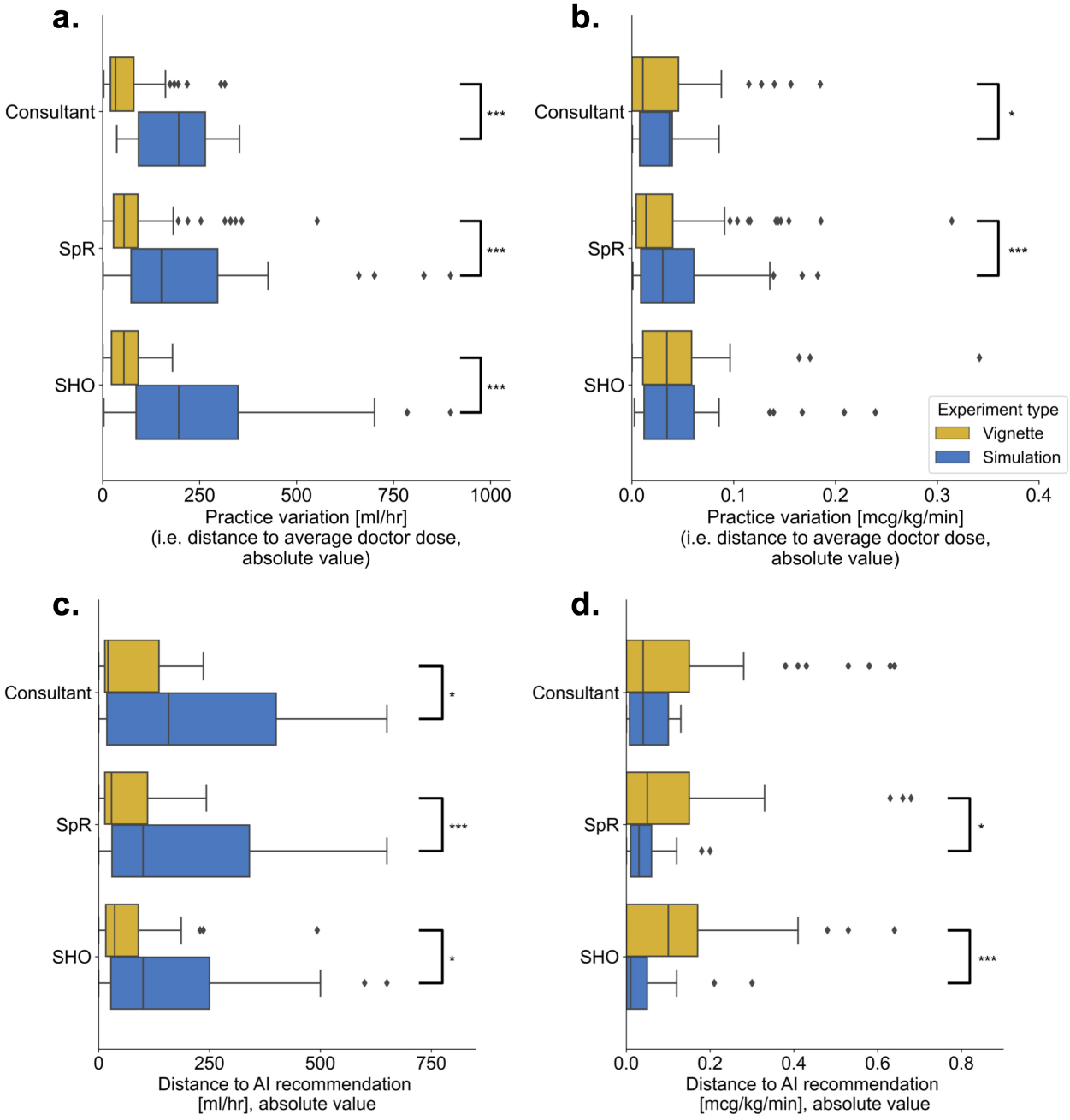


Fig. 4.16 **Comparison between prescription behaviour between vignette and simulation settings.** Differences in the distribution of practice variation for (a) fluid and (b) vasopressor. Differences in the distribution of adherence to AI for (c) fluid and (d) vasopressor. Vignette data is sourced from from Chapter 2 (see 2.3.

video recordings of staff and patients. Second, although unsafe AI recommendations attracted more attention (as proxied by gaze fixations) than safe ones, there was no significant increase in attention to any of the four types of explanations during unsafe scenarios. This suggests no extra reliance on explanations in unsafe situations. Third, the self-reported utility of explanations by physicians did not correlate with the attention they devoted to them. This suggests that self-reports alone are not sufficient for evaluating XAI as crucial elements of the human-AI interaction may be missed.

Fourth, physicians did not spend significantly more time reviewing non-AI clinical data after encountering an unsafe AI suggestion to try and understand why the AI suggestion was so extreme. However, more often than not, the unsafe AI suggestions were not acted upon. It therefore remains unclear from the eye-tracking data what the cognitive processes underlying the rejection of unsafe AI suggestions were (or whether enough of the clinical data could be maintained in short-term memory that physicians felt no need to look back after seeing an unsafe AI suggestion). Fifth, I did not find strong patterns between eye-tracking metrics and either clinical practice variation among physicians or their influenceability to AI recommendations. Sixth, there were notable differences in prescription variability (both with and without XAI) between the high-fidelity simulation format in this experiment and data from my previous lower-fidelity vignette experiment suggesting that fidelity is an important modifier of the behavioural results that emerge from clinical evaluations of XAI and thereby justifies the extra resource cost associated with high-fidelity evaluation.

4.4.2 Limitations

There are several limitations which should be borne in mind when interpreting the findings from this experiment. First, while the fidelity of the simulation suite is much closer to real-world practice than a vignette experiment, it still lacks key elements of the actual hospital environment. For example, the ability to dynamically examine a patient,

observe them over time (rather than as a snapshot scenario) and interact with multiple colleagues within a multi-disciplinary team before acting upon an AI recommendation. Unfortunately, these same elements make it very difficult to standardise an experiment in a real hospital, requiring prohibitively large sample sizes. As a result, simulation experiments remain crucial for the initial exploration of human-AI interaction dynamics prior to larger-scale real-world studies.

Second, the explanations were presented simultaneously. It is therefore difficult to disentangle the marginal contribution of each explanation if seen in isolation. Doing so would have required a factorial trial design with many more arms (and a much larger sample size). This means that these results, strictly speaking, only apply to the overall impact of seeing all four explanations together. Third, the sample size was small so I cannot exclude the possibility that some comparisons might have been significantly different if more physicians had been recruited, particularly more senior and experienced ones. The variable performance of the eye-tracking glasses and software in consistently detecting pupils led to the exclusion of some physicians, further reducing the sample size available for analysis. As above, I therefore cannot definitively say whether non-significant findings are due to a lack of a true effect or an insufficient number of participants.

Fourth, the demarcation between safe or unsafe AI recommendation necessarily imposes an arbitrary bound on a continuous spectrum (where no 'gold standard' answers exist). I used deliberately extreme unsafe scenarios to capture lower bounds on the tendency to inadvertently follow dangerous AI advice. This was achieved predominantly with extreme vasopressor doses as this is more imminently dangerous than extreme fluid doses. Therefore, the findings in this study might be more extreme for borderline AI recommendations, which many clinicians might feel uncomfortable about following but would not have met our strict definition of 'unsafe'. Fifth, some explanations were mostly graphical in form (feature importance), while others were comparatively more

text-heavy (influential training examples). It is likely that less gaze would be required to parse the former than the latter and this could have potentially confounded the comparisons between explanations. A related confounder which I did mitigate against was location bias (i.e. reading left to right and top to bottom for example, giving more weight to top left corner). The position of explanations was rotated between trials (i.e., the XAI based on feature importance would not always be in the same corner of the screen).

Sixth, the explanations for both safe and unsafe conditions in the same patient scenario were identical. This leads to the inadvertent introduction of another experimental condition (quality of explanation). In some studies, explanation quality has been the exclusive focus of the experiment [200]. An experiment with more arms might have better assessed this condition (e.g. six arms: safe/unsafe AI x good quality/poor quality/absent XAI). Seventh, the large display screen used in the study is not typical of the screens used for most hospital EHR systems which are ususally displayed on laptops or desktop monitors. The large screen was chosen to maximise gaze discrimination between the explanation types in each corner of the screen. Future experiments should use smaller screens and, far more importantly for true fidelity, embed the XAI being evaluated into a dummy version of the real hospital EHR software.

4.4.3 Implications and comparison with other studies

While acknowledging these limitations, reviewing the study findings alongside existing literature nonetheless provides critical insights into how the deployment of XAI-based clinical decision support tools could be improved. The behavioural intuition that explanations ought to help users correctly reject unsafe or poor AI recommendations often follows this chain: (i) the user is presented with an unsafe AI suggestion, (ii) the user is provided an explanation for the AI suggestion, (iii) the user notices a deficiency in the explanation or fails to find mitigating circumstances that explain

why the AI suggestion might be inappropriate, (iv) the user rejects the unsafe AI suggestion. A study involving non-clinicians suggests that the causal link between (iii) and (iv) might in fact be absent [150]. The authors found a greater influence from AI advice when provided alongside an explanation. However, the quality of the explanation did not seem to make a difference. In other words, participants may have fallen into an automation bias trap by equating the mere presence of an explanation as a heuristic for AI suggestions that they ought to follow, rather than actually evaluating the explanation. This danger of automation bias has also been highlighted in many clinical studies [151, 201].

Further evidence on this topic includes an experiment evaluating a mental health drug decision support tool where explanations did not rescue clinical users from following intentionally poor AI recommendations [68]. In my study, I confirmed that the higher rejection rate of unsafe compared to safe advice could not clearly be explained by either greater reliance on, or attention devoted to, simultaneously presented explanations. Unlike previous experiments, this study addressed the inherent weaknesses of self-report metrics by assessing the trifecta of what physicians *did* (i.e. their actual prescription decisions), *how* they did it (i.e. what they looked at during the decision-making process) and what they *said* (i.e. their subjective ratings for the explanations after the experiment ended).

Algorithmic aversion was not specifically assessed in this study (this is the opposite phenomenon to automation bias where there is instead inappropriate reluctance to use AI) because in a clinical setting, this would tend toward the current standard of care (status quo) and therefore be less risky, at least from a medico-legal standpoint [202]. However, the behavioural mechanisms underlying algorithmic aversion have been investigated by others and seem related to the subjective difficulty of understanding AI and the presumed illusory subjective understanding of human decision-making itself [203]. It seems likely that progress in medical XAI will need to address both algorithmic

aversion and automation bias. Consequently, identifying where a physician falls on this spectrum and possessing an ability to individualise the XAI output specifically to that physician might be important in future. I attempted to investigate whether eye-tracking of behaviour could facilitate achieving this goal.

One notable example of using eye-tracking to evaluate AI includes work by Cao and colleagues who used a spatial reasoning task and found a positive correlation between the percentage of gaze on the AI suggestion and both perceived user reliance on AI as well as their agreement with AI suggestions [182]. However, gaze did not appear to correlate with perceived trust. Although I did not measure perceived trust directly, my study did find no correlation between subjective explanation ratings and the number of gaze fixations on the AI explanations. One potential advantage of evaluating XAI with eye-tracking is that it occurs in real-time (unlike post-hoc human agreement with AI suggestions or subjective ratings) and could therefore form a key part of an adaptive collaboration feedback loop [182]. Unfortunately, my findings suggest that before eye-tracking can become a central aspect of such a feedback system incorporating XAI, we need to first establish more robust phenotypic patterns of eye movements that at the very least accurately categorise users and ideally even predict their behavioural interactions with AI.

4.4.4 Conclusion

Taken together, these findings build on the work in Chapters 2 and 3 with important insights into the design and evaluation of real-world clinical XAI systems. The key takeaways are that (i) it is feasible to use eye-tracking to evaluate human behaviour with simultaneously presented XAI, (ii) that the response to safe or unsafe AI recommendations is identifiably different, (iii) that the lack of clear rescue provided by XAI calls into question its utility as a mitigation against harm from physicians erroneously following poor-quality AI advice and (iv) that a simulation suite setting provides a

better estimate for the real-world variability of human decision-makers over simpler vignette formats. This final point might be due to more accurately reflecting the situational awareness and cognitive load that experts face in real life [204].

Chapter 5

Large language model derived explanations

This chapter is closely based on a manuscript under preparation on which I am the first author and my supervisors are co-authors:

Interaction behaviour of physicians with large language model derived XAI-decision support. Submission to a peer-reviewed journal specialising in digital health is planned in due course.

5.1 Introduction

5.1.1 The advantages of dynamic explainability

Traditional static explanations (of the form used in Chapters 2 and 4) have shown limited effectiveness in influencing clinical prescriptions as confirmed in my previous experiments [138, 188]. One possible reason might be that the static nature of the XAI fails to capture the nuances of individual patient cases compared to the highly dynamic process of clinical reasoning that occurs during the course of a human physician assessing a patient. In contrast, a dynamic XAI would allow clinicians to ask follow-up

questions, request clarifications and receive more tailored follow-up explanations based on their specific concerns or the unique aspects of a patient scenario. This could be achieved by means of a dialogue system that more closely resembles the natural language process which clinicians use to communicate to each other (and to internally formalise their own thoughts). Indeed, a dialogue system is highlighted by Sokol and Flach as a key means of facilitating higher level human-AI collaboration [205]. Furthermore, the conversation transcripts themselves might contain important insights into human-AI interaction behaviour in a clinical setting (for example, vulnerability to cognitive biases).

5.1.2 Large language models

Large language models (LLMs) work by predicting the most likely next word in a sequence, based on the context provided by the preceding words. This is achieved through a neural network architecture known as a transformer which processes input text in parallel rather than sequentially. The network is therefore able to capture complex and subtle relationships between words and phrases because it can consider the entire context simultaneously. During training, the model's internal parameters (consisting predominantly of weights and biases) are adjusted to minimise the difference between predicted and actual next words. By repeating this process at scale, the LLM is able to learn higher level concepts as well as patterns of language use, factual knowledge and even basic reasoning skills. When given a question (known as a prompt), the model uses its learned parameters to generate a response by repeating the next word prediction for the length of the answer.

The scale of these models and the combined diversity and sheer volume of their training data (i.e. the entire corpora of text available on the public internet) allow them to exhibit remarkable flexibility and apparent understanding across a wide range of topics and tasks. Crucially, this does not require task-specific training. As an

example, PaLM is a model trained by Google containing 540 billion model parameters [206]. LLMs captured global attention with the release of ChatGPT, a commercially available AI chatbot released by OpenAI launched in November 2022. ChatGPT reached 100 million monthly active users just two months after its launch making it the fastest-growing consumer application in history and sparking widespread interest in AI capabilities beyond the scientific community [207].

In healthcare applications, LLMs have been applied to a range of tasks with variable performance thus far including medical literature summarisation [208], peer-review of medical articles [209], medical licensing exams [210, 211], patient triage [212], diagnosis [213], clinical decision-making [214], drug discovery [215], clinical trial recruitment [216] and medical coding [217].

5.1.3 Explainability with LLMs

The ability of LLMs to understand and produce natural language makes them particularly suited for generating interactive clinical explanations. For example, they can adapt their explanations to the physician's level of expertise and provide rationale in a form that more closely mirrors human clinical reasoning. This latter point might also lead to better integration of XAI suggestions into clinical workflows [57]. However, automation bias is an even larger concern than with traditional XAI given the persuasive and human-like conversational nature of LLMs. For example, some clinicians have raised concerns that LLM suggestions will be challenged less frequently when they align with what physicians want to hear (i.e. confirmation bias) [218].

Another question that arises is "who watches the watchmen". In other words, if LLMs are used as a post-hoc explainability method for an RL agent, is there any recourse to determining the explainability of the LLM itself? Given the scale of LLMs, the answer is unsurprisingly no, at least for now. Researchers at OpenAI attempted to address a simpler version of this problem by using GPT-4 to automatically generate and evaluate

natural language explanations for individual neurons in the much more basic GPT-2 [219]. Their method involved a three-step process: (i) generating explanations based on neuron activations, (ii) simulating neuron behaviour based on these explanations and (iii) scoring the explanations by comparing simulated activations to real activations, thereby quantifying the quality of the explanations. Overall, while they could explain the activations of certain neurons, the approach struggled with many neurons whose functions remained unclear or complex, likely due to either polysemantic behaviour or representing concepts that are difficult to articulate in natural language [219]. These findings underscore the difficulty of truly interpreting LLMs and suggest that post-hoc explanations, while imperfect, may be a reasonable approach for providing XAI for non-LLM derived AI clinical decision support tools.

In this final experiment, I aimed to evaluate how LLM-derived XAI influences physicians in a high-fidelity clinical setting. Specifically, the study sought to assess the impact on prescription decisions, examine the dynamics of physician-AI collaboration and measure cognitive engagement through eye-tracking metrics, while also comparing these findings to my previous experiments with different XAI approaches.

5.2 Methods

5.2.1 Patient scenario selection

Unlike in the previous simulation suite experiment (see Chapter 4, section 4), this experiment used real patient scenarios because the intent was to assess LLM-driven XAI on real AI treatment suggestions. A subset of six patients from the 24 used in the vignettes experiment (see Chapter 2 methods, section 2.2) were selected aiming to have a range of scenarios (i.e. a range of noradrenaline suggestions, a range of patient acuity from mild illness to very sick).

5.2.2 Experiment format

Physicians who participated in the experiment first completed a standardised experiment briefing (see appendix F) and a pre-experiment questionnaire (see appendix F) on demographics and attitudes toward AI . This time, the questionnaire also included questions on risk tolerance and appetite for uncertainty in clinical practice (adapted from three validated scales: 'Tolerance for Ambiguity' (TFA), 'Pearson Risk Attitude' (PRA) and 'Ambiguity Aversion in Medicine' (AA-Med), see work by Han et al. [220]). The reason for including these questions was to assess whether these are characteristics associated with particular patterns of interaction behaviour with AI/XAI decision support tools.

Physicians were then taken into the simulation suite for orientation before conducting the eye-tracking calibration exercises (see Chapter 3 Methods, section 3.2). Eye-tracking was conducted throughout the experiment as detailed in the Chapter 3 Methods, section 3.2). I played the role of the bedside ICU nurse throughout all scenarios. Within each of the six scenarios, physicians were tasked with performing an assessment, which included reviewing the available patient data and examining the patient mannequin. Afterward, they were asked by the bedside nurse for a fluid and vasopressor prescription for the next hour of the patient's admission (to be entered into the bedside laptop containing the AI/XAI-decision support tool).

Physicians were then shown either a plain text AI recommendation or the plain text AI recommendation accompanied by an explanation provided by a LLM (see details below). A chat window allowed participants to have a conversation with the LLM on the patient scenario and explanation. There was no limit for time or conversation length in place. AI and XAI conditions alternated and physicians were recruited sequentially to one of two groups (order of six cases either starting with AI and alternating or starting with XAI and alternating). Regardless, all six patient cases were experienced

in the same order. Before completing the scenario, the physician had to finalise their prescription doses in light of either the AI or XAI information.

A post-experiment questionnaire (see appendix F) included previously validated questions on various domains that contribute to high-quality explanations (also known as the System Causability Scale (SCS), see work by Holzinger et al. [221]).

5.2.3 Generating LLM-derived explanations

Using an LLM to provide an explanation for an AI suggestion from an RL model is a necessarily post-hoc approach (i.e. the interpretability is not inherent to the source model but provided by a second model). This method, common in XAI, involves a secondary model interpreting the primary model's output. However, the fidelity of this interpretation to the original model can vary significantly. At one extreme, providing the LLM with only the AI suggested doses would likely result in either severe hallucination or overly vague statements, neither of which would be clinically useful. Therefore, to maximise the LLM's fidelity to the RL model, it was crucial to supply it with comprehensive information from the RL training process. This included Q-value differences, state transitions, influential training cases, and feature distributions over clustered states. This approach aimed to create explanations that were both accurate representations of the RL model's decision-making process and valuable to clinicians.

The main issue was how to provide the RL data to the LLM. Initially, I attempted a direct approach by providing the LLM with raw training data (an example can be seen in appendix F) consisting of a comma separated value (CSV) file with predominantly numerical data and some accompanying prompts. However, this method proved ineffective, likely due to the difficulty in the LLM processing and interpreting numerical data and complex code structures.

Recognising that LLMs are optimised for natural language processing, I instead developed an alternative approach that transformed the numerical and categorical data

from the RL process into prose text. This method involved creating a series of Python functions that converted various aspects of the RL model output and patient data into human-readable paragraphs. These functions addressed the following key areas: (i) patient specific summary, (ii) state summary, (iii) influential training case(s) summary, (iv) AI recommendation certainty (proxied by Q-value difference) and (v) future state predictions/alternate trajectories if non-optimal actions were pursued.

A key aspect of this approach was the careful translation of numerical values into clinically relevant language. I developed a system to categorise continuous variables based on their deviation from the mean, using standard deviations as a measure of abnormality. This approach was applied to various clinical parameters. For example, instead of a raw lactate number, lactate levels were described as "normal" if within 1 SD of the mean, "mildly raised" if between 1 and 2 SDs above the mean, "moderately raised" if between 2 and 3 SDs, and "very high" if more than 3 SDs above the mean. This standardised approach was applied across various clinical parameters. I created z-scores for each feature ((value - mean) / standard deviation) and used these to identify significantly deviating variables. Variables with z-scores exceeding ± 1 were flagged as abnormal and included in the patient summary. This method allowed for a consistent and clinically meaningful way to highlight important deviations across different physiological parameters in the final prose explanation.

I also implemented context-specific interpretations for certain variables. For example, fluid balance was categorised as "low," "medium," "high," "very high," or "extremely high" based on predefined thresholds that corresponded to clinically relevant cutoffs. Similarly, vasopressor dosage was expressed in familiar clinical units (mcg/kg/min) and described in terms of its intensity (e.g., "minimal support" vs. "high dose"). For categorical variables like ventilation status, I used boolean logic to translate numerical indicators into clear statements such as "is mechanically ventilated" or "is not mechanically ventilated" for example. Time-based variables, like the patient's progress

in their sepsis course, were mapped to descriptive phases such as "very early," "early," "middle phase," or "late" based on the time step in the patient's episode.

The final 'explain' function combined outputs from all the aforementioned subsidiary functions to generate a cohesive explanation (an example can be seen in appendix F). By presenting the RL model's complex data in a format more aligned with the LLM's natural language processing capabilities, this approach aimed to enhance the accuracy and relevance of the generated explanations while keeping the simplicity of a zero-shot approach (i.e. not relying on further fine-tuning of the underlying LLM). The final prompt is also shown in appendix F.

5.2.4 Integration with commercially available LLMs

I developed a JSPsych-based web application that integrated application programming interfaces (APIs) from Anthropic's Claude and OpenAI's GPT-4. I specifically invested time ensuring that two independent LLM APIs were working to ensure resilience for the experiment in the event that an upgrade (occurring frequently even back in late 2023/early 2024) resulted in a software error that would have left me unable to run the experiment. Had there been longer to run the experiment, this also left the option open to compare LLMs as one possible factorial arm (in addition to AI versus XAI with LLMs) though this was not ultimately pursued. Below, I outline the technical approach I took to implement this application, focusing on the integration with the Claude and GPT-4 APIs.

Application architecture

The application was structured around a Flask web server, which manages requests and responses between the client-side JavaScript and the AI models. Flask was configured to handle CORS (Cross-Origin Resource Sharing) to enable smooth communication between the frontend and backend, with sessions managed using Flask-Session to

maintain state across requests (important for experiment with multiple scenarios and prescription doses carried forward pre- and post-AI suggestion reveal).

API integration with Claude API (from Anthropic)

To interact with Anthropic's Claude v2.1 API, I utilized Amazon's Bedrock service (bedrock-2023-05-31) to manage requests (at the time, direct access to Claude API was severely limited via a long waiting list). The integration involved setting up a boto3 client to communicate with the Bedrock runtime in a specified AWS region ('us-east-1'). For each clinical case, I defined a structured prompt (see previous section) that incorporated patient-specific data and AI recommendations. The Claude API was invoked using 'invoke_model_with_response_stream', which streams responses back to the client in real-time.

This approach allowed for continuous, word-by-word rendering of Claude's output in the browser, simulating a live conversation with the AI. I spent time executing the word-by-word approach as opposed to waiting for the entire message to finish generating and displaying it all at once in a single block because I felt the former better simulated the anthropomorphic nature of having a conversation with another entity versus a static search on the web for a guideline document for example.

The API request included parameters such as 'max_tokens_to_sample', 'temperature', and 'top_p' to control the generation characteristics, ensuring the responses were tailored and relevant to the physician's task. The responses were streamed using 'stream_with_context' to ensure continuity and maintain the state of the conversation throughout the session.

API integration with GPT-4 API (from OpenAI)

For the GPT-4 API integration, I used OpenAI's Python client to facilitate communication. The API call was made via client.chat.completions.create, specifying the

model version (gpt-4-1106-preview), message history, and other parameters such as 'max_tokens' and 'stream'. The responses from GPT-4 were streamed to the frontend in a manner similar to Claude's integration. This ensured a consistent user experience across both AI models and ensured there would be minimal noticeable difference had I needed to switch models.

User interaction and Data Flow

On the client side, the experiment was structured using JSPsych, a JavaScript library for creating and running behavioural experiments that was also used in the vignette experiment (see Chapter 2 Methods, section 2.2). Interactive elements were designed using HTML and JavaScript. For instance, physician responses in the chat were processed by Flask routes ('/get_first_response' and '/get'). These routes handled the initial and follow-up AI responses, respectively. Data from each interaction, including user inputs and AI responses, was captured using JSPsych's data management features and saved via AJAX requests to the locally running server. Conversations were serialised to JavaScript Object Notation (JSON) format for ease of later analysis.

5.2.5 Additional analyses

Influence of AI was again calculated on a continuous scale from 0 (completely ignoring advice) to 1 (completely relying on advice) per work by Yaniv et al. [141] (as well as the alternative analysis with distance between final prescription and AI/XAI recommendation, both per Chapter 4 methods, section 4.2). Additional analyses in this experiment beyond those from Chapters 3 and 4 included quantitative text analysis of parsed conversations with the LLM.

After filtering to remove the initial system prompt and the first AI response (the core explanation) for each user-case combination, the following metrics were obtained: (i) word count per response, (ii) conversation length per conversation, (iii) sentiment

analysis (using the TextBlob library [222]) to calculate polarity (measured on a scale from -1 very negative sentiment to +1 very positive sentiment) and subjectivity (measured on a scale from 0 to +1 where a score of 0 suggests a completely objective message with only factual information and no opinions or emotions while a score of +1 indicates a completely subjective text, filled with personal opinions, emotions, and/or biases).

Latent Dirichlet Allocation (LDA), a generative statistical modelling technique, was also used. LDA can be used to discover underlying topics in a document collection by representing them as a mixture of topics, with each topic consisting of a distribution over words. LDA assigns words to topics based on their co-occurrence patterns within the text, allowing for the identification of abstract themes and patterns in large volumes of text. I anticipated that it might not work well here due to short responses but the plan was to attempt the analysis nonetheless. Persuasion scores were calculated using a predefined list of persuasive phrases (should, must, consider, recommend, suggest, clearly, evident, as you know, important, in my opinion, therefore, because, as a result) and normalised to total word count of the response.

5.2.6 Participant recruitment and experiment conduct

ICU physicians were recruited as participants using both convenience sampling and targeted advertising to a local NHS trust (Imperial College Healthcare NHS Trust). The following inclusion criteria were used: (i) practising physician, (ii) has worked for at least 2 months in an adult ICU, (iii) currently works in ICU or has worked in ICU within the last 6 months. Electronically recorded informed consent was obtained from all physicians. Each experiment lasted approximately 60 minutes in total and participants were compensated for their time.

5.3 Results

Nineteen ICU physicians were included (13 male, 6 female) of which ten had eye-tracking data available (8 male, 2 female). Mean physician age was 34 years (standard deviation (SD) 5 years). Mean total clinical experience was 8.2 years (SD 4.7 years, range 1.5 years to 22 years). Mean ICU experience was 4.1 years (SD 3.5 years, range 3 months to 12 years), see Figure 5.1a. The attitudes to AI from the pre-experiment questionnaire (see Figure 5.1b) once again followed a broadly similar pattern to the previous studies. The mean practice variation for fluid was 160 ml/hr (SD 165 ml/hr) and for vasopressor was 0.06 mcg/kg/min (SD 0.13 mcg/kg/min). All physicians completed the task successfully with a mean scenario completion time of 4.9 minutes (SD 0.9 minutes).

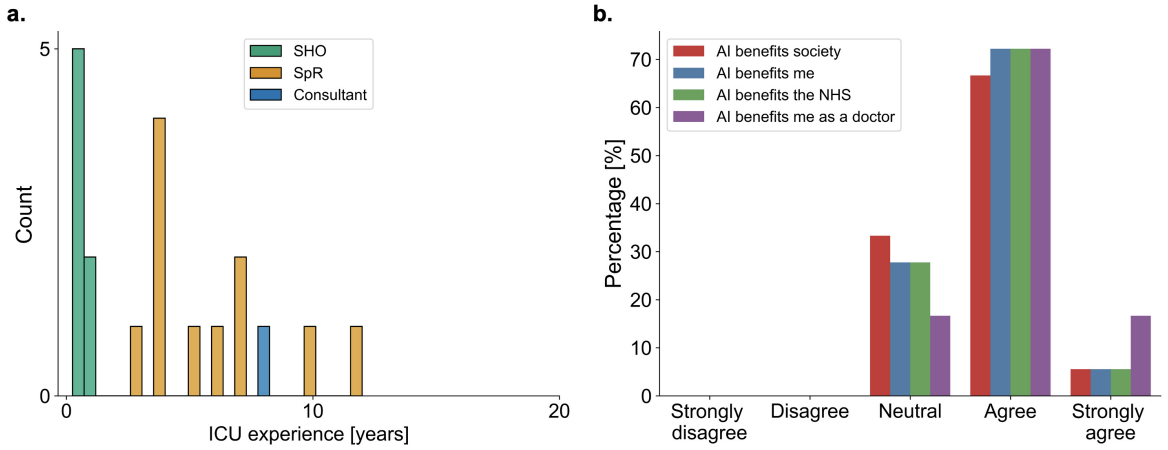


Fig. 5.1 **Physician experience and attitudes.** Physician experience levels in (a) and answers to AI attitude questions in (b).

5.3.1 Eye-tracking metrics

Fixation count per scenario is compared both pre- and post-reveal of the AI/XAI suggestion in Figures 5.2a and b respectively. As expected, fixation count is similar in the traditional clinical (non-AI) ROIs pre AI/XAI reveal. Post AI/XAI reveal, physicians hardly looked back to non-AI ROIs at all (and where they did, this was

most likely for the ICU chart). The fixation count per scenario on the AI screen was significantly higher in the XAI condition versus the AI condition (mean 1363 [95% confidence interval (CI) 852 to 1874] vs. 235 [95% CI 162 to 309] respectively, p=0.0004, Figure 5.2b). However, considerably more screen time was spent after AI/XAI reveal in the XAI condition (6.9 minutes, SD 3.3 minutes) than in the AI condition (1.3 minutes, SD 0.5 minutes) and therefore longer gaze duration confounds the raw fixation count comparison.

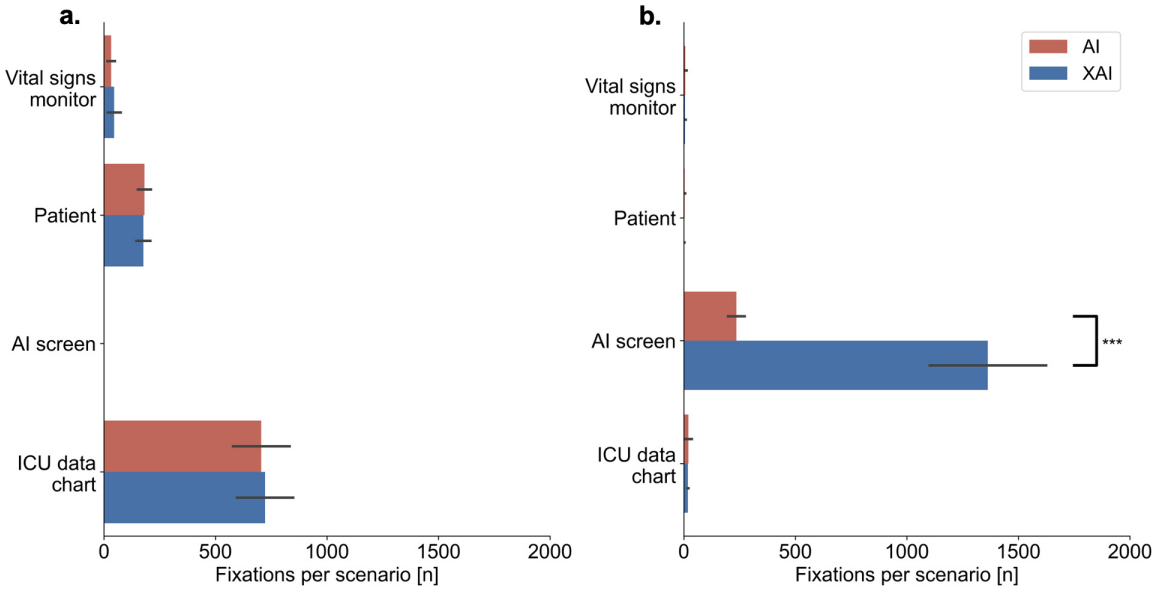


Fig. 5.2 **Fixations per scenario by AI or XAI condition of scenario**. Mean and SEM error bars. The AI screen is a super-set encompassing the AI suggestion and all four XAI ROIs. Significance levels are three stars for p<0.001 (based on Student's t-test).

Mean fixation duration was broadly similar at around 120 to 130ms for all ROIs (see 5.3a). Mean blink rate was lowest for the ICU chart (8.2 blinks per minute (bpm)). Unlike in the previous simulation suite experiment, the blink rate on the AI screen was the next lowest at 13.2 bpm (SD 9.9), see Figure 5.3b showing both ICU chart and AI screen within the lower blink rate region synonymous with higher concentration tasks such as reading.

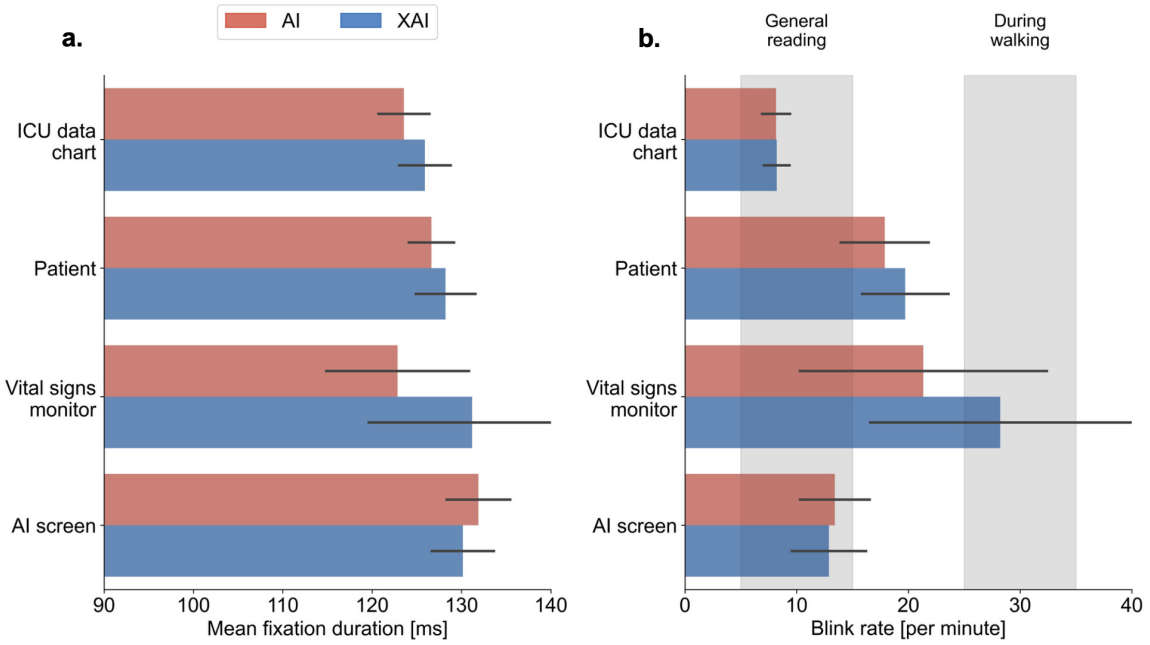


Fig. 5.3 **Mean fixation duration and blink rate by region of interest.** Mean and SEM error bars. Red bars refer to the AI condition while blue bars refer to the XAI condition. In (a), fixation duration is shown while in (b), blink rate is shown. The grey shaded regions show typical blink rate estimates associated with either general reading (requiring more concentration) and walking (less concentration) as per Chidi-Egboka et al. [169].

For all ROIs (this time including the patient mannequin unlike in the previous simulation suite experiment), there was a significantly higher actual than chance gaze proportion, see Figure 5.4a. Furthermore, the actual gaze proportion on the AI screen was higher than on the ICU data chart (the opposite was true in the previous experiment). The ratio of actual to chance gaze was 2.3, 1.6, 9.3, 2.0 respectively for the major ROIs (AI screen, ICU chart, vital signs monitor, patient mannequin, see Figure 5.4b).

5.3.2 Clinical practice variation among physicians

Practice variation was again (per section 4.2) defined as the dose distance of an individual physician from the average of all pre-AI reveal physician prescriptions in any

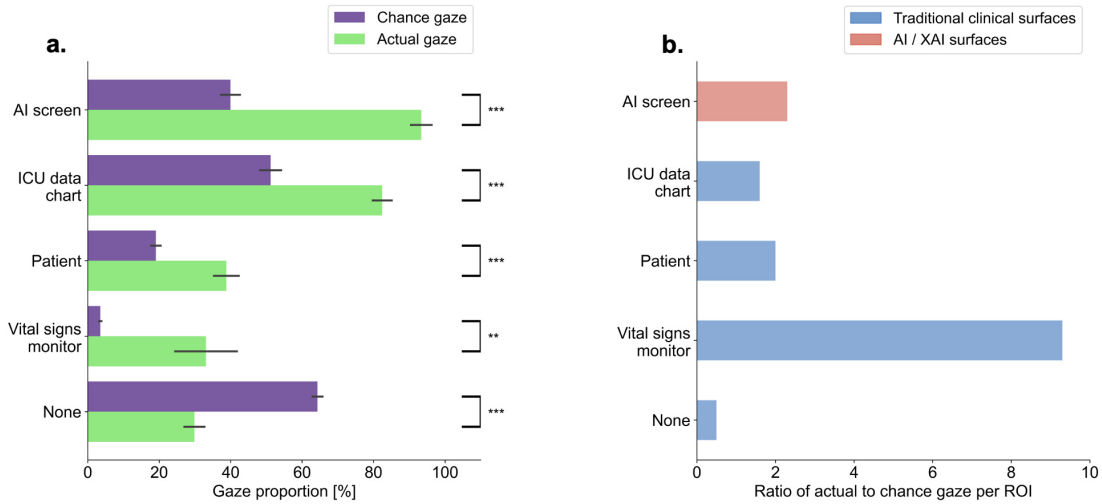


Fig. 5.4 **Gaze proportion per ROI and ratio of actual to chance gaze per ROI.** Mean and SEM error bars. Chance gaze is the proportion expected based on the area occupied by an ROI within the visual field (i.e. if ROI takes up 50% of the screen, 50% of the time, we would expect gaze to fall within it 25% of the time by chance alone). Actual gaze is the proportion observed during the experiment. Sub-figure (a) shows the chance and actual gaze proportions while sub-figure (b) shows the ratio of the two. Significance levels are one star for $p<0.05$, two stars for $p<0.01$ and three stars for $p<0.001$ (based on Student's t-test).

given trial/scenario (higher dose distance suggesting that the physician was more of an outlier in that particular trial and vice versa). When assessing blink rate on the clinical ROIs, there was weak positive correlation for fluid (r=0.074, p=0.57, Figure 5.5a) and vasopressor (r=0.088, p=0.50, Figure 5.5b). When assessing fixation duration on the clinical ROIs, there was weak negative correlation for both fluid (r=-0.23, p=0.08, Figure 5.5c) and vasopressor (r=-0.21, p=0.11, Figure 5.5d). Overall, the sample was too small to reliably conclude whether there was a pattern between the eye-tracking metrics (blink rate and fixation duration) and the degree of clinical practice variation (see Figure 5.5).

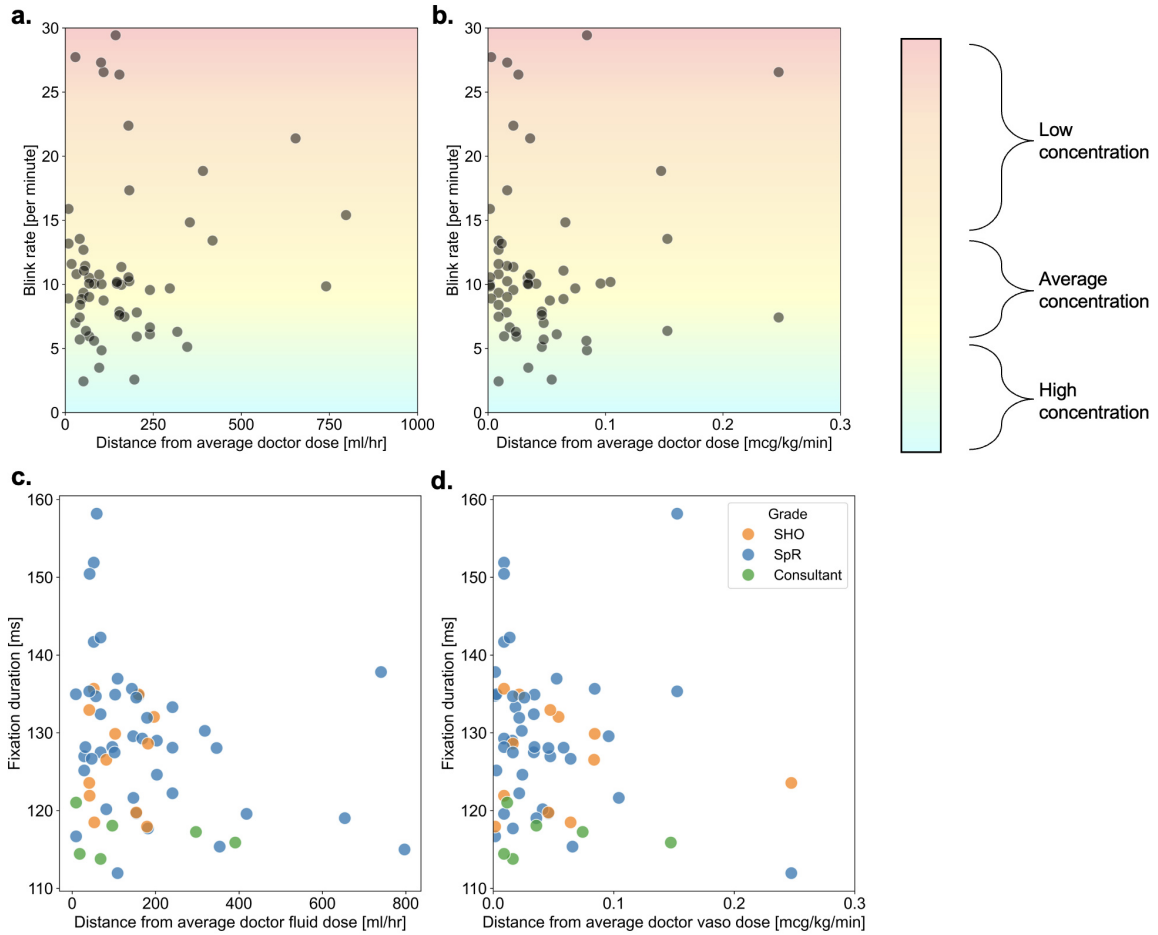


Fig. 5.5 **Correlation between practice variation and eye tracking metrics, for both fluid and vasopressor.** In (a), blink rate versus practice variation for fluid. In (b), blink rate versus practice variation for vasopressor. In (c), fixation duration versus practice variation for fluid. In (d), fixation duration versus practice variation for vasopressor. The degree of practice variation is defined by the per trial distance between a physician's prescription and the average prescription dose of all physicians. The background shading of plots (a) and (b) gives an indicator for expected blink rate depending on level of focus/concentration. Typical spontaneous blink rate is around 8-15 with lower levels suggesting higher focus/concentration and vice versa.

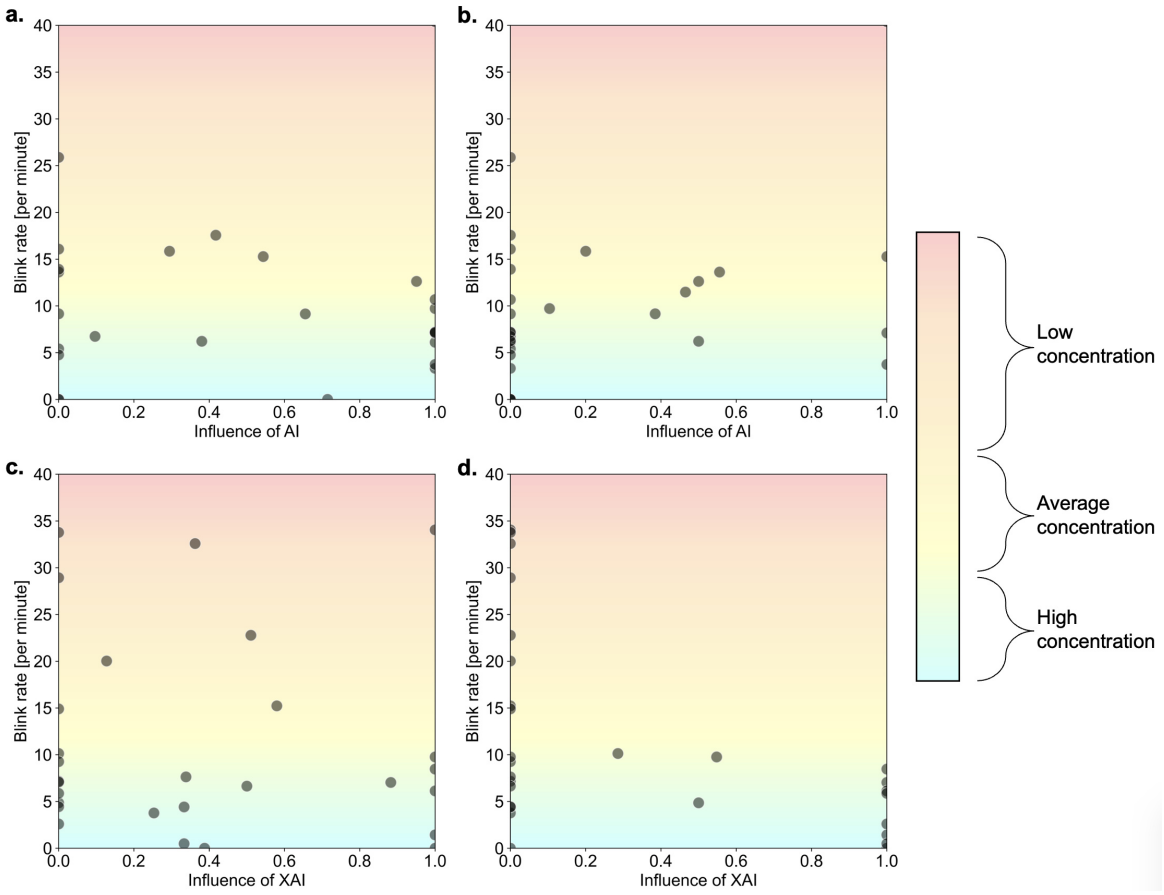


Fig. 5.6 **Blink rate by influence of AI/XAI for both fluid and vasopressor.** In (a), blink rate versus influence of AI for fluid. In (b), blink rate versus influence of AI for vasopressor. In (c), blink rate versus influence of XAI for fluid. In (d), blink rate versus influence of XAI for vasopressor. Typical spontaneous blink rate is around 8-15 with lower levels suggesting higher focus/concentration and vice versa.

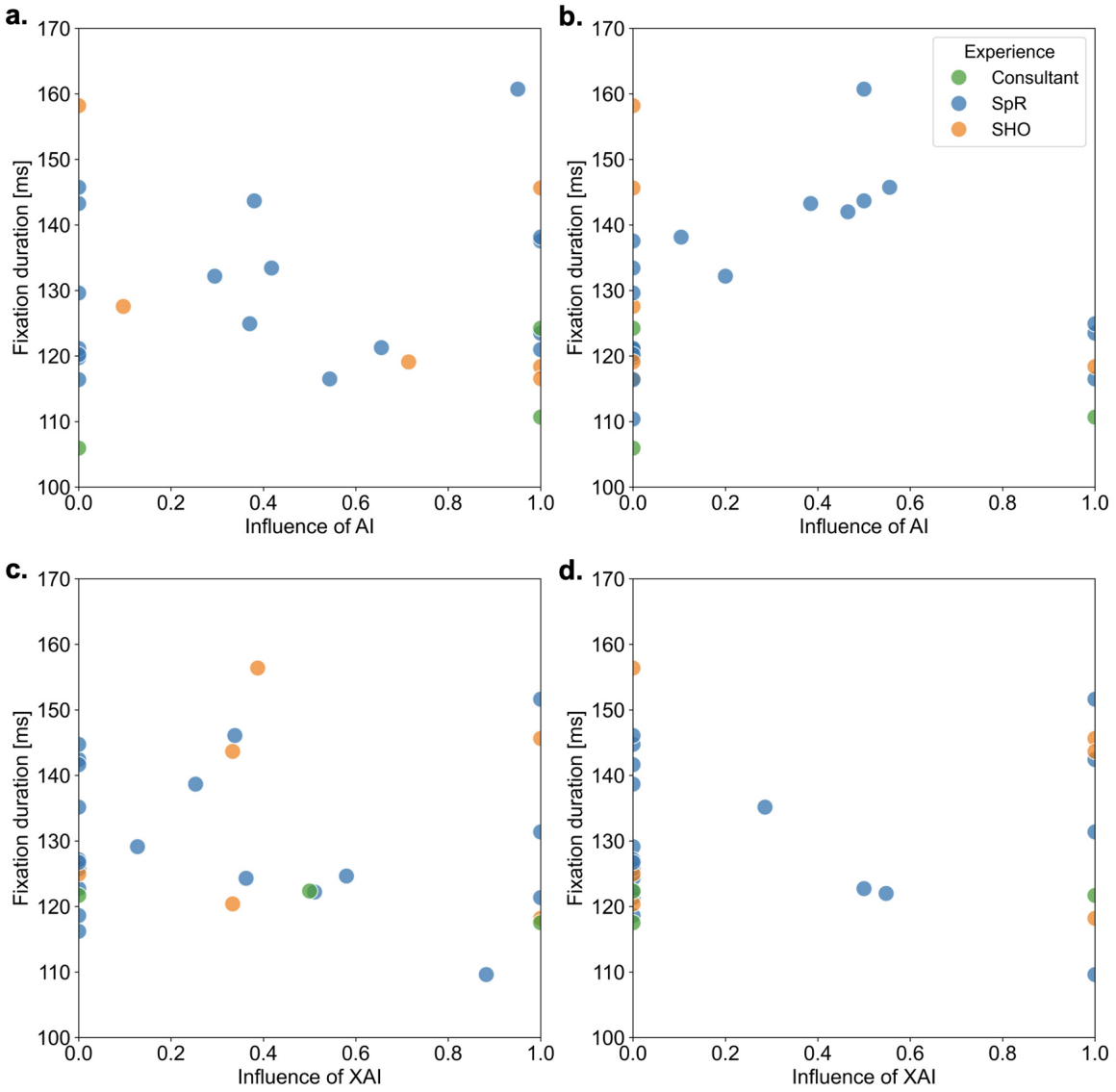


Fig. 5.7 **Fixation duration by influence of AI/XAI for both fluid and vasopressor.** In (a), fixation duration versus influence of AI for fluid. In (b), fixation duration versus influence of AI for vasopressor. In (c), fixation duration versus influence of XAI for fluid. In (d), fixation duration versus influence of XAI for vasopressor. Points are also categorised by grade of experience from lowest (SHO) to highest (consultant).

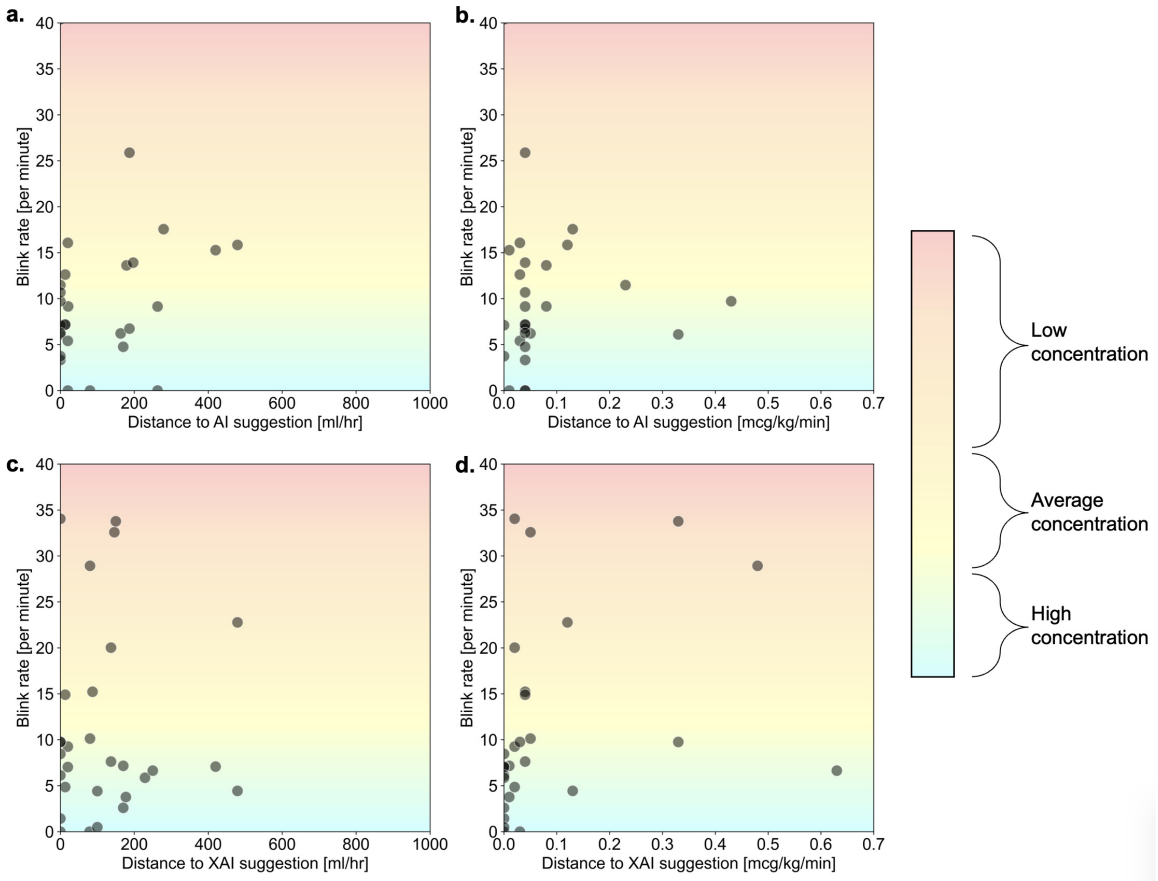


Fig. 5.8 **Blink rate by degree of adherence to AI/XAI for both fluid and vasopressor.** The degree of adherence to AI/XAI is defined by the per trial distance between a doctor's prescription and the AI/XAI suggestion. The larger the gap/distance between them, the lower the adherence to AI/XAI and vice versa. In (a), blink rate versus AI adherence for fluid. In (b), blink rate versus AI adherence for vasopressor. In (c), blink rate versus XAI adherence for fluid. In (d), blink rate versus XAI adherence for vasopressor. The background shading of the plots gives an indicator for expected blink rate depending on level of focus/concentration. Typical spontaneous blink rate is around 8-15 with lower levels suggesting higher focus/concentration and vice versa. Correlation measures: (AI fluid: r=0.18, p=0.36), (AI vasopressor: r=-0.11, p=0.56), (XAI fluid: r=0.24, p=0.20), (XAI vasopressor: r=0.16, p=0.41).

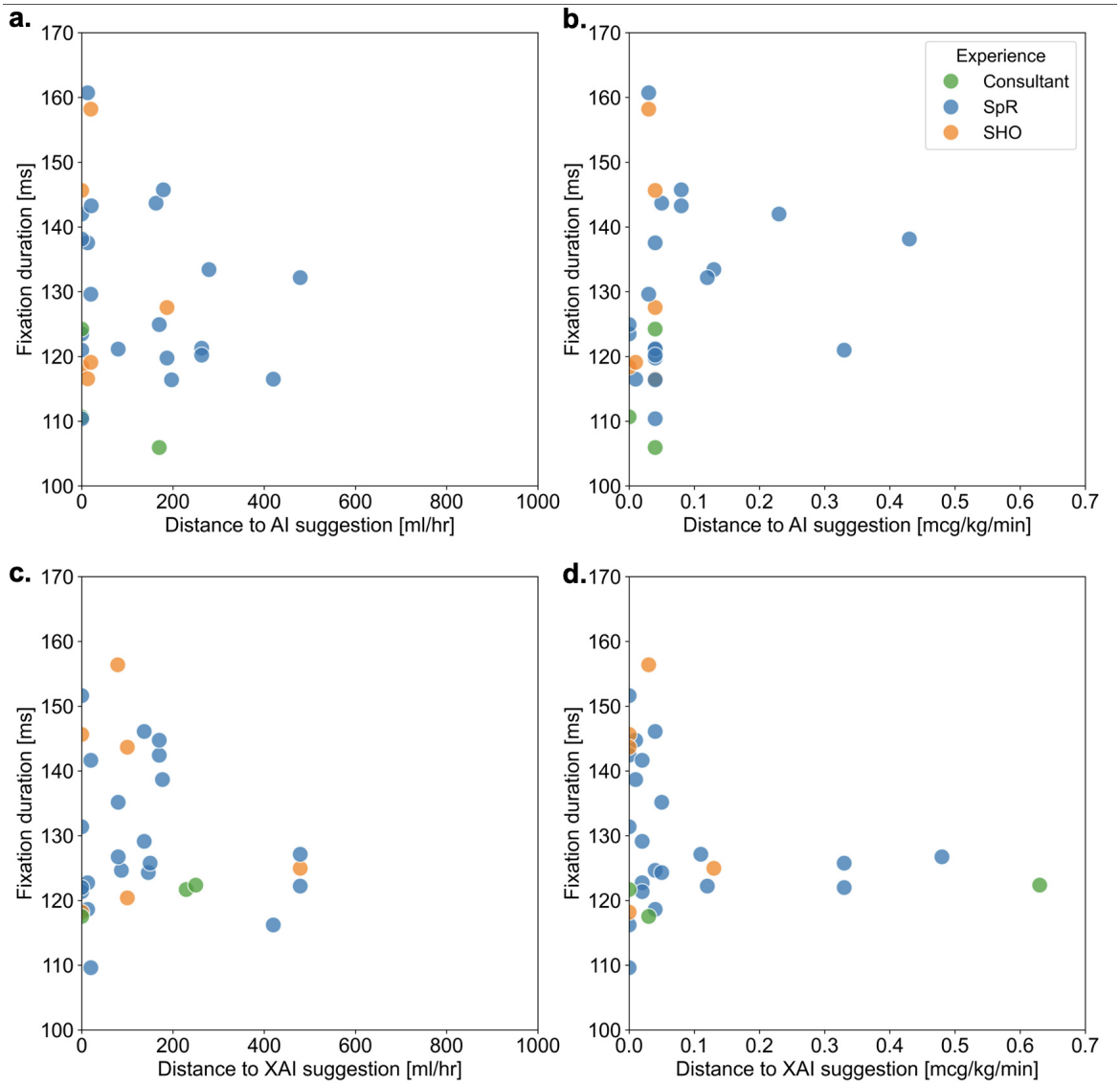


Fig. 5.9 **Fixation duration by degree of adherence to AI/XAI for both fluid and vasopressor.** The degree of adherence to AI/XAI is defined by the per trial distance between a doctor's prescription and the AI/XAI suggestion. The larger the gap/distance between them, the lower the adherence to AI/XAI and vice versa. In (a), fixation duration versus AI adherence for fluid. In (b), fixation duration versus AI adherence for vasopressor. In (c), fixation duration versus XAI adherence for fluid. In (d), fixation duration versus XAI adherence for vasopressor. Points are also categorised by grade of experience from lowest (SHO) to highest (consultant). Correlation measures: (AI fluid: r=-0.16, p=0.40), (AI vasopressor: r=0.20, p=0.30), (XAI fluid: r=-0.15, p=0.44), (XAI vasopressor: r=-0.27, p=0.15).

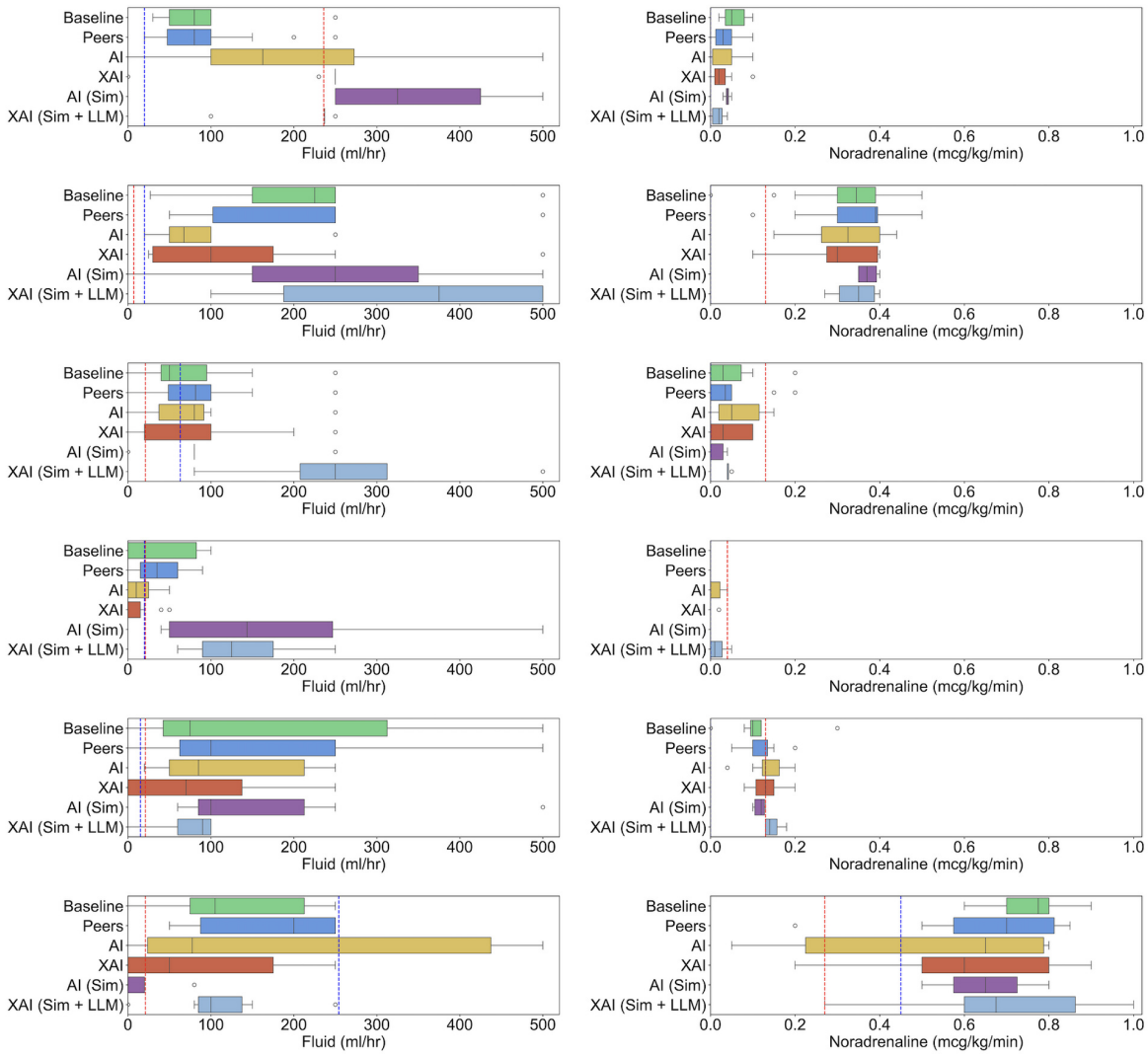


Fig. 5.10 **Boxplots of prescription dose distributions by trial arm comparing vignette and simulation (LLM) experiments.** For each boxplot, the centre line represents the median, box edges represent upper and lower quartiles, whiskers represent 1.5x inter-quartile range and diamonds are outliers. Blue dashed line represents the median of the peer distribution data (only available to those in the 'Peers' arm). Red dashed line represents the AI suggested dose (only available to those in the 'AI' or 'XAI' arms).

5.3.3 Influence of AI and XAI

The influence of AI and XAI for fluid were not significantly different (mean 0.48 for AI with SD 0.44 vs. mean 0.37 for XAI with SD 0.40, p=0.31). The same was true for influence of AI and XAI on vasopressor prescription (mean 0.26 for AI with SD 0.38 vs. mean 0.32 with SD 0.44, p=0.55). Given that the patient scenarios used were a subset of those from the vignette experiment in Chapter 2, we can also compare the AI and XAI prescription distributions from this experiment to the four arms from the vignette experiment (baseline, peer information, AI and feature importance XAI) for the same six patients, see Figure 5.10. The AI and XAI prescription distributions from this experiment are not infrequently different to those from the vignette experiment again highlighting that (i) experiment setting/fidelity is important and (ii) that the form of XAI makes a difference (i.e. not all XAI can be considered together under the same banner).

There was no strong evidence of correlation between blink rate and influence of AI/XAI regardless of drug (see Figure 5.6). Nor was there strong evidence of correlation between fixation duration and influence of AI/XAI regardless of drug (see Figure 5.7). There was no change when instead looking at the distance between a physician's final prescription (having had the opportunity to view the AI/XAI recommendation) and the value of the AI/XAI recommendation rather than influence of AI/XAI per se (see Figures 5.8 and 5.9). The absolute distance between final prescription and AI recommendation for fluid was 107 mL/hr and 133 mL/h respectively for AI and XAI conditions (SD 132 mL/h and 150mL/h respectively, p=0.48). For vasopressor, the distance between a physician's final prescription and the value of the AI/XAI recommendation was 0.09 mcg/kg/min and 0.10 mcg/kg/min respectively for AI and XAI conditions (SD 0.13 mcg/kg/min and 0.19 mcg/kg/min respectively, p=0.60).

5.3.4 Self-reported utility of XAI

The overall mean post-experiment utility rating for the XAI was 2.8 (SD 0.7) on a 0-4 scale with higher value implying the XAI was more useful (median 3, IQR 3-4). Although limited by sample size, there was no obvious visual correlation between eye-tracking metrics (fixation duration and blink rate) and XAI utility rating, see Figure 5.11a and b. Data were also collected on more specific elements of domains of XAI utility per the methods. This is broken down into a heatmap in Figure 5.12 with the vertical axis sorted by highest rated at the top (except for overall rating) and horizontal axis sorted by ICU experience from least to most. Timeliness was the most highly rated (mean 3.1, SD 0.8) despite at least 20-30 seconds lag required between responses while requirement for further guidelines to use in conjunction with the XAI was the lowest rated (mean 1.6, SD 1.4). Some of the dynamic elements of the XAI including ability to modify detail and integrate with a physician's own knowledge base were reasonably well rated relatively speaking (mean 2.6 and 2.9, SD 0.8 and 0.3 respectively). There was no obvious impact in this small sample of duration of ICU experience on the responses. Consistency and precision were both rated towards the lower end (these are known weaknesses of LLMs, especially consistency).

5.3.5 LLM conversations

The recorded conversation logs from the XAI condition were parsed and analysed. Excluding the initial prompt and initial explanation provided by the LLM, there were a total of 193 AI responses and 125 human responses over all 114 trials (19 physicians and 6 cases each). Median conversation length was 2 responses (IQR 2 to 4, range 2 to 12). Median word count per response was 11 (IQR 7 to 16, range 2 to 35). Mean polarity was 0.06 (SD 0.15) (i.e. very close to neutral). Mean subjectivity score was 0.21 (SD 0.23) again indicating relatively neutral and non-opinion based content (likely as most responses from physicians were typically questions rather than fully elucidating their

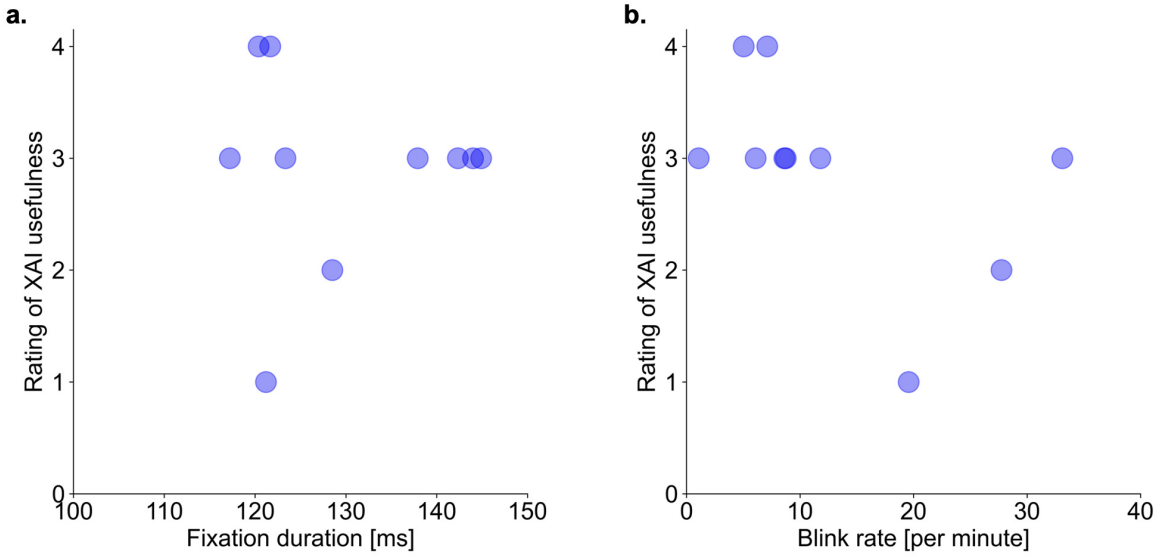


Fig. 5.11 **Overall self-reported XAI usefulness rating against eye-tracking metrics.** Fixation duration against XAI utility in (a) and blink rate against XAI utility in (b).

clinical thoughts and rationale). LDA analysis of the text was used to identify potential topics within the response dataset. Despite adjusting the number of components, the analysis yielded no significant insights or distinct niche topics, likely due to the small sample size limiting the depth of text available for analysis. Raw persuasion score was higher for the LLM side of the conversation than the human (mean 0.7 vs 0.1 respectively, SD 1.1 vs. 0.3). However, when normalised for word count given the human responses were much shorter the LLM normalised persuasion score was actually lower than the human (0.004 vs. 0.011, SD 0.007 vs. 0.032).

Selected examples of conversations follow with associated commentary (AI in orange, human physician in blue):

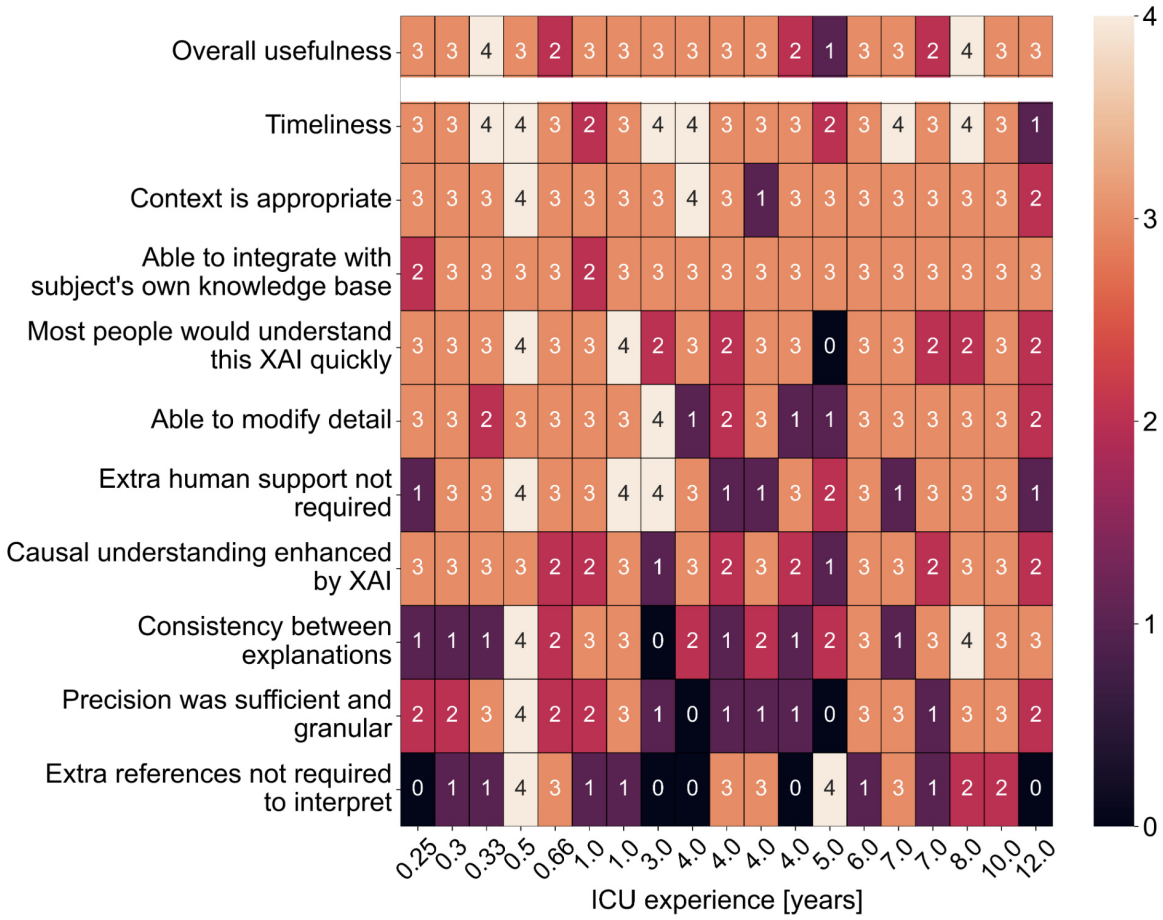


Fig. 5.12 **Heatmap for the breakdown of specific XAI rating questions.** All questions on a 0-4 scale with higher value implying XAI more positive (i.e. more useful/better/easier) ICU experience on x-axis increases from left to right. Excluding the 'overall usefulness' rating, all others appear in descending order according to mean rating (i.e. highest rated at the top and vice versa.

Subject C19, Case 1

what about a fluid bolus ?

Based on the information provided, I do not have enough data to recommend a fluid bolus at this time. The patient is currently demonstrating good urine output with the fluids they are receiving, suggesting adequate perfusion. Key factors are their lack of mechanical ventilation, normal lactate, and moderate fluid balance compared to the typical patient cluster.

While a fluid bolus could help improve perfusion, the data does not indicate signs of inadequate perfusion that would clearly warrant additional fluids. Their MAP is 63, which could be improved, but their good urine output and comparison to the patient cluster provides reassurance.

Moreover, the 90th percentile certainty recommendation balances optimizing circulation while mitigating risks of fluid overload. The state transition data shows only a 2% risk of death and 65% chance of remaining stable. Together this suggests stability without a need for aggressive interventions.

the urine output is still below 0.5ml/kg/hour

You're right, upon further review a urine output less than 0.5 ml/kg/hr does indicate this patient would likely benefit from additional fluid resuscitation.

Looking back at the patient cluster data, typical urine output is less than 100 ml/hr. And the most influential training case had a peak fluid balance of 0 L and low quartile urine output of 30 ml/hr.

Even though this recommendation has 90th percentile certainty, the patient's oliguria and positive fluid responsiveness compared to the cluster supports a trial of fluid bolus therapy. Their MAP is also on the low end.

While there are always risks of fluid overload, this patient's lack of mechanical ventilation and stability of vitals suggests their intravascular volume may be on the decline and could be boosted with bolus while continuing to monitor for fluid overload.

so we can give a trial of fluid bolus?

Yes, I agree that we can give a trial of fluid bolus for this patient. The key factors in this decision are:

- The patient's urine output is below 0.5 ml/kg/hr, indicating the potential need for further fluid resuscitation.
- The most influential training case had low urine output and fluid responsiveness to additional fluids. This suggests this patient may also respond positively.
- The patient cluster has a typical urine output of <100 ml/hr. So a trial of fluid bolus could help bring this patient's output closer to the cluster's level.
- The patient's MAP is on the lower end at 63. A fluid bolus could help improve blood pressure.
- While the 90th percentile certainty recommendation did not originally include a bolus, the patient's oliguria and comparisons to the cluster and training case suggest there is room for a fluid trial.

In the first exemplar chat (see 5.3.5), the LLM provides a good rationale for avoiding a fluid bolus initially before being further challenged by the user on the fact that the urine output is below 0.5ml/kg/hour (a commonly prescribed target by sepsis guidelines [223]). The urine output in this case was around 0.5ml/kg/hour (0.44, 0.54, 0.38 and 0.53 ml/kg/hour over the preceding four hours). The physician's question triggers the LLM to immediately accept the truth of the physician's statement and pivot to agreeing to a fluid bolus. While a further interrogation of the cluster data provided support in this case, the LLM was not able to confidently rebut the physician with a quantification of the fact that urine output was probably near enough to the target to be acceptable.

Subject C3, Case 1

Would additional investigation findings presently unavailable (renal function, ECHO, or CXR) change your thinking with this patient? And if so, would it be present to err on the side of being conservative first of all?

Based on the available information, additional investigations like renal function tests, echocardiography or chest x-rays would not significantly change my recommendations at this time. The patient comparisons to the surviving cluster case, the certainty level of the dose recommendation, and the low predicted risk of death provide reasonable confidence in proceeding with the recommended doses. However, it is wise to remain vigilant and conservative given this patient's advanced age, high comorbidity burden and early progression of sepsis. If the patient's hemodynamics worsen despite our interventions, that would prompt reconsideration and further investigative findings could better inform adjustments to the treatment plan. We aim to avoid under-resuscitation while also preventing fluid overload in this vulnerable patient. I would monitor this patient closely and use all clinical data available in regularly re-assessing appropriateness of the fluid and vasopressor doses. Please let me know if you have any other questions!

In the second exemplar chat (see 5.3.5), the LLM responds to a request for additional data (such as echocardiography and a chest radiograph) that would indeed be useful as not being pertinent to affecting the AI suggestion. This is a good example of the LLM confidently stating something that is, on the balance of probabilities, unlikely to be true. An echocardiogram suggesting a severely reduced ejection fraction would probably affect the propensity to aggressive fluid resuscitation in this patient. However, the LLM does agree with the physician that a conservative approach would be prudent (albeit in a long-winded response).

Subject C8, Case 4

Has this patient been eating and drinking in the days leading up to the operation?

Unfortunately the information provided does not specify details on this patient's oral intake in the days leading up to being mechanically ventilated for sepsis treatment. The prompt compared this patient to a typical cluster with a heart rate of 79, MAP of 76, and a norad dose of 0.11 mcg/kg/min. It also referenced training cases that survived with average norad doses from 0.09 to 0.13 mcg/kg/min. While further details on oral intake would be helpful for context, the recommendation aims to align care with the representative cluster and training cases based on the vitals, dosing and mortality data provided. I apologize, but without explicit oral intake details for this patient's history, I do not have enough context to factor that into my explanation. Please let me know if any additional clinical details can be provided.

What might happen if I give this patient another 1000ml of fluid over the next hour?

Given this patient is 91 years old with medium comorbidities and is moderately oliguric while mechanically ventilated for sepsis, aggressively giving an additional 1000mL fluid over the next hour beyond the recommended 21mL/hr risks volume overload complications.
The patient details compare to a cluster receiving less than 100mL/hr fluids on average. The influential training cases received peak fluid balances from 7.7-13.7L, far below an extra 1L pushing this patient's cumulative balance from -1L towards positive.
Moreover, the model has only a 1% chance of transitioning this patient to lower mortality states with more fluids. There is high confidence in the moderate 21mL/hr dose. Additional fluids risk volume overload and worsening respiratory failure without clear benefit. The safer approach is adhering to the recommended conservative fluids aligned to the cluster and training data. Transition chances and mortality risks reinforce this. Please let me know if you have any other questions!

In the third exemplar chat (see 5.3.5), the LLM appropriately states that while the additional information requested might be useful, it is unable to speculate on how it might affect the final treatment suggestion. When confronted with a relatively aggressive fluid suggestion by the physician, the LLM initially gives a good rebuttal

armed with the training and cluster data. However, it then confuses the peak fluid balances in the training cases with the current fluid balance of the current case (which was actually positive 1L rather than negative). It is unclear why this error/hallucination occured.

Subject C6, Case 4

What is the percentage of your moderate certainety for improving outcomes?

Based on the information provided, my "moderate certainty" recommendation has a 60% percentile rank for 'certainty' based on the Q value difference compared to alternative actions. Additionally, the next most likely state is to remain in the current state (58% chance). Of the remaining top next states, 5 are lower mortality states while 2 are higher. The risk of entering the death state is 1% while chance of entering survival state is 3%. So in summary, I have 60% confidence in my dose recommendation to improve outcomes, with a 58% chance of maintaining current state, and only a 1% risk of death and 3% chance of survival to discharge based on state transitions.

What about with my suggestions? 500mls/hr and 0.42 of norad. what are the outcomes from your dataset?

Based on the information provided in the original prompt, I unfortunately do not have enough data to definitively state the outcomes for a fluids rate of 500 ml/hr and a norad dose of 0.42 mcg/kg/min. A few key points:

- The reference cluster had average norad doses ranging from 0.09 to 0.13 mcg/kg/min, significantly lower than the suggested 0.42 mcg/kg/min. So there is no clear comparison data available.
- The two most influential training cases had peak fluid balances ranging from 7.7 to 13.7 liters total. A rate of 500 ml/hr would accumulate very quickly, and likely far exceed those peak balances from the influential cases.
- My moderate 60% certainty recommendation was based on typical doses for similar patients in the cluster, and positive outcomes in influential cases using lower doses. Deviating significantly from those doses does not have support in the available data.

And with her vitals becoming worse and correcponding metabolic changes such as increasing lactate, why are we cutting back rather than escalating?

Based on the patient's worsening vitals and increasing lactate levels, escalating care would be more appropriate than cutting back. A few key points:

- The reference cluster showed patients had an average lactate level of 3.4. So this patient's rising lactate to 3.0 indicates their condition is worsening compared to similar patients.
- The state transition data showed only a 1% chance of death if my original dose recommendations were followed. But with worsening vitals, the chance of death would likely increase if doses were reduced.
- The two influential training cases used norad doses up to 0.13 mcg/kg/min, higher than this patient's current 0.39 mcg/kg/min. So escalation aligned with those cases could be reasonable.
- My moderate 60% certainty was for improving outcomes. But with the patient deteriorating, certainty in further escalation should increase to meet the elevated risks.

In the fourth exemplar chat (see 5.3.5), the LLM is able to present a strong data-driven argument for the moderate certainty for the AI treatment suggestion. It is also able to use these data to make a reasonable attempt at forecasting the potential impact of a dose change suggested by the physician (albeit while clearly acknowledging the uncertainty of doing so). Ultimately, the LLM sides with the user when confronted with further evidence of clinical deterioration which prompts a repeat comparison with the reference cluster and influential training case data.

Subject C6, Case 6

Why such a drop in your norad dose? Shouldnt we think if weaning more slowly?

Based on the details provided, a few key factors support recommending a lower vasopressor dose of 0.27 mcg/kg/min compared to the patient's current 0.9 mcg/kg/min:
1) The patient is currently maintaining adequate blood pressure (MAP 79) and urine output (100-200ml/hr) on 0.9 mcg/kg/min norad. This indicates they may tolerate a lower dose without compromise.
2) Compared to the representative patient cluster, this patient is on a higher norad dose (0.9 vs 0.09 mcg/kg/min) despite having a comparable MAP. This suggests potential to decrease.
3) The most influential training case survived on a median norad dose of just 0.21 mcg/kg/min. As this case is most relevant, targeting a similar lower dose is reasonable.
4) This dose recommendation has 50th percentile certainty, so

What about if we met in the middle with fluids (i.e 90mls/hr) and go 0.6 with norad. What does this look like from your data?

Based on the details provided, going to an intermediate fluid rate of 90mls/hr and a norad dose of 0.6 mcg/kg/min balances multiple considerations:

1. A fluid rate of 90mls/hr is in between the recommended very restrictive 0 ml/hr and the patient's current 100-200ml/hr. This allows some continued fluid resuscitation while preventing significant positive fluid balance.
2. A norad dose of 0.6 mcg/kg/min is in between the patient's current 0.9 mcg/kg/min and the recommended 0.27 mcg/kg/min. This prevents an overly aggressive wean while moving towards the representative cluster dose.
3. The most influential training case survived on median 0.21 mcg/kg/min norad but peaked at 0.43 mcg/kg/min. Targeting 0.6 mcg/kg/min balances.

In the fifth and final exemplar chat (see 5.3.5), the LLM initially provides a counter-argument to the physician's suggestion of weaning noradrenaline more slowly. However, the physician then proposes a compromise with a dose between the AI and the initial physician dose with a request for the LLM to see what the impact might be. The LLM is agreeable to the compromise suggestion and provides corroborating evidence.

5.4 Discussion

5.4.1 Summary of results

This study provides key insights into the interaction behaviour of physicians with LLM-derived clinical XAI suggestions. First, there was no clear evidence of a significantly different influence from LLM-XAI on prescription decisions as compared to only an AI suggestion. Whether this was a true null effect or that there were too few subjects to determine either an overall effect (or indeed an effect within a subgroup of interest such as junior physicians) is unclear. Second, the experiment revealed a dynamic interaction between physicians and the LLM-XAI, where physicians sometimes challenged or provided additional context to AI suggestions leading to revised AI recommendations. This suggests a more nuanced human-AI collaboration is possible by highlighting the potential for AI to enhance clinical reasoning through offering alternative perspectives or highlighting less obvious data patterns. Third, eye-tracking metrics that proxy concentration suggested more cognitive attention was paid to the LLM-XAI screen in this experiment than to the four simultaneously presented XAIs in the previous experiment. Fourth, once again there were no obviously discernible patterns between eye-tracking metrics and either clinical practice variation or final AI/XAI prompted prescriptions though comparisons were severely limited by small sample size. Fifth, the AI and XAI prescription distributions from this experiment often differed from those in the vignette experiment, despite identical patient scenarios. This reinforces two key points: (i) experiment fidelity significantly influences physician behaviour and (ii) different types of XAI can have varied impacts on physicians. As a consequence, using 'XAI' as a blanket term risks overlooking the finer details of specific explainable AI techniques, much in the same way as describing all AI algorithms simply as 'AI' would obscure their important individual characteristics.

5.4.2 Limitations

Several limitations should be acknowledged before interpreting the findings of this experiment. First, the limitations discussed in the previous chapter (see section 4.4) on fidelity of the simulation suite still apply. The setting is easy to standardise for experiments and is far more life-like than vignettes but still does not compare to true real-world clinical practice. Second, the limitations in Chapter 4 (see section 4.4) on sample size and eye-tracking inclusions once again apply. In that experiment, 50% of the sample was available for eye-tracking analysis (19/38) compared to 53% (10/19) in this experiment. Once again, I therefore cannot definitively say whether non-significant findings are due to a lack of a true effect or an insufficient number of participants.

Third, as discussed in Chapter 2 (see section 2.4), the patient scenarios while attempting to represent a broad range of clinical presentations nonetheless do not capture the full diversity of sepsis presentations. In the vignette experiment, the format lent itself to more patient scenarios whereas the extra time required in the simulation suite makes it difficult to assess more than six patient scenarios for any given physician. If more physicians could have been recruited then a reasonable compromise would have been to sample from all 24 patient scenarios used in the vignette experiment as opposed to only six of them.

Fourth, the exact decision-making process and clinical rationale for the decisions made by physicians is not easy to elucidate as a think-aloud protocol was not used. We can infer some of the decision-making from a combination of the prescription changes and conversation logs but it is likely that a formal qualitative interview study with a sub-sample of physicians in this experiment would be a better way to shed light on the cognition of subjects during the task (reporting bias not withstanding). Fifth, while the AI-decision support tool was more realistically presented here on a laptop screen as opposed to the large TV display in the previous experiment, it was still not embedded into a dummy version of the EHR for maximum fidelity. Furthermore, despite

optimising the code for speed, there was a 10-30 second delay between physicians submitting a question and obtaining a reply from the LLM. This delay may have not only contributed to smaller conversation lengths or a reduced propensity to ask questions compared to faster commercial products, but it could also disrupt the natural flow of physician-AI interaction, potentially affecting the quality and depth of these conversations.

5.4.3 Implications

Notwithstanding these limitations, there are still important implications for those considering the use of LLMs for XAI-driven decision support tools. For example, some have proposed using LLM-XAI as an aggregator of multimodal and longitudinal EHR data. In a recent review, the authors argue that such a system would be a more dynamic and uncertainty-aware manner in which to convey AI-derived insights to clinicians [224]. The dynamic nature of the XAI in my experiment compared to the previous ones reveals how physicians and LLMs can engage in a collaborative decision-making process. This suggests that the integration of AI in clinical settings might not necessarily lead to simple automation bias, but rather a more balanced interaction where both human expertise and AI insights contribute to the final decision (and where the bias could act in both directions) [225].

The chat logs reveal instances where physicians challenge AI suggestions or provide additional context, sometimes leading to revised recommendations (albeit from the LLM rather than the AI Clinician RL agent). This dynamic interaction presents an opportunity for improving the underlying AI Clinician RL agent. By recording these conversations and converting the natural language exchanges into structured, RL-compatible data, we could create a feedback loop for the original RL agent [226]. This process would be analogous to how LLMs are fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [227], but with a crucial difference. Instead

of ranking LLM outputs, the feedback would come from the strength and frequency of physician counter-arguments or the need for recommendation revisions. This approach could allow the AI Clinician RL agent to learn from these clinical discussions, refining its understanding of when initial suggestions may be suboptimal or require additional context. Over time, this could lead to an RL agent that is more aligned with clinical expertise and better able to account for the intricacies of scenarios that might not be fully captured in the initial training data. For example, if the physician was able to perform a bedside echocardiogram, this would be useful additional data that was not present in the original training data set for the AI Clinician.

In some cases, the LLM provided data-driven arguments that might not have been immediately apparent to physicians. This suggests that AI could potentially enhance clinical reasoning by offering alternative perspectives or highlighting less obvious data patterns at the bedside [228]. In short, it offers an opportunity to prompt the physician to think again and might in fact be an antidote to some of the common cognitive biases that can warp clinical decision-making by improving reasoning skills [229]. There is therefore also a powerful educational opportunity too, particularly for junior physicians. This could be further enhanced by developing a system that remembers interactions with specific users and incorporates insights from risk tolerance and uncertainty questionnaires as in this experiment. For instance, if a physician is identified as overly cautious, the LLM could offer personalised suggestions that gently challenge this tendency, encouraging a more balanced approach [13]. Conversely, for physicians inclined toward aggressive prescriptions, the system could emphasise more conservative options. This personalised, adaptive learning tool could help trainees develop a more nuanced understanding of clinical decision-making, counterbalancing their pre-existing biases and might also be useful in situations where a senior decision-maker is not immediately available [230]. However, the not infrequent inconsistencies and errors in the LLM responses, such as confusing current patient data with historical

data, underscores the need for extreme caution to be exercised before any full clinical roll-out can be considered. For example, in one recent study by Cabral and colleagues, while an LLM was superior to physicians in clinical reasoning, it also had more instances of outright incorrect clinical reasoning than human physicians [231].

5.4.4 Conclusions

In summary, this study provides key insights into the interaction behaviour of physicians with LLM-derived clinical XAI suggestions. While there was no significant difference in prescription decisions or data verification behaviours between AI and XAI conditions, eye-tracking metrics suggested increased cognitive attention to LLM-XAI compared to the simpler and less dynamic XAIs in the previous experiment. The study also challenges concerns about automation bias with LLM-XAIs, revealing instances where physicians in fact influenced the LLM's recommendations through collaborative discussion and additional contextual input. This dynamic interaction highlights the potential for AI to enhance clinical reasoning by offering alternative perspectives and prompting physicians to reconsider their decisions. However, inconsistencies in LLM responses underscore the need for caution before real-world implementation.

Chapter 6

Discussion

6.1 Summary of findings from experiments

The experiments conducted in this thesis provide valuable insights into the influence of XAI-driven decision support systems on clinical decision-making (and specifically prescription decisions with an ICU sepsis context). The key findings can be summarised into six discrete themes:

Impact of AI recommendations on clinical decision-making

AI-driven recommendations significantly influenced prescription decisions as evidenced by dose shifts. However, the extent of this influence varied depending on the experimental setup (i.e. fidelity) and the nature of the AI suggestions (i.e. whether accompanied by XAI and if so, which type).

Evaluation methodology of healthcare-AI

The experiments demonstrated both the feasibility and added value of using non-standard methodologies such as eye-tracking and high-fidelity simulation environments to evaluate physician interaction with AI-driven decision support tools. These methods

provided objective measures of attention and cognitive load which offer insights that self-reports alone would not capture. While this necessarily implies extra resource cost (both financial and time) to performing experiments, it is likely to be worthwhile in terms of presenting study estimates that are closer to real-world practice.

Perception of explainable AI

The experiments revealed an unexpected relationship between XAI and its subjectively perceived utility. There was a consistent lack of correlation between physician self-reported usefulness of XAI and its actual influence on decision-making or the objectively measured attention devoted by physicians to explanations. This core finding calls into question the reliability of self-reports as a metric for evaluating the impact of XAI on clinical decision-making. This is not to say that self-reported measures of XAI utility should not be used at all. Rather it suggests that self-reports should be only one (and arguably not the primary) metric among several balanced evaluation metrics for XAI in clinical settings. There are many situations in which human participants report satisfaction despite this contradicting their true feelings [232–234].

Human-AI behavioural interaction

The experiments with more dynamic interaction methods shed light on how physician-AI behaviour evolves over the course of several responses. For example, physicians were observed challenging AI suggestions and providing additional context, leading to revised recommendations. This interaction highlights the potential for AI to enhance clinical reasoning by offering alternative perspectives and highlighting data patterns that physicians may not have been cognisant of.

Variability in XAI effectiveness

The experiments reinforced that different types of XAI can have varied impacts on physician decision-making. Simple feature importance explanations, simultaneous presentation of multiple XAI types, and LLM-derived explanations each showed distinct patterns of influence (or absence thereof).

Influence of physician characteristics on interaction with AI

Contrary to expectations, neither clinical experience nor attitudes toward AI showed strong correlations with the influence of AI-driven treatment suggestions. The most likely explanation for this phenomenon was that variability in adherence to AI treatment suggestions might instead be more dependent on specific patient scenarios. However, some subtle differences were observed such as (i) senior doctors exhibiting higher cognitive load when reviewing data-rich information sources and (ii) junior physicians being more likely to seek second opinions when conflicted with 'unsafe' AI suggestions.

6.2 Strengths and limitations

The key strengths of the experimental approaches employed in this thesis include the following:

Methodological innovation

Eye-tracking has been used very infrequently with regard to clinical AI evaluation (and especially so within free-movement high-fidelity simulation settings). The validation of this approach represents a significant advancement in evaluating human-AI interaction in healthcare. This approach provided objective measures of attention and cognitive load, offering insights that self-reports alone could not capture.

Evaluation trifecta

The studies in this thesis evaluated the full trifecta of what physicians *did* (i.e. their actual prescription decisions), *how* they did it (i.e. what they looked at during the decision-making process) and what they *said* (i.e. their subjective ratings for the explanations after the experiment ended). This provides a truly multi-faceted view of the interaction between physicians and AI-driven decision support tools that is absent from most studies in high-stakes decision settings.

Diverse XAI modalities

The inclusion of various XAI types (feature importance, mortality based predictions, Q-value differences, influential training examples, LLM-derived explanations) allowed for conclusions that more broadly reflect XAI approaches rather than being limited to the weaknesses of any single approach. This contributes significantly to the broader understanding of XAI in healthcare especially in acute settings such as intensive care.

However, as candidly acknowledged throughout the chapters of this thesis, there were also several limitations that fall into four broad themes.

Fidelity

I tried to address the fidelity limitations of the vignette by upgrading to a simulation suite but this nonetheless is only one prior step before real-world studies and is in no way a substitution for such work.

Eye-tracking issues

This is a novel technology to apply to the clinical XAI setting and therefore there are likely many technical improvements that are possible. Efforts to address the significant

drop-off in participants includable due to technical issues with pupil identification would be valuable in increasing the sample size for future clinical studies.

Sample size

While the sample recruited for the vignette is comparatively large compared to some similar studies, the small sample sizes in the simulation studies limited statistical power and the ability to draw definitive conclusions from non-significant findings. I pursued both convenience sampling and targeted advertising within the local NHS trusts of the simulation suite (i.e. Imperial College Healthcare NHS Trust) but it was nonetheless difficult to recruit physicians, even when offering considerable financial remuneration for participant inconvenience. Unlike the vignette experiment, it was less feasible to recruit physicians who were based far from Imperial College Healthcare NHS Trust as they were generally less willing to travel to the simulation suite for the experiment. However, the counter-argument is that given the novelty in this field, it was important to learn from multiple studies rather than concentrating the limited recruitment potential into a single large study.

Insight into physician cognitive processes and decision-making

Ultimately, the evidence in the studies within this thesis tend to focus on the distal end of the decision-making process (visual attention on ROIs and prescription decisions). The intricate cognitive process of weighing up information and gradually coming to a decision is comparatively opaque. Although the LLM experiment does suggest some intermediate decision-making steps in so far as physicians asked questions to the LLM, this is nonetheless patchy and not uniformly present for all the patient scenarios that were undertaken. Chapter 3 dealt with the disadvantages of think aloud protocols which would be the natural alternative (or complementary approach) to what I executed in these studies. Another approach would be dedicated qualitative studies with thematic

interviews. Ultimately, the well acknowledged difficulty in truly elucidating the human decision-making process without impact from reporting and hindsight biases is difficult to counteract.

6.3 Overarching implications and future directions

The findings from this thesis offer significant insights into the potential and challenges of implementing XAI-CDSS in a clinical setting such as the intensive care unit (which combines the triumvirate of high-stakes decision-making, high clinical uncertainty and large volumes of multimodal data). Three overarching themes emerge from my work, each with important implications for future research and real-world implementation of XAI-CDSS.

First, these studies reveal a nuanced interplay between clinicians and AI-driven decision support tools that goes beyond simple automation bias or algorithmic aversion. The dynamic nature of human-AI collaboration, especially evident in the LLM-based XAI experiment, suggests that these systems can potentially serve as partners in the decision-making process rather than mere tools. This collaborative approach has the potential to highlight alternative clinical perspectives and less obvious data patterns to decision-makers, potentially mitigating common cognitive biases that occur in medical decision-making. However, there is also a need to design XAI systems that can adapt to individual characteristics in physician personality, such as risk tolerance and appetite for uncertainty, if we truly want to minimise the harmful impact of cognitive biases in decision-making. The ethics of this latter aspect (i.e. specifically targeting AI suggestions in an individualised manner to clinician users based on their personality) has not been well explored and may be more controversial than it initially appears.

Second, the studies in this thesis highlight the complexity and difficulty of evaluating XAI effectiveness in clinical settings as opposed to straightforward non-clinical XAI tasks evaluated over the internet on any willing member of the public. I have demonstrated

how traditional metrics such as self-reports are inadequate in capturing the true impact of XAI on decision-making. The innovative use of eye-tracking technology in this work offers a promising avenue for more objective assessment of XAI influence. However, further work is required before true gaze-related behavioural phenotypes can be validated for clinical use (and ideally also for providing feedback to an RL-agent on how end-users respond to RL treatment suggestions). In Chapter 1, I quoted Doshi-Velez and Kim who made the important point that the highest level in their taxonomy of XAI evaluation was application-grounded evaluation using real humans and real tasks and that *"we as a community respect the time and effort involved to do such evaluations, and also demand high standards of experimental design when such evaluations are performed"* [56]. The work in this thesis has attempted to live up to this aspiration for real-world XAI evaluation.

Third, the studies in this thesis support the educational implications for XAI in clinical settings. Junior physicians, who are still developing their clinical reasoning skills, may require extra support (or at least the availability of such support for situations where the AI suggestion is grossly discordant to their preferred course of action). To the extent that in Chapter 1 an advantage of AI-CDSS was its ability to improve the resource constraints associated with limited clinical experts, some of the findings in this study challenge whether that advantage will be realised in the near-term. However, there is potential for a CDSS to grow beyond simply providing isolated decision-support and instead be trained to a user, much like a social media algorithm is adapted to present a user with homepage content that keeps them maximally engaged.

What is not yet clear is how the educational training of early- and mid-stage clinical trainees will need to adapt in the age of AI-driven clinical decision support tools to equip physicians with the necessary knowledge on two fronts: (i) how they might be led astray by LLMs so that they have the necessary cognitive immunisation against automation and aversion bias and (ii) the broad-brush core workings of AI-tools such

as RL agents and LLMs so that physicians understand what the limitations are (and thereby when they should be less trusting of advice from such systems).

6.4 Conclusions

The findings from this thesis have important implications for all healthcare AI stakeholders involved in attempting to bring XAI-CDSS toward real-world implementation. Specifically, the findings in my thesis support an emphasis on high-fidelity simulation, neuro-behavioural objective markers of attention such as real-time eye-tracking and a more nuanced understanding of the interaction behaviour between physicians and AI in high-stakes healthcare settings. As the field progresses, it will be crucial to use these insights to design larger-scale, real-world studies that evaluate the impact of XAI in a live clinical environment.

References

[1] Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. Lancet (London, England). 2020 Jan;395(10219):200-11. Place: England.

[2] Angus DC, van der Poll T. Severe sepsis and septic shock. The New England journal of medicine. 2013 Aug;369(9):840-51. Place: United States.

[3] Prescott HC, Iwashyna TJ. Improving Sepsis Treatment by Embracing Diagnostic Uncertainty. Annals of the American Thoracic Society. 2019 Apr;16(4):426-9. Place: United States.

[4] Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. The New England Journal of Medicine. 2001 Nov;345(19):1368-77.

[5] Byrne L, Van Haren F. Fluid resuscitation in human sepsis: Time to rewrite history? Annals of Intensive Care. 2017 Dec;7(1):4.

[6] Waechter J, Kumar A, Lapinsky SE, Marshall J, Dodek P, Arabi Y, et al. Interaction between fluids and vasoactive agents on mortality in septic shock: a multicenter, observational study. Critical Care Medicine. 2014 Oct;42(10):2158-68.

[7] Marik PE, Linde-Zwirble WT, Bittner EA, Sahatjian J, Hansell D. Fluid administration in severe sepsis and septic shock, patterns and outcomes: an analysis of a large national database. Intensive care medicine. 2017 May;43(5):625-32. Place: United States.

[8] Li Y, Li H, Zhang D. Timing of norepinephrine initiation in patients with septic shock: a systematic review and meta-analysis. Critical care (London, England). 2020 Aug;24(1):488. Place: England.

[9] Permpikul C, Tongyoo S, Viarasilpa T, Trainarongsakul T, Chakorn T, Udompanturak S. Early Use of Norepinephrine in Septic Shock Resuscitation (CENSER). A Randomized Trial. American Journal of Respiratory and Critical Care Medicine. 2019 May;199(9):1097-105.

[10] Russell JA, Gordon AC, Walley KR. Early May Be Better: Early Low-Dose Norepinephrine in Septic Shock. American Journal of Respiratory and Critical Care Medicine. 2019 May;199(9):1049-51.

[11] National Heart, Lung, and Blood Institute Prevention and Early Treatment of Acute Lung Injury Clinical Trials Network, Shapiro NI, Douglas IS, Brower RG, Brown SM, Exline MC, et al. Early Restrictive or Liberal Fluid Management for Sepsis-Induced Hypotension. The New England Journal of Medicine. 2023 Feb;388(6):499-510.

[12] Saugel B, Vincent JL. Cardiac output monitoring: how to choose the optimal method for the individual patient. Current Opinion in Critical Care. 2018 Jun;24(3):165-72.

[13] Chen Y, Nagendran M. Thinking About Thinking: A Prescription for Healthcare Improvement. CRC Press; 2024. Available from: https://books.google.com/books?hl=en&lr=&id=f0INEQAAQBAJ&oi=fnd&pg=PT8&dq=info:90i_SdvanWcJ:scholar.google.com&ots=QhS6-6HuX4&sig=dm7D1m3yQYWzGahvrMwfVI98Hfs.

[14] Mamdani M, Slutsky AS. Artificial intelligence in intensive care medicine. Intensive Care Medicine. 2021 Feb;47(2):147-9.

[15] Fleuren LM, de Bruin DP, Tonutti M, Lalisang RCA, Elbers PWG, Dutch ICU Data Sharing Collaborators. Large-scale ICU data sharing for global collaboration: the first 1633 critically ill COVID-19 patients in the Dutch Data Warehouse. Intensive Care Medicine. 2021 Apr;47(4):478-81.

[16] Gutierrez G. Artificial Intelligence in the Intensive Care Unit. Critical Care (London, England). 2020 Mar;24(1):101.

[17] Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016 May;3:160035. Place: England.

[18] Ridley S, Morris S. Cost effectiveness of adult intensive care in the UK. Anaesthesia. 2007 Jun;62(6):547-54.

[19] Wong DJN, Popham S, Wilson AM, Barneto LM, Lindsay HA, Farmer L, et al. Postoperative critical care and high-acuity care provision in the United Kingdom, Australia, and New Zealand. British Journal of Anaesthesia. 2019 Apr;122(4):460-9.

[20] Wunsch H, Angus DC, Harrison DA, Linde-Zwirble WT, Rowan KM. Comparison of medical admissions to intensive care units in the United States and United Kingdom. American Journal of Respiratory and Critical Care Medicine. 2011 Jun;183(12):1666-73.

[21] Harris S, Singer M, Sanderson C, Grieve R, Harrison D, Rowan K. Impact on mortality of prompt admission to critical care for deteriorating ward patients: an instrumental variable analysis using critical care bed strain. Intensive Care Medicine. 2018 May;44(5):606-15.

[22] Ananadaciva S. Critical care services in the english NHS; 2020. Available from: https://www.kingsfund.org.uk/publications/critical-care-services-nhs.

[23] NHS-England. Critical care bed capacity and urgent operations cancelled 2019-20 data; 2020. Available from: https://www.england.nhs.uk/statistics/statistical-work-areas/critical-care-capacity/critical-care-bed-capacity-and-urgent-operations-cancelled-2019-20-data/.

[24] Proudfoot AG, O'Brien B, Schilling R, Gould DW, McGlennan A, Collaborating authors. Rapid establishment of a COVID-19 critical care unit in a convention centre: the Nightingale Hospital London experience. Intensive Care Medicine. 2021 Mar;47(3):349-51.

[25] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019 Jan;25(1):44-56.

[26] Patel BV, Haar S, Handslip R, Auepanwiriyakul C, Lee TML, Patel S, et al. Natural history, trajectory, and management of mechanically ventilated COVID-19 patients in the United Kingdom. Intensive Care Medicine. 2021 May;47(5):549-65.

[27] Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nature Medicine. 2020 Mar;26(3):364-73.

[28] Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Medicine. 2020 Mar;46(3):383-400.

[29] van de Sande D, van Genderen ME, Huiskens J, Gommers D, 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. 2021 Jul;47(7):750-60. Place: United States.

[30] Liu S, See KC, Ngiam KY, Celi LA, Sun X, Feng M. Reinforcement Learning for Clinical Decision Support in Critical Care: Comprehensive Review. Journal of Medical Internet Research. 2020 Jul;22(7):e18477.

[31] Zampieri FG, Salluh JIF, Azevedo LCP, Kahn JM, Damiani LP, Borges LP, et al. ICU staffing feature phenotypes and their relationship with patients' outcomes: an unsupervised machine learning analysis. Intensive Care Medicine. 2019 Nov;45(11):1599-607.

[32] Seymour CW, Kennedy JN, Wang S, Chang CCH, Elliott CF, Xu Z, et al. Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. JAMA. 2019 May;321(20):2003-17. Available from: https://doi.org/10.1001/jama.2019.5791.

[33] Schrittwieser J, Antonoglou I, Hubert T, Simonyan K, Sifre L, Schmitt S, et al. Mastering Atari, Go, chess and shogi by planning with a learned model. Nature. 2020 Dec;588(7839):604-9.

[34] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature. 2017 Oct;550(7676):354-9.

[35] Vinyals O, Babuschkin I, Czarnecki WM, Mathieu M, Dudzik A, Chung J, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature. 2019 Nov;575(7782):350-4.

[36] Bennett CC, Hauser K. Artificial intelligence framework for simulating clinical decision-making: a Markov decision process approach. Artificial Intelligence in Medicine. 2013 Jan;57(1):9-19.

[37] Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. Nature Medicine. 2019 Jan;25(1):16-8.

[38] Padmanabhan R, Meskin N, Haddad WM. Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. Mathematical Biosciences. 2017 Nov;293:11-20.

[39] Mishra A, Soni U, Huang J, al e. Why? Why not? When? Visual explanations of agent behavior in reinforcement learning. arXiv:210402818. 2021.

[40] Bothe MK, Dickens L, Reichel K, Tellmann A, Ellger B, Westphal M, et al. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. Expert Review of Medical Devices. 2013 Sep;10(5):661-73.

[41] Prasad N, Cheng LF, Chivers C, al e. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. arXiv:170406300. 2017.

[42] Peine A, Hallawa A, Bickenbach J, Dartmann G, Fazlic LB, Schmeink A, et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. NPJ digital medicine. 2021 Feb;4(1):32. Place: England.

[43] Eghbali N, Alhanai T, Ghassemi MM. Patient-Specific Sedation Management via Deep Reinforcement Learning. Frontiers in Digital Health. 2021;3:608893.

[44] Bica I, Jarrett D, Hüyük A, al e. Learning "what-if" explanations for sequential decision-making. arXiv:200713531. 2021.

[45] Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nature medicine. 2018 Nov;24(11):1716-20. Place: United States.

[46] Jeter R, Lehman LW, Josef C, Shashikumar S, Nemati S. Learning to Treat Hypotensive Episodes in Sepsis Patients Using a Counterfactual Reasoning Framework. medRxiv; 2021. Pages: 2021.03.03.21252863. Available from: https://www.medrxiv.org/content/10.1101/2021.03.03.21252863v1.

[47] Zhang K, Wang Y, Du J, al e. Identifying decision points for safe and interpretable reinforcement learning in hypotension treatment. arXiv:210103309. 2021.

[48] Roggeveen L, El Hassouni A, Ahrendt J, Guo T, Fleuren L, Thoral P, et al. Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. Artificial Intelligence in Medicine. 2021 Feb;112:102003.

[49] Li L, Komorowski M, Faisal AA. The actor search tree critic (ASTC) for off-policy POMDP learning in medical decision making. arXiv:180511548. 2018.

[50] Li L, Komorowski M, Faisal AA. Optimizing sequential medical treatments with auto-encoding heuristic search in pomdps. arXiv:190507465. 2019.

[51] Jones C, Thornton J, Wyatt JC. Enhancing trust in clinical decision support systems: a framework for developers. BMJ health & care informatics. 2021 Jun;28(1):e100247.

[52] Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. Applied Sciences. 2021;11(11).

[53] Lu M, Shahn Z, Sow D, al e. Is deep reinforcement learning ready for practical applications in healthcare? A sensitivity analysis of duel-DDQN for hemodynamic management in sepsis patients. arXiv:200504301. 2020.

[54] Arrieta AB, Díaz-Rodríguez N, Ser JD, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58:82-115. Available from: https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[55] Biran O, Cotton CV. Explanation and justification in machine learning: a survey. IJCAI-17 workshop on explainable AI (XAI) 2017; 8, 1. 2017.

[56] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:170208608. 2017.

[57] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. 2019 Feb;267:1-38. Available from: https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[58] Lundberg S, Lee S. A unified approach to interpreting model predictions. arXiv:170507874. 2017.

[59] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Kdd '16. New York, NY, USA: Association for Computing Machinery; 2016. p. 1135-44. Number of pages: 10 Place: San Francisco, California, USA. Available from: https://doi.org/10.1145/2939672.2939778.

[60] Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature Communications. 2020 Jul;11(1):3852.

[61] Puiutta E, Veith EM. Explainable reinforcement learning: a survey. arXiv:200506247. 2020.

[62] Wells L, Bednarz T. Explainable AI and Reinforcement Learning-A Systematic Review of Current Approaches and Trends. Frontiers in Artificial Intelligence. 2021;4:550030.

[63] Beyret B, Shafti A, Faisal AA. Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation. IROS 2019; 5014-5019. 2019.

[64] Xu X, Huang T, Wei P, al e. Hierarchical reinforcement learning in StarCraft II with human expertise in subgoals selection. arXiv:200803444. 2020.

[65] Chou YL, Moreira C, Bruza P, al e. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. arXiv: 210304244. 2021.

[66] Madumal P, Miller T, Sonenberg L, al e. Explainable reinforcement learning through a causal lens. arXiv:190510958. 2019.

[67] Madumal P, Miller T, Sonenberg L, al e. Distal explanations for model-free explainable reinforcement learning. arXiv:200110284. 2020.

[68] Jacobs M, Pradier MF, McCoy THJ, Perlis RH, Doshi-Velez F, Gajos KZ. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. Translational psychiatry. 2021 Feb;11(1):108. Place: United States.

[69] Lage I, Chen E, He J, al e. An evaluation of the human-interpretability of explanation. arXiv:190200006. 2019.

[70] Tonekaboni S, Joshi S, McCradden MD, al e. What clinicians want: Contextualizing explainable machine learning for clinical end use. Proc Mach Learn Res 2019; 106, 359-80. 2019.

[71] Ronan CE, Crable EL, Drainoni ML, Walkey AJ. The impact of clinical decision support systems on provider behavior in the inpatient setting: A systematic review and meta-analysis. Journal of Hospital Medicine. 2022;17(5):368-83. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jhm.12825. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jhm.12825.

[72] Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. BMJ. 2020 Sep;370:m3216. Publisher: British Medical Journal Publishing Group Section: Research. Available from: https://www.bmj.com/content/370/bmj.m3216.

[73] Laka M, Milazzo A, Merlin T. Factors That Impact the Adoption of Clinical Decision Support Systems (CDSS) for Antibiotic Management. International Journal of Environmental Research and Public Health. 2021 Jan;18(4):1901. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/1660-4601/18/4/1901.

[74] Westerbeek L, Ploegmakers KJ, de Bruijn GJ, Linn AJ, van Weert JCM, Daams JG, et al. Barriers and facilitators influencing medication-related CDSS acceptance according to clinicians: A systematic review. International Journal of Medical Informatics. 2021 Aug;152:104506. Available from: https://www.sciencedirect.com/science/article/pii/S1386505621001325.

[75] Vasey B, Ursprung S, Beddoe B, Taylor EH, Marlow N, Bilbro N, et al. Association of Clinician Diagnostic Performance With Machine Learning–Based Decision Support Systems: A Systematic Review. JAMA Network Open. 2021 Mar;4(3):e211276. Available from: https://doi.org/10.1001/jamanetworkopen.2021.1276.

[76] Wood W. Attitude change: persuasion and social influence. Annual Review of Psychology. 2000;51:539-70.

[77] Nolan JM, Schultz PW, Cialdini RB, Goldstein NJ, Griskevicius V. Normative social influence is underdetected. Personality & Social Psychology Bulletin. 2008 Jul;34(7):913-23.

[78] Schultz PW, Nolan JM, Cialdini RB, Goldstein NJ, Griskevicius V. The constructive, destructive, and reconstructive power of social norms. Psychological Science. 2007 May;18(5):429-34.

[79] Hallsworth M, Chadborn T, Sallis A, Sanders M, Berry D, Greaves F, et al. Provision of social norm feedback to high prescribers of antibiotics in general

practice: a pragmatic national randomised controlled trial. Lancet (London, England). 2016 Apr;387(10029):1743-52.

[80] Barnett Michael L , Olenski Andrew R , Jena Anupam B . Opioid-Prescribing Patterns of Emergency Physicians and Risk of Long-Term Use. New England Journal of Medicine. 2017;376(7):663-73. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMsa1610524. Available from: https://www.nejm.org/doi/full/10.1056/NEJMsa1610524.

[81] Burton JH, Hoppe JA, Echternach JM, Rodgers JM, Donato M. Quality Improvement Initiative to Decrease Variability of Emergency Physician Opioid Analgesic Prescribing. Western Journal of Emergency Medicine. 2016 May;17(3):258-63. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4899055/.

[82] Boyle KL, Cary C, Dizitzer Y, Novack V, Jagminas L, Smulowitz PB. Reduction of opioid prescribing through the sharing of individual physician opioid prescribing practices. The American Journal of Emergency Medicine. 2019 Jan;37(1):118-22.

[83] Suffoletto B, Landau A. Nudging Emergency Care Providers to Reduce Opioid Prescribing Using Peer Norm Comparison Feedback: A Pilot Randomized Trial. Pain Medicine. 2020 Jul;21(7):1393-9. Available from: https://doi.org/10.1093/pm/pnz314.

[84] Yuan CT, Bradley EH, Nembhard IM. A mixed methods study of how clinician 'super users' influence others during the implementation of electronic health records. BMC Medical Informatics and Decision Making. 2015 Apr;15(1):26. Available from: https://doi.org/10.1186/s12911-015-0154-6.

[85] Conlon JJ, Mani M, Rao G, Ridley MW, Schilbach F. Not Learning from Others [Working Paper]. National Bureau of Economic Research; 2022. Available from: https://www.nber.org/papers/w30378.

[86] Gaertig C, Simmons JP. Do People Inherently Dislike Uncertain Advice? Psychological Science. 2018 Apr;29(4):504-20. Publisher: SAGE Publications Inc. Available from: https://doi.org/10.1177/0956797617739369.

[87] Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ digital medicine. 2021 Feb;4(1):31. Place: England.

[88] Agarwal N, Moehring A, Rajpurkar P, Salz T. Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology [Working Paper]. National Bureau of Economic Research; 2023. Available from: https://www.nber.org/papers/w31422.

[89] Choudhury A. Factors influencing clinicians' willingness to use an AI-based clinical decision support system. Frontiers in Digital Health. 2022 Aug;4. Publisher:

Frontiers. Available from: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2022.920662/full.

[90] Shamszare H, Choudhury A. Clinicians' Perceptions of Artificial Intelligence: Focus on Workload, Risk, Trust, Clinical Decision Making, and Clinical Integration. Healthcare. 2023 Jan;11(16):2308. Number: 16 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/2227-9032/11/16/2308.

[91] Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P. Heterogeneity and predictors of the effects of AI assistance on radiologists. Nature Medicine. 2024 Mar;30(3):837-49. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41591-024-02850-w.

[92] Gaube S, Suresh H, Raue M, Lermer E, Koch TK, Hudecek MFC, et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. Scientific reports. 2023 Jan;13(1):1383. Place: England.

[93] Schwartz JM, George M, Rossetti SC, Dykes PC, Minshall SR, Lucas E, et al. Factors Influencing Clinician Trust in Predictive Clinical Decision Support Systems for In-Hospital Deterioration: Qualitative Descriptive Study. JMIR Human Factors. 2022 May;9(2):e33960. Company: JMIR Human Factors Distributor: JMIR Human Factors Institution: JMIR Human Factors Label: JMIR Human Factors Publisher: JMIR Publications Inc., Toronto, Canada. Available from: https://humanfactors.jmir.org/2022/2/e33960.

[94] Wysocki O, Davies JK, Vigo M, Armstrong AC, Landers D, Lee R, et al. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. Artificial Intelligence. 2023 Mar;316:103839. Available from: https://www.sciencedirect.com/science/article/pii/S0004370222001795.

[95] Amann J, Vetter D, Blomberg SN, Christensen HC, Coffee M, Gerke S, et al. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. PLOS Digital Health. 2022 Feb;1(2):e0000016. Publisher: Public Library of Science. Available from: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000016.

[96] Badal K, Lee CM, Esserman LJ. Guiding principles for the responsible development of artificial intelligence tools for healthcare. Communications Medicine. 2023 Apr;3(1):1-6. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s43856-023-00279-9.

[97] Brown DG, Wobst HJ, Kapoor A, Kenna LA, Southall N. Clinical development times for innovative drugs. Nature Reviews Drug Discovery. 2021 Nov;21(11):793-4. Bandiera_abtest: a Cg_type: From The Analyst's Couch

Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/d41573-021-00190-9.

[98] Joshi G, Jain A, Araveeti SR, Adhikari S, Garg H, Bhandari M. FDA-Approved Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices: An Updated Landscape. Electronics. 2024 Jan;13(3):498. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/2079-9292/13/3/498.

[99] Nannini L, Alonso-Moral JM, Catala A, Lama M, Barro S. Operationalizing Explainable AI in the EU Regulatory Ecosystem. IEEE Intelligent Systems. 2024:1-13. Conference Name: IEEE Intelligent Systems. Available from: https://ieeexplore.ieee.org/abstract/document/10489989.

[100] Chung NC, Chung H, Lee H, Brocki L, Chung H, Dyer G. False Sense of Security in Explainable Artificial Intelligence (XAI). arXiv; 2024. ArXiv:2405.03820 [cs]. Available from: http://arxiv.org/abs/2405.03820.

[101] Nahar N, Rowlett J, Bray M, Omar ZA, Papademetris X, Menon A, et al. Regulating Explainability in Machine Learning Applications – Observations from a Policy Design Experiment. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. FAccT '24. New York, NY, USA: Association for Computing Machinery; 2024. p. 2101-12. Available from: https://doi.org/10.1145/3630106.3659028.

[102] Schneeberger D, Röttger R, Cabitza F, Campagner A, Plass M, Müller H, et al. The Tower of Babel in Explainable Artificial Intelligence (XAI). In: Holzinger A, Kieseberg P, Cabitza F, Campagner A, Tjoa AM, Weippl E, editors. Machine Learning and Knowledge Extraction. Cham: Springer Nature Switzerland; 2023. p. 65-81.

[103] Stogiannos N, Malik R, Kumar A, Barnes A, Pogose M, Harvey H, et al. Black box no more: a scoping review of AI governance frameworks to guide procurement and adoption of AI in medical imaging and radiotherapy in the UK. British Journal of Radiology. 2023 Dec;96(1152):20221157. Available from: https://doi.org/10.1259/bjr.20221157.

[104] Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine. 2022 May;28(5):924-33.

[105] McNamara SL, Yi PH, Lotter W. The clinician-AI interface: intended use and explainability in FDA-cleared AI devices for medical image interpretation. npj Digital Medicine. 2024 Mar;7(1):1-5. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41746-024-01080-1.

[106] EU. Article 6: Classification Rules for High-Risk AI Systems | EU Artificial Intelligence Act; 2024. Available from: https://artificialintelligenceact.eu/article/6/.

[107] EU. Article 50: Transparency Obligations for Providers and Users of Certain AI Systems and GPAI Models | EU Artificial Intelligence Act; 2024. Available from: https://artificialintelligenceact.eu/article/50/.

[108] Sovrano F, Sapienza S, Palmirani M, Vitali F. Metrics, Explainability and the European AI Act Proposal. J. 2022 Mar;5(1):126-38. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/2571-8800/5/1/10.

[109] Kiseleva A. Making AI's transparency transparent: notes on the EU Proposal for the AI Act; 2021. Available from: https://europeanlawblog.eu/2021/07/29/making-ais-transparency-transparent-notes-on-the-eu-proposal-for-the-ai-act/.

[110] Gyevnar B, Ferguson N, Schafer B. Bridging the Transparency Gap: What Can Explainable AI Learn From the AI Act?; 2024. Available from: https://ar5iv.labs.arxiv.org/html/2302.10766.

[111] Petkovic D. It is Not "Accuracy vs. Explainability"—We Need Both for Trustworthy AI Systems. IEEE Transactions on Technology and Society. 2023 Mar;4(1):46-53. Conference Name: IEEE Transactions on Technology and Society. Available from: https://ieeexplore.ieee.org/abstract/document/10029927.

[112] Papadakis T, Christou IT, Ipektsidis C, Soldatos J, Amicone A. Explainable and transparent artificial intelligence for public policymaking. Data & Policy. 2024 Jan;6:e10. Available from: https://www.cambridge.org/core/journals/data-and-policy/article/explainable-and-transparent-artificial-intelligence-for-public-policymaking/51D4C6E27CFDEB3CD19EC5E1A6F4FAE7#.

[113] Bertrand A, Belloum R, Eagan JR, Maxwell W. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. Oxford United Kingdom: ACM; 2022. p. 78-91. Available from: https://dl.acm.org/doi/10.1145/3514094.3534164.

[114] Jones-Jang SM, Park YJ. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. Journal of Computer-Mediated Communication. 2023 Jan;28(1):zmac029. Available from: https://doi.org/10.1093/jcmc/zmac029.

[115] Vicente L, Matute H. Humans inherit artificial intelligence biases. Scientific Reports. 2023 Oct;13(1):15737. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41598-023-42384-8.

[116] FDA. Artificial Intelligence and Machine Learning in Software as a Medical Device. FDA. 2024. Publisher: FDA. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.

[117] MHRA. Impact of AI on the regulation of medical products; 2024. Available from: https://www.gov.uk/government/publications/impact-of-ai-on-the-regulation-of-medical-products.

[118] EMA. The use of Artificial Intelligence (AI) in the medicinal product lifecycle; 2023. Available from: https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle.

[119] Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. npj Digital Medicine. 2022 Oct;5(1):1-15. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41746-022-00699-2.

[120] Cervantes JA, López S, Rodríguez LF, Cervantes S, Cervantes F, Ramos F. Artificial Moral Agents: A Survey of the Current Status. Science and engineering ethics. 2020 Apr;26(2):501-32. Place: England.

[121] van Wynsberghe A, Robbins S. Critiquing the Reasons for Making Artificial Moral Agents. Science and engineering ethics. 2019 Jun;25(3):719-35.

[122] Wikipedia. Nazi human experimentation; 2021. Available from: https://en.wikipedia.org/wiki/Nazi_human_experimentation.

[123] Wikipedia. Tuskegee Syphilis Study; 2021. Available from: https://en.wikipedia.org/wiki/Tuskegee_Syphilis_Study.

[124] Association WM. WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects; 2018. Available from: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/.

[125] (GMC) GMC. Good Medical Practice; 2021. Available from: https://www.gmc-uk.org/ethical-guidance/ethical-guidance-for-doctors/good-medical-practice.

[126] Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making. 2020 Nov;20(1):310.

[127] (NHS) NHS. Liberating the NHS: No decision about me, without me; 2012. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/216980/

Liberating-the-NHS-No-decision-about-me-without-me-Government-response.pdf.

[128] London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. The Hastings Center report. 2019 Jan;49(1):15-21. Place: United States.

[129] Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. NPJ digital medicine. 2020;3:47.

[130] Bigman YE, Gray K. People are averse to machines making moral decisions. Cognition. 2018 Dec;181:21-34. Place: Netherlands.

[131] Fund K. Improving the allocation of health resources in England: How to decide who gets what; 2013. Available from: https://www.kingsfund.org.uk/sites/default/files/field/field_publication_file/improving-the-allocation-of-health-resources-in-england-kingsfund-apr13.pdf.

[132] Scientist N. Discriminating algorithms: 5 times AI showed prejudice; 2018. Available from: https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/.

[133] (CQC) CQC. The duty of candour; 2016. Available from: https://www.cqc.org.uk/sites/default/files/Duty-of-Candour-2016-CQC-joint-branded.pdf.

[134] Carter SM, Rogers W, Win KT, Frazer H, Richards B, Houssami N. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. Breast (Edinburgh, Scotland). 2020 Feb;49:25-32.

[135] Islam KM, Anggondowati T, Deviany PE, Ryan JE, Fetrick A, Bagenda D, et al. Patient preferences of chemotherapy treatment options and tolerance of chemotherapy side effects in advanced stage lung cancer. BMC cancer. 2019 Aug;19(1):835.

[136] Zidaru T, Morrow EM, Stockley R. Ensuring patient and public involvement in the transition to AI-assisted mental health care: A systematic scoping review and agenda for design justice. Health Expectations: An International Journal of Public Participation in Health Care and Health Policy. 2021 Aug;24(4):1072-124.

[137] Morrow E, Zidaru-Bărbulescu T, Stockley R. 4 priorities to reaffirm patient voice in the coming era of AI healthcare; 2021. Available from: https://blogs.lse.ac.uk/impactofsocialsciences/2021/06/18/4-priorities-to-reaffirm-patient-voice-in-the-coming-era-of-ai-healthcare/.

[138] Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision

making. npj Digital Medicine. 2023 Nov;6(1):206. Available from: https://doi.org/10.1038/s41746-023-00955-z.

[139] Jussupow E, Spohrer K, Heinzl A, Gawlitza J. Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. Information Systems Research. 2021 Sep;32(3):713-35. Publisher: INFORMS. Available from: https://doi.org/10.1287/isre.2020.0980.

[140] Moussaïd M, Kämmer JE, Analytis PP, Neth H. Social influence and the collective dynamics of opinion formation. PloS one. 2013;8(11):e78433. Place: United States.

[141] Yaniv I. Receiving other people's advice: Influence and benefit. Organizational Behavior and Human Decision Processes. 2004 Jan;93(1):1-13. Available from: https://www.sciencedirect.com/science/article/pii/S0749597803001018.

[142] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016 Feb;315(8):801-10. Place: United States.

[143] Garriga R, Mas J, Abraha S, Nolan J, Harrison O, Tadros G, et al. Machine learning model to predict mental health crises from electronic health records. Nature medicine. 2022 Jun;28(6):1240-8. Place: United States.

[144] Lam JY, Shimizu C, Tremoulet AH, Bainto E, Roberts SC, Sivilay N, et al. A machine-learning algorithm for diagnosis of multisystem inflammatory syndrome in children and Kawasaki disease in the USA: a retrospective model development and validation study. The Lancet Digital health. 2022 Oct;4(10):e717-26. Place: England.

[145] Laatifi M, Douzi S, Ezzine H, Asry CE, Naya A, Bouklouze A, et al. Explanatory predictive model for COVID-19 severity risk employing machine learning, shapley addition, and LIME. Scientific reports. 2023 Apr;13(1):5481. Place: England.

[146] Deng H, Eftekhari Z, Carlin C, Veerapong J, Fournier KF, Johnston FM, et al. Development and Validation of an Explainable Machine Learning Model for Major Complications After Cytoreductive Surgery. JAMA network open. 2022 May;5(5):e2212930. Place: United States.

[147] Triberti S, Torre DL, Riva G. The Artificial Face (ART-F) Project: Addressing the Problem of Interpretability, Interface, and Trust in Artificial Intelligence. Cyberpsychology, behavior and social networking. 2023 Apr;26(4):318-20. Place: United States.

[148] Bienefeld N, Boss JM, Lüthy R, Brodbeck D, Azzati J, Blaser M, et al. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. NPJ digital medicine. 2023 May;6(1):94. Place: England.

[149] Festor P, Luise G, Komorowski M, Faisal AA. Enabling risk-aware Reinforcement Learning for medical interventions through uncertainty decomposition. arXiv preprint arXiv:210907827. 2021.

[150] Shafti A, Derks V, Kay H, Faisal AA. The response shift paradigm to quantify human trust in AI recommendations. arXiv preprint arXiv:220208979. 2022.

[151] Micocci M, Borsci S, Thakerar V, Walne S, Manshadi Y, Edridge F, et al. Attitudes towards Trusting Artificial Intelligence Insights and Factors to Prevent the Passive Adherence of GPs: A Pilot Study. Journal of clinical medicine. 2021 Jul;10(14). Place: Switzerland.

[152] Brenner PS, DeLamater J. Lies, damned lies, and survey self-reports? Identity as a cause of measurement bias. Social psychology quarterly. 2016 Dec;79(4):333-54. Place: United States.

[153] Li AC, Kannry JL, Kushniruk A, Chrimes D, McGinn TG, Edonyabo D, et al. Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. International Journal of Medical Informatics. 2012 Nov;81(11):761-72. Available from: https://www.sciencedirect.com/science/article/pii/S138650561200041X.

[154] Anjara SG, Janik A, Dunford-Stenger A, Kenzie KM, Collazo-Lorduy A, Torrente M, et al. Examining explainable clinical decision support systems with think aloud protocols. PLOS ONE. 2023 Sep;18(9):e0291443. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0291443.

[155] Fox MC, Ericsson KA, Best R. Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. Psychological Bulletin. 2011 Mar;137(2):316-44.

[156] Ball LJ, Richardson BH. Eye movement in user experience and human–computer interaction research. In: Stuart S, editor. Eye Tracking: Background, Methods, and Applications. New York, NY: Springer US; 2022. p. 165-83. Available from: https://doi.org/10.1007/978-1-0716-2391-6_10.

[157] Harston JA, Faisal AA. Methods and models of eye-tracking in natural environments. In: Stuart S, editor. Eye Tracking: Background, Methods, and Applications. New York, NY: Springer US; 2022. p. 49-68. Available from: https://doi.org/10.1007/978-1-0716-2391-6_4.

[158] Auepanwiriyakul C, Harston A, Orlov P, Shafti A, Faisal AA. Semantic fovea: real-time annotation of ego-centric videos with gaze context. In: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications. Warsaw, Poland: Association for Computing Machinery; 2018. p. Article 87. Type:

10.1145/3204493.3208349. Available from: https://doi.org/10.1145/3204493.3208349.

[159] Makrigiorgos A, Shafti A, Harston A, Gerard J, Faisal AA. Human visual attention prediction boosts learning & performance of autonomous driving agents. arXiv preprint arXiv:190905003. 2019.

[160] Ranti C, Jones W, Klin A, Shultz S. Blink rate patterns provide a reliable measure of individual engagement with scene content. Scientific reports. 2020 May;10(1):8267. Place: England.

[161] Harston JA, Auepanwiriyakul C, Faisal A. Prediction of visual attention in embodied real-world tasks. Journal of Vision. 2021 Sep;21(9):2741-1. Available from: https://doi.org/10.1167/jov.21.9.2741.

[162] Causse M, Lancelot F, Maillant J, Behrend J, Cousy M, Schneider N. Encoding decisions and expertise in the operator's eyes: Using eye-tracking as input for system adaptation. International Journal of Human-Computer Studies. 2019 May;125:55-65. Available from: https://www.sciencedirect.com/science/article/pii/S1071581918306888.

[163] Pauszek JR. An introduction to eye tracking in human factors healthcare research and medical device testing. Human Factors in Healthcare. 2023 Jun;3:100031. Available from: https://www.sciencedirect.com/science/article/pii/S2772501422000288.

[164] Labs P. Pupil labs core hardware; 2024. Available from: https://docs.pupil-labs.com/core/hardware/.

[165] Labs P. Pupil labs core specifications & performance; 2024. Available from: https://pupil-labs.com/products/core/tech-specs.

[166] Kassner M, Patera W, Bulling A. Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. arXiv; 2014. ArXiv:1405.0006 [cs]. Available from: http://arxiv.org/abs/1405.0006.

[167] Sirois S, Brisson J. Pupillometry. Wiley Interdisciplinary Reviews Cognitive Science. 2014 Nov;5(6):679-92.

[168] Eckstein MK, Guerra-Carrillo B, Miller Singley AT, Bunge SA. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? Developmental Cognitive Neuroscience. 2017 Jun;25:69-91.

[169] Chidi-Egboka NC, Jalbert I, Chen J, Briggs NE, Golebiowski B. Blink rate measured in situ decreases while reading from printed text or digital devices, regardless of task duration, difficulty, or viewing distance. Investigative ophthalmology & visual science. 2023 Feb;64(2):14. Place: United States.

[170] Chen X, Wang S, Yang X, Yu C, Ni F, Yang J, et al. Utilizing artificial intelligence-based eye tracking technology for screening ADHD symptoms in children. Frontiers in Psychiatry. 2023 Nov;14. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2023.1260031/full.

[171] Liu Z, Yang Z, Gu Y, Liu H, Wang P. The effectiveness of eye tracking in the diagnosis of cognitive disorders: a systematic review and meta-analysis. PLOS ONE. 2021 Jul;16(7):e0254059. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0254059.

[172] Tourangeau M, Tanoubi I, Perron R, Bélanger M Sodoké K, Drolet P, et al. Comparing the visual perception of novice and expert clinicians: An exploratory simulation-based study. Anaesthesia Critical Care & Pain Medicine. 2020 Apr;39(2):247-9. Available from: https://www.sciencedirect.com/science/article/pii/S2352556820300266.

[173] Chen HE, Sonntag CC, Pepley DF, Prabhu RS, Han DC, Moore JZ, et al. Looks can be deceiving: Gaze pattern differences between novices and experts during placement of central lines. The American Journal of Surgery. 2019 Feb;217(2):362-7. Available from: https://www.sciencedirect.com/science/article/pii/S0002961018306809.

[174] Li S, Duffy MC, Lajoie SP, Zheng J, Lachapelle K. Using eye tracking to examine expert-novice differences during simulated surgical training: a case study. Computers in Human Behavior. 2023 Jul;144:107720. Available from: https://www.sciencedirect.com/science/article/pii/S0747563223000717.

[175] King AJ, Hochheiser H, Visweswaran S, Clermont G, Cooper GF. Eye-tracking for clinical decision support: a method to capture automatically what physicians are viewing in the EMR. AMIA Summits on Translational Science Proceedings. 2017 Jul;2017:512-21. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5543363/.

[176] Visweswaran S, King AJ, Tajgardoon M, Calzoni L, Clermont G, Hochheiser H, et al. Evaluation of eye tracking for a decision support application. JAMIA Open. 2021 Jul;4(3):ooab059. Available from: https://doi.org/10.1093/jamiaopen/ooab059.

[177] Lohmeyer Q, Schiess C, Wendel Garcia PD, Petry H, Strauch E, Dietsche A, et al. Effects of tall man lettering on the visual behaviour of critical care nurses while identifying syringe drug labels: a randomised in situ simulation. BMJ quality & safety. 2023 Jan;32(1):26-33.

[178] Hsieh C, Luís A, Neves J, Nobre IB, Sousa SC, Ouyang C, et al. EyeXNet: Enhancing abnormality detection and diagnosis via eye-tracking and x-ray fusion. Machine Learning and Knowledge Extraction. 2024 Jun;6(2):1055-71. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/2504-4990/6/2/48.

[179] Karargyris A, Kashyap S, Lourentzou I, Wu JT, Sharma A, Tong M, et al. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for AI development. Scientific Data. 2021 Mar;8(1):92. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41597-021-00863-5.

[180] Shamyuktha RS, Amudha J, Aiswariya Milan K. A machine learning framework for classification of expert and non-experts radiologists using eye gaze data. In: 2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE). MANGALORE, India: IEEE; 2022. p. 314-20. Available from: https://ieeexplore.ieee.org/document/10054277/.

[181] Buettner R. Cognitive workload of humans using artificial intelligence systems: Towards objective measurement applying eye-tracking technology. In: Timm IJ, Thimm M, editors. KI 2013: Advances in Artificial Intelligence. Berlin, Heidelberg: Springer; 2013. p. 37-48.

[182] Cao S, Huang CM. Understanding user reliance on AI in assisted decision-making. Proc ACM Hum-Comput Interact. 2022;6(CSCW2):Article 471. Publisher: Association for Computing Machinery Type: 10.1145/3555572. Available from: https://doi.org/10.1145/3555572.

[183] Cioffi GM, Pinilla-Echeverri N, Sheth T, Sibbald MG. Does artificial intelligence enhance physician interpretation of optical coherence tomography: insights from eye tracking. Frontiers in Cardiovascular Medicine. 2023 Dec;10. Publisher: Frontiers. Available from: https://www.frontiersin.org/articles/10.3389/fcvm.2023.1283338.

[184] Arsiwala-Scheppach LT, Castner NJ, Rohrer C, Mertens S, Kasneci E, Cejudo Grano de Oro JE, et al. Impact of artificial intelligence on dentists' gaze during caries detection: a randomized controlled trial. Journal of Dentistry. 2024 Jan;140:104793. Available from: https://www.sciencedirect.com/science/article/pii/S0300571223003792.

[185] Marshall SP. The index of cognitive activity: Measuring cognitive workload. In: Proceedings of the IEEE 7th conference on Human Factors and Power Plants. IEEE; 2002. p. 7-7. Available from: https://ieeexplore.ieee.org/abstract/document/1042860/?casa_token=B-lvAM06unIAAAAA:ysdRMlooDv3FuDJ-3r1lpMoBI0TgPF4uGJDTqUIu8_wG2vMRHzo5LCgBa_mj6Aax5Wuo8qR7.

[186] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al.. Segment Anything. arXiv; 2023. ArXiv:2304.02643 [cs]. Available from: http://arxiv.org/abs/2304.02643.

[187] Gegenfurtner A, Seppänen M. Transfer of expertise: An eye tracking and think aloud study using dynamic medical visualizations. Computers & Education. 2013 Apr;63:393-403. Available from: https://www.sciencedirect.com/science/article/pii/S0360131513000055.

[188] Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations. npj Digital Medicine. 2024 Aug;7(1):1-10. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41746-024-01200-x.

[189] Festor P, Nagendran M, Gordon AC, Faisal AA, Komorowski M. Safety of human-AI cooperative decision-making within intensive care: A physical simulation study. PLOS Digital Health. 2025 Feb;4(2):e0000726. Publisher: Public Library of Science. Available from: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000726.

[190] Platt ML, Huettel SA. Risky business: the neuroeconomics of decision making under uncertainty. Nature neuroscience. 2008 Apr;11(4):398-403. Place: United States.

[191] Toomaari A, Cafazzo J. Contrasts in Safety Management: Safety-Critical Industries vs. Healthcare; 2019. Available from: https://www.csagroup.org/wp-content/uploads/CSA-Group-Research-Safety-Management-Healthcare.pdf.

[192] Buçinca Z, Lin P, Gajos KZ, Glassman EL. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In: Proceedings of the 25th International Conference on Intelligent User Interfaces; 2020. p. 454-64. ArXiv:2001.08298 [cs]. Available from: http://arxiv.org/abs/2001.08298.

[193] Schoonderwoerd TAJ, Jorritsma W, Neerincx MA, van den Bosch K. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. International Journal of Human-Computer Studies. 2021 Oct;154:102684. Available from: https://www.sciencedirect.com/science/article/pii/S1071581921001026.

[194] Cato DL, Murray M. Use of simulation training in the intensive care unit. Critical care nursing quarterly. 2010 Mar;33(1):44-51. Place: United States.

[195] Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA. 2011 Sep;306(9):978-88. Place: United States.

[196] Chen Y, Harris S, Rogers Y, Ahmad T, Asselbergs FW. Nudging within learning health systems: next generation decision support to improve cardiovascular care. European Heart Journal. 2022 Feb;43(13):1296-306. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8971005/.

[197] UKRI. Eligibility of technology readiness levels (TRL); 2024. Available from: https://www.ukri.org/councils/stfc/guidance-for-applicants/check-if-youre-eligible-for-funding/eligibility-of-technology-readiness-levels-trl/.

[198] Festor P, Jia Y, Gordon AC, Faisal AA, Habli I, 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 Jul;29(1). Place: England.

[199] Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. This Looks Like That: Deep Learning for Interpretable Image Recognition. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc.; 2019. Available from: https://proceedings.neurips.cc/paper_files/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html.

[200] Morrison K, Spitzer P, Turri V, Feng M, Kühl N, Perer A. The Impact of Imperfect XAI on Human-AI Decision-Making. Proceedings of the ACM on Human-Computer Interaction. 2024 Apr;8(CSCW1):1-39. ArXiv:2307.13566 [cs]. Available from: http://arxiv.org/abs/2307.13566.

[201] Panigutti C, Beretta A, Giannotti F, Pedreschi D. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. New Orleans, LA, USA: Association for Computing Machinery; 2022. p. Article 568. Type: 10.1145/3491102.3502104. Available from: https://doi.org/10.1145/3491102.3502104.

[202] Price WNn, Gerke S, Cohen IG. Potential Liability for Physicians Using Artificial Intelligence. JAMA. 2019 Nov;322(18):1765-6. Place: United States.

[203] Cadario R, Longoni C, Morewedge CK. Understanding, explaining, and utilizing medical artificial intelligence. Nature human behaviour. 2021 Dec;5(12):1636-42. Place: England.

[204] Hollands JG, Spivak T, Kramkowski EW. Cognitive Load and Situation Awareness for Soldiers: Effects of Message Presentation Rate and Sensory Modality. Human Factors. 2019 Aug;61(5):763-73. Publisher: SAGE Publications Inc. Available from: https://doi.org/10.1177/0018720819825803.

[205] Sokol K, Flach P. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. In: Proceedings of the 2020 Conference

on Fairness, Accountability, and Transparency; 2020. p. 56-67. ArXiv:1912.05100 [cs, stat]. Available from: http://arxiv.org/abs/1912.05100.

[206] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al.. PaLM: Scaling Language Modeling with Pathways. arXiv; 2022. ArXiv:2204.02311 [cs]. Available from: http://arxiv.org/abs/2204.02311.

[207] Hu K. ChatGPT sets record for fastest-growing user base - analyst note. Reuters. 2023 Feb. Available from: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[208] Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. npj Digital Medicine. 2023 Aug;6(1):1-8. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41746-023-00896-7.

[209] Liang W, Zhang Y, Cao H, Wang B, Ding DY, Yang X, et al. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. NEJM AI. 2024 Jul;1(8):AIoa2400196. Publisher: Massachusetts Medical Society. Available from: https://ai.nejm.org/doi/full/10.1056/AIoa2400196.

[210] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023 Aug;620(7972):172-80.

[211] Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, et al. Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. JAMA internal medicine. 2023 Sep;183(9):1028-30.

[212] Williams CYK, Zack T, Miao BY, Sushil M, Wang M, Kornblith AE, et al. Use of a Large Language Model to Assess Clinical Acuity of Adults in the Emergency Department. JAMA Network Open. 2024 May;7(5):e248895. Available from: https://doi.org/10.1001/jamanetworkopen.2024.8895.

[213] Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023 Jul;330(1):78-80.

[214] Zakka Cyril, Shad Rohan, Chaurasia Akash, Dalal Alex R , Kim Jennifer L , Moor Michael, et al. Almanac — Retrieval-Augmented Language Models for Clinical Medicine. NEJM AI. 2024 Jan;1(2):AIoa2300068. Publisher: Massachusetts Medical Society. Available from: https://doi.org/10.1056/AIoa2300068.

[215] Oniani D, Hilsman J, Zang C, Wang J, Cai L, Zawala J, et al. Emerging opportunities of using large language models for translation between drug molecules and indications. Scientific Reports. 2024 May;14(1):10738. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41598-024-61124-0.

[216] Unlu Ozan, Shin Jiyeon, Mailly Charlotte J , Oates Michael F , Tucci Michela R , Varugheese Matthew, et al. Retrieval-Augmented Generation–Enabled GPT-4 for Clinical Trial Screening. NEJM AI. 2024 Jun;1(7):AIoa2400181. Publisher: Massachusetts Medical Society. Available from: https://doi.org/10.1056/AIoa2400181.

[217] Soroush A, Glicksberg BS, Zimlichman E, Barash Y, Freeman R, Charney AW, et al. Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying. NEJM AI. 2024 Apr;1(5):AIdbp2300040. Publisher: Massachusetts Medical Society. Available from: https://ai.nejm.org/doi/full/10.1056/AIdbp2300040.

[218] Schwartz IS, Link KE, Daneshjou R, Cortés-Penfield N. Black Box Warning: Large Language Models and the Future of Infectious Diseases Consultation. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America. 2024 Apr;78(4):860-6.

[219] Bills S, Cammarata N, Mossing D, Tillman H, Gao L, Goh G, et al.. Language models can explain neurons in language models; 2023. Publisher: OpenAI. Available from: https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

[220] Han PKJ, Schupack D, Daggett S, Holt CT, Strout TD. Temporal changes in tolerance of uncertainty among medical students: insights from an exploratory study. Medical Education Online. 2015 Sep;20:10.3402/meo.v20.28285. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4565063/.

[221] Holzinger A, Carrington A, Müller H. Measuring the Quality of Explanations: The System Causability Scale (SCS). KI - Künstliche Intelligenz. 2020 Jun;34(2):193-8. Available from: https://doi.org/10.1007/s13218-020-00636-z.

[222] Loria t. TextBlob: Simplified Text Processing — TextBlob 0.18.0.post0 documentation; 2024. Available from: https://textblob.readthedocs.io/en/dev/.

[223] Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock 2021. Critical care medicine. 2021 Nov;49(11):e1063-143. Place: United States.

[224] Pahud de Mortanges A, Luo H, Shu SZ, Kamath A, Suter Y, Shelan M, et al. Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. npj Digital Medicine. 2024 Jul;7(1):1-10. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41746-024-01190-w.

[225] Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. Journal of the American Medical Informatics Association: JAMIA. 2017 Mar;24(2):423-31.

[226] Christiano P, Leike J, Brown TB, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. arXiv; 2023. ArXiv:1706.03741 [cs, stat]. Available from: http://arxiv.org/abs/1706.03741.

[227] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al.. Training language models to follow instructions with human feedback. arXiv; 2022. ArXiv:2203.02155 [cs]. Available from: http://arxiv.org/abs/2203.02155.

[228] Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. JAMA. 2018 Dec;320(21):2199-200.

[229] Abdulnour REE, Parsons AS, Muller D, Drazen J, Rubin EJ, Rencic J. Deliberate Practice at the Virtual Bedside to Improve Clinical Reasoning. New England Journal of Medicine. 2022 May. Publisher: Massachusetts Medical Society. Available from: https://www.nejm.org/doi/full/10.1056/NEJMe2204540.

[230] Wartman SA, Combs CD. Medical Education Must Move From the Information Age to the Age of Artificial Intelligence. Academic Medicine: Journal of the Association of American Medical Colleges. 2018 Aug;93(8):1107-9.

[231] Cabral S, Restrepo D, Kanjee Z, Wilson P, Crowe B, Abdulnour RE, et al. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. JAMA Internal Medicine. 2024 May;184(5):581-3. Available from: https://doi.org/10.1001/jamainternmed.2024.0295.

[232] Furnham A. Response bias, social desirability and dissimulation. Personality and Individual Differences. 1986;7(3):385-400. Place: Netherlands Publisher: Elsevier Science.

[233] Nisbett RE, Wilson TD. Telling more than we can know: Verbal reports on mental processes. Psychological Review. 1977;84(3):231-59. Place: US Publisher: American Psychological Association.

[234] Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. The Journal of Applied Psychology. 2003 Oct;88(5):879-903.

Appendix A

Thesis outputs

Publications

Directly related to PhD thesis

1. Festor P*, **Nagendran M***, Gordon AC, Faisal AA, Komorowski M. Safety of human-AI cooperative decision-making within intensive care: a physical simulation study

   PLOS Digit Health. 2025; 4(2): e0000726.

   *Joint first author

2. **Nagendran M**, Festor P, Komorowski M, Gordon AC, Faisal AA. Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations.

   npj Digit Med. 2024; 7(1): 202.

3. **Nagendran M**, Festor P, Komorowski M, Gordon AC, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision making.

   npj Digit Med. 2023; 6(1): 206.

Indirectly related to PhD thesis

1. Catling FJR, **Nagendran M**, Festor P, Bien Z, Harris S, Faisal AA, Gordon AC, Komorowski M. Can Machine Learning Personalize Cardiovascular Therapy in Sepsis?

   Crit Care Explor. 2024; 6(5): e1087.

2. Vasey B, **Nagendran M**, Campbell B, Clifton DA, Collins GS, Denaxas S, Denniston AK, Faes L, Geerts B, Ibrahim M, Liu X, Mateen BA, Mathur P, McCradden MD, Morgan L, Ordish J, Rogers C, Saria S, Ting DS, Watkinson P, Weber W, Wheatstone P, McCulloch P, DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI.

   Nat Med. 2022; 28(5): 924-933.

   BMJ. 2022; 377: e070904. (joint publication)

Conference papers and presentations

1. **Nagendran M**, Festor P, Komorowski M, Gordon AC, Faisal AA. Interaction of doctors with explainable RL decision support via behavioural readouts of eye-tracking.

   Conference poster - 16th European Workshop on Reinforcement Learning (EWRL 2023), Brussels, Belgium

2. **Nagendran M**, Festor P, Komorowski M, Gordon AC, Faisal AA. Eye-tracking of clinician behaviour with explainable AI decision support: a high-fidelity simulation study.

   Conference poster (and oral talk delivered by Paul Festor) - 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH) 2023, International Conference on Machine Learning (ICML), Hawaii, USA

3. **Nagendran M**, Festor P, Komorowski M, Gordon AC, Faisal AA. Eye-tracking of doctors exposed to XAI decision support: a high-fidelity simulation study.
   Conference poster - UKRI Joint AI CDT Conference 2023 - York, UK.

4. **Nagendran M**, Festor P, Komorowski M, Gordon AC, Faisal AA. Eye-tracking of doctors exposed to XAI decision support: a high-fidelity simulation study.
   Conference poster - University of Bayreuth Digital Health Conference 2023 - Bayreuth, Germany.

5. **Nagendran M**, Gordon AC, Faisal AA. Impact of AI dose suggestions on the prescriptions of ICU doctors.
   Conference video - Machine Learning for Healthcare 2022 - Durham, USA.

6. **Nagendran M**, Gordon AC, Faisal AA. Impact of XAI dose suggestions on the prescriptions of ICU doctors.
   Conference poster - UKRI Joint AI CDT Conference 2022 - Milton Keynes, UK.

7. **Nagendran M**, Gordon AC, Faisal AA. Impact of XAI dose suggestions on the prescriptions of ICU doctors.
   Conference poster - Cognitive Computational Neuroscience, 2022 - San Francisco, USA.

Reproducibility

- The code that supports the findings in this thesis will be made available on Github after the completion of a 12-month embargo period.
- The code and datasets that support the findings of both published studies can be viewed at the links below:
  - Quantifying the impact of AI recommendations with explanations on prescription decision making
  - Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations

Appendix B

Chapter 1: Background

Patient and public involvement (PPI)

PPI application form

Included overleaf.

PPI briefing form

PRE-SESSION INFORMATION SHEET FOR FOCUS GROUP PARTICIPANTS (22nd FEB) on *"Early diagnosis and decision support tools in sepsis using artificial intelligence (AI)"*

We look forward to welcoming you to the focus group session and are grateful for your participation. To ensure we make the most of the session we have prepared some brief background reading material that we hope will be of interest.

Sepsis is a life-threatening condition where an infection triggers an excessive and potentially harmful reaction by the body which can be lethal. Septic patients are different from one another in many ways: their organ problems vary and their immune systems behave differently. Given these differences it may be that a more personalised and tailored approach to treatment improves outcomes. We want to help doctors decide on the right treatment for each patient by applying artificial intelligence (AI) techniques to data from patient monitors and blood tests.

What is so great about AI? Because of the volume of data collected these days an AI can be trained on potentially hundreds of thousands of patient episodes far more than any human could see in their entire clinical lifetime. AI can also be better at spotting patterns among data than humans and suffering less from e.g. tiredness or fatigue during a night shift. There are downsides too though e.g. inability to talk things through with you and AI still needs testing to make sure the suggestions are actually safe before it's implemented in the real world. The best solution in the future is likely to be clinicians and AI working together to cover each other's blind spots and offer the best care possible.

To make sure that this work serves patients in the best way possible we want to hear from you and are particularly interested in the following themes:

a) **How would you feel about an AI algorithm suggesting diagnoses or treatments to the clinical team looking after you or a family member/friend?**

   We foresee several potential issues here. One is how large a role the AI has. Is it merely offering a 'humble' suggestion is it saying that a certain treatment is the best in this case or should it direct the clinical team to do a certain action? Would you expect to be told when a clinician has used an AI as part of the diagnostic/treatment process or only if it deviates from the current status quo/best practice guidelines?

b) **To what extent does the performance of the AI matter?** Most of us would rather hear the opinion of a senior consultant than a first year doctor because they consultant has more experience and is more likely to be correct. If clinical studies proved that the AI made suggestions that were superior (on average) to even senior doctors then would that make you feel more comfortable about them having a bigger role? Should the clinical team always have the final say before discussing with you?

   The AI will use data from blood tests monitors and the clinical record to make its suggestions and recommendations. It is not (yet!) capable of having a chat with you. So we also wonder what the best way to incorporate your views into the decision making process might be? At the moment for example the clinical team will discuss diagnoses and treatments with you to determine what your preferences and feelings are and how they best fit into a management plan for you. How would you feel about submitting a brief questionnaire to your clinical record on e.g. your thoughts about risks/benefits how much you value quantity

over quality of life etc. Or would you rather the clinician communicates the AI suggestions and discusses this with you?

c) **If we could design the AI to explain its decisions what sort of explanation would be helpful to you?**

One of us is researching the best way for the AI to explain its decisions to clinicians. For example if the question for the AI is what volume of fluid to give or what dose of drug to give there are several possible ways for the AI to display this:

- "Give 200ml fluid over the next hour"
- "These are the top 3 actions for the next hour..."
- "These are the top 3 actions and the percentage chance they are the best thing to do..."
- "Here is a graph of what I ('the AI') predict will happen if you perform the best action and if you do not"
- "I think X is the best action and these are the observations/blood results I've seen that make me think this..."

How important is it for you that an AI is able to explain why it is recommending something. Ultimately if you ask a clinician for an explanation and they have used an AI then they will need to convey some of this information to you.

Does it matter how the AI decided something as long as we can tell from studies that it seems to be the right action? Some very complex AI algorithms have internal workings that rely on complicated maths and we can't often tell exactly how they're weighing things up. But we can see the outputs and whether it's working. In other words should the AI be explaining its decision making process or persuading the clinical team (and by extension you) about how good/effective its decision is regardless of how it came to make it.

The AI can also be trained to value certain goals more highly than others and the aim would be to prioritise patient centred outcomes (e.g. survival time spent on a ventilator time in the ICU/in hospital) as opposed to short term number based outcomes (what your blood pressure is on a given day how much urine has been produced whether there are fevers). We know that sometimes aiming for the

short term numbers to look good can actually make the longer term outcomes worse. Would an explanation that says it is trying to maximise a patient-centred outcome be helpful to know?

d) **How would you feel if the AI made a recommendation that turned out to be a mistake?**

Despite the pace of discovery there is still a lot of uncertainty in healthcare. Human clinicians despite our best intentions still can and do make mistakes. When this happens there is now a much stronger culture of honesty and a duty of candour.

How would you feel if there was an incorrect recommendation or suggestion from an AI? Would this depend on whether it was just a suggestion (in which case the clinician acting on it should take most of the responsibility) or more of a command (in which case the AI and by extension the makers of the AI are responsible)?

The example of self-driving cars is interesting to think about in this regard. We know most road accidents are the result of human error. Even if self-driving cars make mistakes they are probably much less likely to cause harm than humans. However many of us still find the idea of a self-driving car causing an accident very disturbing and any such reports get a lot of media and press attention.

Do you think these attitudes might be shaped more generally about overall views on AI in society and how big a role humans should keep in some fields e.g. healthcare vs. manufacturing automation?

e) **Data available from patients' blood samples such as genetic information may be able to help doctors decide if a patient is likely to respond to a treatment. Are you comfortable with your data being used in this way? Would use of genetic information cause any concern?**

PPI transcripts

PPI SESSION HELD REMOTELY VIA ZOOM ON 22 FEB 2021

Comments from three participants (all former ICU patients)

Topic 1: AI decision support tools in general

- **How would you feel about an AI algorithm suggesting diagnoses or treatments to the clinical team looking after you or a family member/friend?**

  Participant 1 feels this is a good idea and can help the doctors targeting/diagnosing. Could also be quicker than humans can do it. Doesn't scare her.

  Participant 2 asked if all data from all patients could be added into the system to help with diagnosis especially around genetic testing. How would this work with consent especially if this from NOK.

  Participant 3 agreed with points above but felt it was important how it was put across to patients. Felt that if it was put across in a positive way.

  Participant 1 felt although she is not a 'techie' person she felt we have to move with the times and if it is used all the time it will become just part of normal care. This would be a development of normal practice and the machines we use already.

  Participant 3 have to trust the experts who know what the right drugs are etc and that this is an extension of that.

  Participant 2 felt that the AI needs lots of testing to be sure that it works before it is fully implemented as it is dealing with people's lives.

  Participant 1 felt she would trust the decisions that the doctors were making based on the AI. She believed it would be tested rigorously before it was implemented.

  Participant 3 would like there to be back up systems so that we weren't reliant on one system/computer.

  Finn – There are two ways these could work a description about how a part of the body could work (mathematical model) or a system that uses the huge amount of data and finds a pattern without reference to what might be going on underneath. Which is more acceptable?

  Participant 2 both equally acceptable but depends on how much it would alter treatment and would need monitoring all the way through would be concerned when the genetic markers are introduced. She thinks this is concerning because she feels many people's relatives would be uncomfortable with this as it may find other things that are unrelated to sepsis and how would these things be

communicated to patients. More used to the data being analysed and the results being given to the doctor who then decides what to do with it.

Participant 1 thinks the AI would allow greater shared knowledge across hospitals especially about rare situations. She is really positive about this.

Participant 3 asked if people could opt out of getting results from genetic information if it was being used.

Participant 2 felt this could be an issue around not being conscious at the time of the results.

Finn – asked about risks and benefits of treatment asked how we might collect ideas about preferences for risks and benefits before going into hospital.

Participant 1 – very different going for scheduled care compared to unplanned sepsis.

Participant 3 – felt and opt out system might work

Participant 2 – Would hope that husband and daughter would know them well enough to make a decision on their behalf.

Myura – Would people feel comfortable incorporating patient wishes into the decision suggestions.

Participant 2 – clinicians need to be the link between the AI and patients

Participant 1 – may be different in 30-40 years' time when AI is more normal part of life for now would need a person explaining it. Everyone likes some human touch. Interaction with patients is important (including with staff) is important to keep patients going this couldn't be done with AI.

Participant 3 – could the system be introduced on general wards before ICU so that patients could get more comfortable with it.

Finn – Asked if a system that suggested guidelines and evidence needed as much scrutiny as a black box system.

Participant 1 – don't want to move away from knowing the patient and becoming reliant on the AI. Clinicians need to be involved to know if it is the right think for the patient. Likened AI to the machines that we are currently used.

Participant 3 – felt that if the AI could display the long-term consequences of the decisions would be helpful

Participant 2 – would depend on the stage that the research is to the extent that the AI can prompt or take over. Needs the baby steps at the beginning. Would the AI be able to keep up in a timely way?

Topic 2: Explainability

- **If we could design the AI to explain its decisions what sort of explanation would be helpful to you?**

  Myura - It is one thing for the AI to give good recommendations but another thing is whether and how clinicians translate these into action. This is influenced by lots of factors. If an AI recommended a treatment how important is it to explain (to patients and relatives) why the treatment needs to be given as long as the answer is right?

  Participant 3 - It would have to be an explanation in very simple terms given how stressful situation is for patient and family e.g. this is the treatment we want to give and here is the and this is what the treatment does

  Participant 2 - I feel that it does have to be explained.

  Participant 1 - As a patient I have lots of trust in the team that are treating me. If they were happy that the AI was recommending the right thing I would trust this without needing an explanation.

  Myura - What might be the best way for an AI to present recommendations to the clinical team?

  Participant 1 - I don't think it's too relevant to me too much as long as it helps the clinician helps make the best decision

  Participant 2 - I feel it's important that the AI presents the clinicians with multiple options

  Myura - Sometimes treatments which 'make the numbers look right' in the short-term don't correlate with improved long term outcomes. In contrast AI might be able to infer links between actions taken now and longer term outcomes and recommend treatments that it thinks will improve these long-term outcomes. In this situation the clinicians looking after the patient might not be able to understand the rationale for the AI's suggested actions and so wouldn't be able to offer an explanation to the patient / family. How do you feel about the fact that we couldn't offer a full explanation here?

Participant 1 - I would be OK with there not being an explanation long as this long term association was proven.

Participant 2 - I'm worried that it might be difficult for the AI to tell the difference between 'chicken' and 'egg'. For example if someone dies then we aren't collecting their data any more and so the AI won't see it.

Participant 3 - In the future it would be really helpful if the AI was able to take psychological data into account as when the staff took this into account it was very therapeutic for me.

Topic 3: AI mistakes or errors

- **How would you feel if the AI made a recommendation that turned out to be a mistake?**

  Finn – How much information would patients want about the level of uncertainty of a decision.

  Participant 2 – there may be problems linking GP to hospital data.

  Participant 2 – many of the treatments are down to personal preference of the medical staff and this usually isn't discussed with patients. The case conference discussions often aren't communicated to patients as it is so what would the difference be. Knowing the uncertainty would make things more complicated. This could feed into the suing culture

  Participant 3 – agrees with Participant 2 need to assume that the clinicians are making the best decisions and rely on what the clinicians are saying.

  Participant 1 – have to trust clinicians that the best decisions are being made even if these turn out not to be the right ones.

  Finn – asked about what they would feel if an AI made a recommendation that turned out to be a mistake.

  Participant 1 – clinicians can also make mistakes so it would still be hard to put blame somewhere. If it is the AI is it any worse than if it is a clinician. Maybe it is best not to know if it is AI.

  Participant 2 – as AI is introduced will this decrease the teaching of lower grade staff so that they won't be as experienced when they become consultants.

Participant 1 – want to make sure that if a mistake is made that there is learning from it.

Participant 3 – agrees that if there is a mistake it needs to be fed back into the system to allow learning.

Participant 2 – the transparency is the aspect where there needs to be the bridge between the medical staff and the relatives. Somethings patients don't need to know if the computer decided it and some patient would want to know even if it was a clinician who made the decision.

Appendix C

Chapter 2: Vignette experiment

Pre- and post-experiment questionnaires

- Pre-experiment:
  - How old are you?
  - For how many years have you been working clinically?
  - What is your current position (or equivalent level)? Options: 'SHO', 'SpR', 'Consultant'
  - What type of ICU do you predominantly work in? Options: 'General', 'Cardiac', 'Neuro', 'Paeds', 'Liver', 'Trauma', 'Mixed'
  - 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
- Post-experiment:
  - How likely would you be to use an AI system like this to help guide your clinical practice in the ICU? Keep in mind that the current interface is a demo and will be replaced by a professionally designed version similar to current electronic health software. 5-point Likert scale ('Very Unlikely', 'Unlikely', 'Neutral', 'Likely', 'Very likely').
  - How important to you using an AI system like this is evidence of improved clinical outcomes (e.g. reduced length of stay or lower mortality) in an observational study? 5-point Likert scale ('Very unimportant', 'Unimportant', 'Neutral', 'Important', 'Very important').

- How important to you using an AI system like this is evidence of improved clinical outcomes (e.g. reduced length of stay or lower mortality) in a randomised trial? 5-point Likert scale ('Very unimportant', 'Unimportant', 'Neutral', 'Important', 'Very important').

- How useful overall were the explanations for the AI suggestions? 5-point Likert scale ('Not at all useful', 'Not useful', 'Neutral', 'Useful', 'Very useful').

Experiment screenshots

Screenshots showing progress through various screens of the interface follow in the pages below.

Selected comments from subjects

- I feel that the AI does help narrow my range of options to make a more evidence-based decision.
- At times I found I agreed with the AI system. At time I found I wildly diverged from it. I think mostly my responses were more extreme ('stop fluid', 'double norad') than the AI's suggestions. I really enjoyed this exercise.
- Different way of approaching decision making than I might normally do, obviously with less inputs than normally available. Prompts interesting thinking about how we make clinical decisions "ahead of time" for patients i.e. in the next hour, rather than being reactive. It would be really interesting to estimate how much effect the tool has on moderating my normal decision making.
- Very interesting. I thought the AI explanations would make me trust it, but it's difficult if you don't understand the algorithm / biological mechanism - I guess I expected it to use cardiac output parameters / SIRS criteria etc.
- Great tool - innovative and barring a few anomalies, feels pretty reflective of what I would do.
- The AI suggestions with explanations were not very well explained. 'Words' were given without context.
- Really interesting concept, I'd be very interested to understand the reasoning that the AI tool uses to get to its suggestions, as sometimes they were very different to what I would have done, so it would be useful to know when I was missing something.
- Always difficult as clinical trajectory is associated with uncertainty. very generous use of noradrenaline overall it seemed and this is not without consequence. Potential to help nurse/junior clinician if validation can be confirmed.
- Very interesting to see how Ai is thinking. Made me realise how my human interaction is very much based on a select few parameters, and different layouts are more likely to lead me to potentially miss, or misunderstand some observations. In my own practice, seeing patients and then setting physiological parameters is common. The ability of AI to adjust automatically to meet this would be a huge help in reducing chances of inappropriate drops/ increases of human physiology.
- I found I often disagreed with the AI clinician and agreed with the colleague input
- It showed me my own proclivities to using fluids rather than NA. I'm more comfortable with a 'wet' approach than a dry one. I think this is a reflection of what I have observed in colleagues and seniors over recent years. Similarly,

my likelihood to use, or ignore, AI suggestions will - I imagine - be determined more than anything else by senior (sensible in my view) colleagues and their acceptance or derision of the technology. Also, the possible criticism clinicians may feel they open themselves up to if they fail to follow the algorithm.

- Excellent premise. I think this has loads of potential. I guess difficulty is knowing what the 'gold standards' are with regards to fluid/vasopressor therapy. The model has been exposed to more patient 'lab data' than any clinician, but how much does clinical examination and the resulting 'data' play a role in these decisions?
- More explanation/data around the evidence for the categories chosen as most important by the AI facility would give me more faith in following those instructions.

Appendix D

Chapter 3: Eye-tracking data

Gaze positions on surface

Field	Description
world_timestamp	world camera timestamp at which the sample was recorded
world_index	frame number of the world camera to which this sample is correlated
gaze_timestamp	
x_norm	x coordinates between 0 and 1, where (0,0) is the bottom left corner of the surface and (1,1) is the top right corner
y_norm	y coordinates between 0 and 1, where (0,0) is the bottom left corner of the surface and (1,1) is the top right corner
x_scaled	same x coordinate as above but coordinates scaled with the size defined for the surface
y_scaled	same y coordinate as above but coordinates scaled with the size defined for the surface
on_surf	boolean for whether gaze is on surface or not

Table D.1 Description of fields in eye tracking gaze position data (here using the example region of interest of AI screen)

Gaze positions

Field	Description
gaze_timestamp	timestamp of the source image frame
world_index	closest world video frame
confidence	computed confidence of gaze (0 to 1 scale, higher = more confidence)
norm_pos_x	x position in the world image frame in normalized coordinates
norm_pos_y	y position in the world image frame in normalized coordinates
base_data	"timestamp-id timestamp-id ..." of pupil data that this gaze position is computed from data made available by the 3D vector gaze mappers

Table D.2 Description of fields in eye tracking gaze position data

Fixations

Field	Description
world_timestamp	world camera timestamp at which the sample was recorded
world_index	frame number of the world camera to which this sample is correlated
fixation_id	unique id for fixation
start_timestamp	timestamp of the first related gaze datum
duration	exact fixation duration, in milliseconds
dispersion	dispersion, in degrees
norm_pos_x	normalized x position of the fixation's centroid
norm_pos_y	normalized y position of the fixation's centroid
x_scaled	same x coordinate as above but coordinates scaled with the size defined for the surface
y_scaled	same y coordinate as above but coordinates scaled with the size defined for the surface
on_surf	boolean for whether fixation is on surface or not

Table D.3 Description of fields in eye tracking fixation data (here using the example region of interest of AI screen)

Surface positions

Field	Description
world_index	frame number of the world camera to which this sample is correlated
world_timestamp	world camera timestamp at which the sample was recorded
img_to_surf_trans	3x3 transformation matrix from camera coordinate space to reference surface coordinate space
surf_to_img_trans	3x3 transformation matrix from reference surface coordinate space to camera coordinate space
num_detected_markers	number of April tags detected for this sample
dist_img_to_surf_trans	3x3 transformation matrix from distorted camera image coordinate space to reference surface coordinate space
surf_to_dist_img_trans	3x3 transformation matrix from reference surface coordinate space to distorted camera image coordinate space
num_definition_markers	[unknown - not used]

Table D.4 Description of fields in eye tracking data (surface positions AI)

Blinks

Field	Description
id	numerical ID of blink
start_timestamp	blink start timestamp (s)
duration	blink duration (s)
end_timestamp	blink end timestamp (s)
start_frame_index	blink start world frame index
index	blink median world frame index
end_frame_index	blink end world frame index
confidence	mean of absolute filter response during blink clamped at 1
filter_response	filter response ('Activity') during blink
base_data	timestamps of data associated with blink

Table D.5 Description of fields in eye tracking blink data

Appendix E

Chapter 4: Simulation suite experiment

Convergence Metric and Jaccard Similarity

Jaccard Similarity:

For two sets A and B, the Jaccard similarity is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

In our context, for each state-action pair (s, a) between two consecutive runs i and $i + 1$, we calculate:

$$J_{i,i+1}(s, a) = \frac{|E_i(s, a) \cap E_{i+1}(s, a)|}{|E_i(s, a) \cup E_{i+1}(s, a)|}$$

Where $E_i(s, a)$ is the set of episodes for state-action pair (s, a) in run i. The overall Jaccard similarity between runs i and $i + 1$ is then the average over all state-action pairs:

$$J_{i,i+1} = \frac{1}{|S \times A|} \sum_{(s,a) \in S \times A} J_{i,i+1}(s, a)$$

Convergence Metric: For a given run n, the convergence metric is:

$$C_n = \frac{1}{|S \times A|} \sum_{(s,a) \in S \times A} \frac{\max_{e \in E_n(s,a)} f_n(e, s, a)}{\sum_{e \in E_n(s,a)} f_n(e, s, a)}$$

Where:

- $S \times A$ is the set of all state-action pairs
- $E_n(s, a)$ is the set of all episodes seen for state-action pair (s, a) up to run n
- $f_n(e, s, a)$ is the frequency of episode e for state-action pair (s, a) up to run n

Trial matrix

Physician	Scenario						Included in eye-tracking analysis?
	1	2	3	4	5	6	
1	S	S	S	S	U	I	No
2	S	S	S	U	S	I	No
3	S	S	U	S	S	I	Yes
4	S	U	S	S	S	I	Yes
5	S	S	S	S	I	U	No
6	S	S	S	I	U	S	No
7	S	S	U	S	I	S	No
8	S	U	S	S	I	S	No
9	S	S	S	U	S	I	Yes
10	S	S	S	I	U	S	Yes
11	S	S	U	I	S	S	Yes
12	S	U	S	I	S	S	Yes
13	S	S	I	S	S	U	Yes
14	S	S	I	S	U	S	Yes
15	S	S	I	U	S	S	No
16	S	U	I	S	S	S	Yes
17	S	I	S	S	S	U	Yes
18	S	I	S	S	U	S	No
19	S	I	S	U	S	S	Yes
20	S	I	U	S	S	S	Yes
21	S	S	S	S	I	U	Yes
22	S	S	S	U	S	I	Yes
23	S	S	U	S	S	I	Yes
24	S	U	S	S	S	I	Yes
25	S	S	S	S	I	U	No
26	S	S	S	U	I	S	No
27	S	S	U	S	I	S	No
28	S	U	S	S	I	S	No
29	S	S	S	I	S	U	No
30	S	S	S	I	U	S	No
31	S	S	U	I	S	S	Yes
32	S	U	S	I	S	S	No
33	S	S	I	S	S	U	No
34	S	S	I	S	U	S	Yes
35	S	S	I	U	S	S	No
36	S	U	I	S	S	S	Yes
37	S	I	S	S	S	U	No
38	S	I	S	S	U	S	No

Fig. E.1 Safe scenarios are denoted by a green 'S', unsafe scenarios by a red 'U' and incentivised unsafe scenarios by a yellow 'I'. The last column indicates whether or not eye-tracking data was ultimately available for analysis.

Standardised experiment briefing

This simulation experiment aims to study how clinicians might interact with an AI decision support tool for patients with sepsis. The AI takes patient variables as input and outputs a treatment recommendation for both fluid and vasopressor. We have limited evidence that the AI might be effective based on retrospective data from an American dataset and further validation on retrospective data from the Netherlands. However, there has not yet been prospective evidence of either effectiveness or safety.

You will conduct a brief ward round review of 6 ICU patients with sepsis. Between each patient, you will exit the room and be called in again to see the next patient by the nurse.

For each patient, you will first be asked:

- Treatment prescriptions for: (i) fluid in ml and (ii) vasopressor in mcg/kg/min, both for the next hour
- Your confidence in the prescription on a scale from 1 (low) to 10 (high)
- Whether or not you want to get advice from another doctor / senior doctor

You will then be shown the AI treatment recommendation on a digital screen. The screen contains the AI dose suggestions in the middle (for fluid and vasopressor) along with 4 explanations for the suggestions (one in each corner).

You will then be asked:

- To what extent you agree with the AI suggestion on a scale from 1 (strongly disagree) to 5 (strongly agree)
- Whether you wish to adjust your treatment prescription for: (i) fluid in ml and (ii) vasopressor in mcg/kg/min, both for the next hour
- Whether your confidence in your prescription (on a scale from 1 [low] to 10 [high]) has changed as a result of seeing the AI suggestion
- Whether or not you want to get advice from another doctor / senior doctor
- If the AI suggested treatment was to be administered to the patient, would you act to stop the administration?

We will now show you an example of the AI screen that you will encounter in the experiment. You can see that there is an AI suggestion in the centre of the screen which is how much fluid and vasopressor the AI recommends over the next hour.

Around the corners of the screen there are four different types of explanation which can be thought of as the AI trying to convey the rationale for its suggested doses:

- The mortality change explanation conveys information on what the AI predicts will happen to the overall mortality risk in the short-term based on potential dose increases or decreases.
- The most influential training examples explanation conveys which three training cases were most helpful for learning the current suggestion in the same way that we might base our own treatment choices on previous notable cases we learnt from.
- The feature importance explanation conveys which were the top five features (or items in the data) that were most useful to the AI in generating its current suggestion.
- The 'AI treatment options gap' explanation conveys how much one treatment strategy looks superior compared to alternatives. If all potential options are similar (i.e. a low gap) then it suggests that the AI has near equipoise for options and you may wish to use your own judgement more strongly (as you will have additional information from examining the patient for example). However, if the gap is high it suggests that the AI has identified one particular treatment strategy as superior to the alternatives and therefore it would be worth considering this recommendation more strongly than with a low gap recommendation.

Pre- and post-experiment questionnaires

Pre-experiment:

- 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 physicians
  - If physicians put too much trust in AI, they won't be able to detect when the AI fails, and it will lead to patient harm

Post-experiment:

- How useful overall were the explanations? (on a 5-point Likert scale ('Not at all useful', 'Not useful', 'Neutral', 'Useful', 'Very useful')
- Please rate the explanation types in usefulness:
  - Repeated separately for each of the four explanations
    - Mortality change
    - Most influential training examples
    - AI treatment options gap
    - Top 5 features

Patient scenarios and rationale

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 tachy. Had stat co-amoxiclav and clarithromycin
- 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
- Lac 3.7 in ED, UO 25ml over last 4 hrs

Mannequin 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

Justification: Middle-aged man in septic shock secondary to community acquired pneumonia. Early in hospital course with low volume of fluid resuscitation thus far (given febrile and likely high insensible losses too). Oliguric and tachypneoic. Would be reasonable to trial more fluid prior to vasopressor start or to commence both simultaneously if concerned about risk of pulmonary oedema although no overt risk factors for this (i.e. no background history of poor cardiac function). Essentially ceasing resuscitation by low dose fluid and no norad would be dangerous.

Patient 2 handover note for participants:

- 84F admitted last night from ED with dysuria, presumed urosepsis. COVID -ve.
- PMHx: 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

Mannequin 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

Justification: Elderly lady with septic shock secondary to gram negative bacteraemia from UTI. Normally hypertensive and oliguric. Yet to respond to reasonable volume of fluid resuscitation. Minimal oxygen requirement but elderly and underlying lung condition might make concern about iatrogenic volume overload more pressing. Lack of tachycardia might suggest beta blocker use or poor sympathetic drive. Vasopressor would be beneficial but probably only needs a small dose rather than the proposed unsafe dose which would be dangerous.

Patient 3 handover note for participants:

- 42F admitted 8d ago from ED with SOB. COVID +ve pneumonia.
- PMHx: 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 tazocin course on admission, currently off antimicrobials
- Fluid balance -250ml last 48 hrs
- Latest lac 2.3, UO 60-70 ml/hr last 4 hrs

Mannequin 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

Justification: Middle aged lady with sepsis secondary to likely ICU acquired infection (could be line related or ventilator-associated). Has been in ICU for over a week so likely to be fluid replete. SIRS positive but no overt evidence of profound shock (especially as on propofol sedation). Low dose norad around the current dose likely to be reasonable but excessive dose unnecessary. Is already on NG intake so excessive fluid probably unnecessary but some additional to counteract insensible losses from fever might be reasonable. High dose norad unnecessary and likely dangerous.

Patient 4 handover note for participants:

- 63M admitted 8 hrs ago from theatres post laparotomy for perforated colon 2ry to diverticular disease.
- PMHx: 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

Mannequin 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

Justification: Middle aged man with septic shock secondary to abdominal sepsis after perforated viscus. Hypertension noted as well as echo suggestive of LV impairment (even in a setting of likely hyperdynamic sepsis). Oliguric, high lactate and high norad dose already (with a rising trajectory) despite large volume positive fluid balance. Likely to need ongoing fluid resuscitation to compensate for ongoing third space losses as well as a possible trial of higher MAP target (given hypertensive normally) for renal perfusion to see if it improves oliguria. Complete cessation of vasopressor would be dangerous.

Patient 5 handover note for participants:

- 33F admitted last night from ED with SOB. COVID -ve.
- PMHx: 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

Mannequin 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

Justification: Young lady with mixed septic and cardiogenic shock secondary to endocarditis. Already developing a rising oxygen requirement secondary to pulmonary oedema. Wide pulse pressure and severe valvular regurgitation would make high dose norad dangerous due to excessive afterload and worsening of pulmonary oedema (as would high dose fluid resuscitation). Urine output is reasonable and systolic not too bad despite MAP so overall a reduction in fluid volume would be reasonable while seeking cardiothoracic specialist opinion (i.e. definitive management).

Patient 6 handover note for participants:

- 29M admitted 8 hrs ago from ED for perineal cellulitis +/- nec fasc.
- CT scanner delay, aiming scan imminently, surgeons finishing prev emergency case
- PMHx: 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

Mannequin 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

Justification: Young man with septic shock secondary to necrotising fasciitis. Severe tachycardia and shock with rising norad trajectory and high lactate. Urine output is good though. Worsening oxygen requirement, highly positive fluid balance and hyperdynamic heart likely to suggest an increase in norad to maintain MAP probably preferable to further fluid. Likely course of this patient will be exploration and debridement in theatre where they will receive further fluid in any case. Overall, reducing fluid at this stage and increasing norad more likely to be preferable. Sudden drop in norad to 0.02 likely to be dangerous.

Appendix F

Chapter 5: Large language model experiment

Standardised experiment briefing

This simulation experiment aims at studying how clinicians interact with an AI decision support tool for patients with sepsis. The AI takes patient variables as input and outputs a treatment recommendation for both fluid and vasopressor. We have some evidence of effectiveness of the AI you will interact with on retrospective data from an American dataset and further validation on retrospective data from the Netherlands. However, there has not yet been prospective evidence of either effectiveness or safety.

You will conduct a brief ward round review of 6 ICU patients with sepsis. Between each patient, you will exit the room and be called in again to see the next patient by the nurse. For each patient, you will first be asked for treatment prescriptions: (i) fluid in ml and (ii) vasopressor in mcg/kg/min, both for the next hour.

You will then be shown the AI treatment recommendation on a laptop. In half of the patient scenarios, you will also receive an explanation for the rationale behind the AI recommendation. This is presented in a chat window where you are able to ask, should you wish to, any follow-up questions regarding the explanation. There is no time limit or conversation limit.

You will then be asked to enter your final prescriptions for fluid and vasopressor. You do not have to change your decision.

Pre- and post-experiment questionnaires

Pre-experiment:

- How old are you?
- For how many years have you been working clinically as a doctor?
- For how many years have you been working in ICU?
- What is your current position (or equivalent level)? Options: 'SHO', 'SpR', 'Consultant'
- Risk and attitude to uncertainty on a 5-point Likert scale ('Strongly disagree', 'Disagree', 'Neutral', 'Agree', 'Strongly agree')
  - I enjoy taking risks
  - I try to avoid situations that have uncertain outcomes
  - Taking risks does not bother me if the gains are high
  - I consider security an important element in every aspect of life
  - People have told me I seem to enjoy taking chances
  - I rarely, if ever, take risks when there are other alternatives
  - I usually feel anxious when I am not sure of a diagnosis
  - I am quite uncomfortable with the uncertainty in patient care
- 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

Post-experiment:

- How useful overall were the explanations? (on a 5-point Likert scale ('Not at all useful', 'Not useful', 'Neutral', 'Useful', 'Very useful')
- Your views on the explanations for the AI suggestions on a 5-point Likert scale ('Strongly disagree', 'Disagree', 'Neutral', 'Agree', 'Strongly agree')
  - I found that the data included all relevant known causal factors with sufficient precision and granularity
  - I understood the explanations within the context of my work
  - I could change the level of detail on demand

- I did not need support to understand the explanations
- I found the explanations helped me to understand causality
- I was able to use the explanations with my knowledge base
- I did not find inconsistencies between explanations
- I think that most people would learn to understand the explanations very quickly
- I did not need more references in the explanations: e.g., medical guidelines, regulations
- I received the explanations in a timely and efficient manner

Initial LLM prompt attempts

We are developing interpretability for a healthcare reinforcement learning (RL) algorithm coded in Python using Q-learning. The algorithm consists of 750 non-terminal states and 25 possible actions. The reward is set as +100 for survival at 90 days and -100 for death at 90 days.

Each action corresponds to a combination of 1 of 5 possible fluid doses and 1 of 5 possible vasopressor doses (hence, 25 possible actions).

The 750 states were formed by applying a k-means clustering algorithm to a dataset consisting of 17,083 patient episodes. Each episode consists of around 5-20 timesteps (of 4 hours each). Each timestep contains data for 42 clinical features.

Data from the RL Process
You will have the following data from the RL learning process:

- **Patient Features:** The feature values for the patient we are interested in (called ‘patient_feats‘ in the explanation CSV).
- **State:** The state that the patient is in (called ‘state‘ in the explanation CSV).
- **Centroid Features:** The feature values for the cluster/state centroid (called ‘centroid_feats‘ in the explanation CSV).
- **Optimal Action:** The optimal action, i.e., max Q value in the Q table for this state (called ‘optimal_action‘ in the explanation CSV).
- **Suggested Fluid Dose:** The fluid dose corresponding to the optimal action (called ‘suggested_fluid‘ in the explanation CSV).
- **Suggested Vasopressor Dose:** The vasopressor dose corresponding to the optimal action (called ‘suggested_vaso‘ in the explanation CSV).
- **Optimal Action Value:** The value of the optimal action (called ‘optimal_action_value‘ in the explanation CSV).
- **Training Cases:** Episode numbers of the 1 to 3 training cases most influential during Q-learning (called ‘training_cases‘ in the explanation CSV).
- **Training Case Features:** The feature values for the 1 to 3 most influential training cases (called ‘training_case_feats‘ in the explanation CSV).

- **Top 5 Next States:** The 5 most likely state transitions (called ‘top_5_next_states‘ in the explanation CSV).
- **Top 5 Alternate Actions:** Derived from the Q table for the 5 most likely states (called ‘top_5_alt_actions‘ in the explanation CSV).

Your explanation must be concise (maximum 2-4 sentences) and reflect that the intended audience is expert clinicians.

Here is the Python code that was run on the original data (dt_history.csv):

```
# Python code for AI Clinician 1.0
# ... omitted for brevity in this appendix ...
```

Action Definitions

- **Fluid:** Dose of intravenous fluid from 1 to 5 (which represents the following doses: 1: 0 ml/hr, 2: 8 ml/hr, 3: 21 ml/hr, 4: 80 ml/hr, 5: 237 ml/hr).
- **Vaso:** Dose of vasopressor (noradrenaline), a type of medication that constricts (narrows) blood vessels to increase blood pressure. 1 to 5 represents the following doses: 1: 0 mcg/kg/hr, 2: 0.04 mcg/kg/hr, 3: 0.13 mcg/kg/hr, 4: 0.27 mcg/kg/hr, 5: 0.68 mcg/kg/hr).
- **State:** The cluster from the k-means clustering algorithm that the patient has been assigned to.
- **Action:** A combination of the 1 to 5 fluid dose and 1 to 5 vasopressor dose (therefore 25 potential combinations). Used for allowing the reinforcement learning algorithm to consider both drugs together.

For example, action 25 is fluid 5 and vaso 5, which corresponds to fluid 237 ml/hr and vasopressor 0.68 mcg/kg/hr. All responses must express the actions in terms of fluid and vasopressor dose rather than 1 to 5 or 1 to 25 integers.

Here is the patient data to explain. Use all the previous prompts provided as context to interpret this data. Keep the explanation concise (max 4 sentences) and focus on the RL aspects rather than merely summarising data.

Final LLM prompts

```
prompt = f"""
        You are a persuasive medical expert.
        Succinctly (maximum 3 sentences) repeat the
        recommended doses and explain this dose
        recommendation from an AI
        reinforcement learner agent to a doctor.
        Go straight into your explanation, no filler.
        {explain}
        """

kwargs = {
           "modelId": "anthropic.claude-v2:1",
           "contentType": "application/json",
           "accept": "*/*",
           "body": "{\"prompt\":\"Human: " + prompt +
               100↪ "\\nAssistant:\",\"max_tokens_to_sample\":300,
           \"temperature\":1,\"top_k\":250,\"top_p\":0.999,
           \"stop_sequences\":[\"\\n\\nHuman:\"],
           \"anthropic_version\":\"bedrock-2023-05-31\"}"
        }

#subsequent prompt following human question
conversations[case].append({
           "role": "human", "content": user_input +
           " Do not say you do not have enough info. " +
           "The original prompt provided you with all
           the explanation data you need. " +
           "Use that (e.g. comparison with cluster,
           training cases, certainty level, state
           transitions)."
           })
```

LLM explanation example

Patient: 7684

AI Clinician recommends fluid: 80 ml/hr and vasopressor: 0.13 mcg/kg/min
This sepsis patient cluster has a typical age of 76, weight of 78kg and very high comorbidity burden.
The patients are in the middle phase in their sepsis course and hardly any are mechanically ventilated.
Heart rate is 76 and MAP is 64 on 0.11 mcg/kg/min of norad. Patients are generally okay with urine output with less than 100ml/hr of fluid being infused.
Overall cumulative balance after 1 to 2 days in ICU is around 4 litres and patients are on a moderate FiO2 ($\sim$ 41%).
Lactate is normal but patients are mildly acidotic. There are no other major abnormalities in this state.
The most influential training case survived but was readmitted. Average norad dose was 0.14 mcg/kg/min (with a peak norad dose of 0.56 mcg/kg/min).
Peak fluid balance was 0.1 litres.
The patient was ventilated with median FiO2 of 40%. Max lactate was 2.6 with lowest pH of 7.23. Lowest quartile urine output was 18 ml/hr and max creat was 309.
This AI dose recommendation is in the 50th percentile rank for 'certainty' based on Q value difference compared to alternative actions.
The next most likely state is to remain in the current state (68% chance). Of the remaining top next states, 6 are lower mortality states while 1 are higher.
Risk of entering the death state is 4% while chance of entering survival state is 4%.

```
explanation_data(60971,11)
```


```
{'patient_feats': 'timestep,episode,gender,age,elix,readm,dead,wt,gcs,hr,sbp,map,dbp,rr,spo2,temp,fio2,k,na,cl,glu,urea,creat,mg,ca,alt,bili,
alb,hb,wcc,plt,aptt,pt,inr,ph,po2,pco2,be,lac,hco3,vent,norad,in,out,balance,fluid,vaso,action,state\n11,60971,1,69,5,0,0,105,6,70,119,69,44,
17,100,37.2,0.4,5.0,134,97,12.0,17,680,2.5,2.1,14,39,31,9.8,13.0,93,34.8,17.5,2.1,7.42,22.93,5.1,0.2,2.1,19,1,0.0,10,0,1.3,2,1,6,144\n',
 'state': 144,
 'centroid_feats': 'timestep,gender,age,elix,wt,gcs,hr,sbp,map,dbp,rr,spo2,temp,fio2,k,na,cl,glu,urea,creat,mg,ca,alt,bili,alb,hb,wcc,plt,apt
t,pt,inr,ph,po2,pco2,be,lac,hco3,vent,norad,in,out,balance\n7.0,0.17,65.0,5.0,84.0,8.0,77.0,113.0,71.0,52.0,17.0,98.0,36.5,0.48,4.5,133.0,10
1.0,7.8,24.0,388.0,2.2,1.86,17.0,8.0,24.0,9.33,14.2,209.0,35.4,14.5,1.3,7.33,16.57,4.79,-5.3,1.7,18.0,0.0,0.0,11.0,5.0,3.3\n',
 'optimal_action': 10,
 'suggested_fluid': 'Fluid: 8 ml/hr',
 'suggested_vaso': 'Vasopressor: 0.68 mcg/kg/min',
 'optimal_action_value': 95.20082574431575,
 'training_cases': {27002},
 'training_case_feats': 'timestep,episode,gender,age,elix,readm,dead,wt,gcs,hr,sbp,map,dbp,rr,spo2,temp,fio2,k,na,cl,glu,urea,creat,mg,ca,al
t,bili,alb,hb,wcc,plt,aptt,pt,inr,ph,po2,pco2,be,lac,hco3,vent,norad,in,out,balance,fluid,vaso,action,state\n1,27002,0,56,7,0,0,66,15,102,93,
72,62,13,99,36.2,1.0,5.0,129,90,8.1,15,574,2.1,1.92,3,6,19,11.2,12.2,213,85.4,14.9,1.4,7.34,28.66,5.47,-3.0,1.6,26,0,0.04,2,0,10.7,2,2,7,309\
n2,27002,0,56,7,0,0,66,15,102,88,68,59,10,100,36.8,1.0,5.0,129,90,8.1,15,574,2.1,1.92,3,6,19,11.2,12.2,213,85.4,14.9,1.4,7.34,28.66,5.47,-3.
0,1.6,26,0,0.04,1,0,10.7,2,2,7,309\n3,27002,0,56,7,0,0,66,15,106,95,73,63,10,97,36.5,1.0,5.1,127,93,10.0,15,578,2.1,1.85,2,6,19,9.03,12.0,20
4,85.4,14.9,1.4,7.39,20.89,5.12,-0.4,1.3,23,0,0.07,325,0,12.0,5,2,22,309\n4,27002,0,56,7,0,0,66,15,103,89,71,60,8,98,36.3,1.0,4.6,129,93,9.6,
16,583,2.0,1.75,1,6,19,8.7,11.7,192,85.4,14.9,1.4,7.4,19.6,5.07,0.0,1.3,23,0,0.08,0,0,12.0,1,3,3,309\n5,27002,0,56,7,0,0,66,9,106,121,82,61,1
2,100,36.4,1.0,4.4,129,93,7.2,16,583,2.0,1.75,1,6,19,7.07,11.7,164,84.2,15.9,1.6,7.33,50.87,5.07,-4.6,3.4,22,1,0.0,0,0,12.0,1,1,1,144\n6,2700
2,0,56,7,0,0,66,3,107,122,75,54,14,100,36.4,0.86,4.7,130,96,6.3,15,526,2.1,1.75,1,6,19,8.89,11.7,114,52.0,14.9,1.4,7.37,54.33,4.98,-2.7,4.1,2
2,1,0.25,0,0,12.0,1,4,4,144\n7,27002,0,56,7,0,0,66,6,108,103,64,47,15,97,36.6,0.4,4.8,130,101,6.9,14,459,2.3,1.75,1,6,19,9.78,11.7,127,47.5,1
4.4,1.3,7.34,51.02,5.24,-3.8,4.2,22,1,0.52,3,0,12.0,2,5,10,144\n8,27002,0,56,7,0,0,66,11,101,106,69,52,15,96,36.7,0.4,4.9,130,101,4.4,14,459,
2.3,1.75,1,6,19,9.36,11.7,127,47.5,14.4,1.3,7.37,22.4,4.93,-3.0,3.8,22,1,0.63,2,0,12.1,2,5,10,165\n9,27002,0,56,7,0,0,67,7,99,106,69,51,16,9
9,36.8,0.4,5.3,132,100,5.0,15,506,2.3,1.75,1,6,19,9.22,15.6,176,42.2,13.7,1.2,7.35,11.8,5.38,-2.8,3.8,22,1,0.63,0,0,12.1,1,5,5,144\n10,27002,
0,56,7,0,0,75,4,88,110,71,51,13,100,36.9,0.4,5.5,133,100,7.4,15,512,2.3,1.75,1,7,19,9.2,16.2,183,41.5,13.6,1.2,7.34,4.93,5.6,-3.0,3.4,22,1,0.
61,4,0,12.1,2,5,10,144\n11,27002,0,56,7,0,0,75,6,72,96,62,46,13,100,37.1,0.4,5.2,133,100,4.7,15,512,2.3,1.83,5,8,19,8.79,16.2,183,41.5,13.6,
1.2,7.36,21.38,5.16,-3.0,3.1,22,1,0.68,0,0,12.1,2,5,10,144\n12,27002,0,56,7,0,0,75,9,80,96,62,46,12,99,37.2,0.4,5.1,133,99,5.3,15,512,2.3,1.9
2,6,8,20,7.96,16.2,183,41.5,13.6,1.2,7.37,25.24,4.84,-3.0,2.7,22,1,0.72,0,0,12.1,1,5,5,144\n13,27002,0,56,7,0,0,75,11,80,105,69,49,15,100,37.
3,0.36,5.1,134,99,6.0,15,512,2.3,2.01,6,8,20,7.96,16.2,183,41.5,13.6,1.2,7.35,22.74,5.05,-3.4,2.3,22,1,0.56,0,2,12.1,1,5,5,664\n14,27002,0,5
6,7,0,0,75,14,83,101,64,45,16,98,37.4,0.39,5.2,134,99,6.4,15,500,2.5,2.1,6,8,20,8.03,16.2,182,41.0,13.5,1.2,7.34,17.47,5.33,-3.0,1.9,22,0,0.6
8,0,0,12.1,1,5,5,689\n15,27002,0,56,7,0,0,75,15,82,104,67,47,15,100,37.6,0.4,5.6,134,99,6.2,13,442,2.5,2.18,6,8,20,8.4,15.9,180,38.2,13.2,1.
1,7.29,8.7,6.33,-3.8,1.6,22,0,0.68,0,0,12.1,1,5,5,689\n',
 'top_5_next_states': {752: 325, 751: 182, 367: 165, 393: 129, 664: 113},
 'top_5_alt_actions': [('Fluid: 21 ml/hr', 'Vasopressor: 0.04 mcg/kg/min'),
  ('Fluid: 80 ml/hr', 'Vasopressor: 0.04 mcg/kg/min'),
  ('Fluid: 80 ml/hr', 'Vasopressor: 0.27 mcg/kg/min')]}
```


Fig. F.1 **Initial explanation data in raw CSV format.** F.1