

1 Evaluating the Human Safety Net: Observational study 2 of Physician Responses to Unsafe AI Recommendations 3 in high-fidelity Simulation

4
5 Paul Fester^{1,2,*}, Myura Nagendran^{1,2,3,*}, Anthony C. Gordon^{1,3}, A. Aldo Faisal^{1,2,4,+} and
6 Matthieu Komorowski^{1,3,+}

7
8 ¹ UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, London,
9 UK

10 ² Brain & Behavior Lab: Departments of Bioengineering and Computing, Imperial College
11 London, London, UK

12 ³ Division of Anaesthetics, Pain Medicine and Intensive Care, Imperial College London,
13 London, UK

14 ⁴ Institute of Artificial and Auman Intelligence, Universität Bayreuth, Bayreuth, Germany

15
16
17 * Equal contribution

18 + These authors jointly supervised this work

19 20 **ORCID IDs:**

21 - Paul: 0000-0002-4856-1822

22 - Myura: 0000-0002-4656-5096

23 - Tony: 0000-0002-0419-547X

24 - Aldo: 0000-0003-0813-7207

25 - Matthieu: 0000-0003-0559-5747

26 27 **Competing interests:**

28 MK has received consulting fees from Philips Healthcare, and speaker honoraria from GE
29 Healthcare. The other authors declare that no competing interests.

30 31 **Open access:**

32 For the purpose of open access and as required by funders (UKRI), the authors have
33 applied a Creative Commons Attribution (CC BY) licence to any 'Author Accepted
34 Manuscript' version arising.

35 36 **Author contributions:**

37 MN, PF, MK, AG and AF conceived the study. MN and MK wrote the experimental vignettes.
38 MN, PF and MK recruited participants and conducted experiments. MN post-processed the
39 eye-tracking data. PF performed the initial data analysis. MN, PF, MK, AG and AF contributed
40 to the subsequent interpretation of the data. PF drafted the initial version of the manuscript.
41 MN, PF, MK, AG and AF contributed to critical revision of the manuscript for important
42 intellectual content and approved the final version.

43 44 **Funding:**

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

45 This work was funded by the University of York and the Lloyd's Register Foundation through
46 the Assuring Autonomy International Programme (Project Reference 03/19/07) and supported
47 by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre
48 (BRC). PF and MN were supported by a PhD studentship of the UKRI Centre for Doctoral
49 Training in AI for Healthcare (EP/S023283/1). ACG was supported by an NIHR Research
50 Professorship (RP-2015-06-018). AAF was supported by a UKRI Turing AI Fellowship
51 (EP/V025449/1). This study/project/report is independent research funded by the NIHR
52 (Artificial Intelligence, 'Validation of a machine learning tool for optimal sepsis treatment',
53 AI_AWARD01869).

54

55 **Data availability:**

56 The data and code (in the form of Jupyter notebooks) to reproduce the results and figures in
57 both the manuscript and the supplementary appendices are available at:
58 <https://figshare.com/s/78c5ff5c6031f701c0d1>

59

60

61

62 **ABSTRACT [153 words]**

63 In the context of Artificial Intelligence (AI)-driven decision support systems for high-stakes
64 environments, particularly in healthcare, ensuring the safety of human-AI interactions is
65 paramount, given the potential risks associated with erroneous AI outputs. To address this,
66 we conducted a prospective observational study involving 38 intensivists in a simulated
67 medical setting.

68 Physicians wore eye-tracking glasses and received AI-generated treatment
69 recommendations, including unsafe ones. Most clinicians promptly rejected unsafe AI
70 recommendations, with many seeking senior assistance. Intriguingly, physicians paid
71 increased attention to unsafe AI recommendations, as indicated by eye-tracking data.
72 However, they did not rely on traditional clinical sources for validation post-AI interaction,
73 suggesting limited "debugging."

74 Our study emphasises the importance of human oversight in critical domains and highlights
75 the value of eye-tracking in evaluating human-AI dynamics. Additionally, we observed human-
76 human interactions, where an experimenter played the role of a bedside nurse, influencing a
77 few physicians to accept unsafe AI recommendations. This underscores the complexity of
78 trying to predict behavioural dynamics between humans and AI in high-stakes settings.

79

80

81

82 INTRODUCTION

83

84 Artificial Intelligence (AI) driven systems are set to take an increasingly prominent supportive
85 role in decision-making including in high-stakes settings^{1,2}. While the final decision remains in
86 human hands, understanding how AI recommendations impact their user's behaviour is
87 crucial. In fact, recent work has highlighted differences in our perception of human and AI
88 advice.^{3,4} In high-risk shared decision-making settings where even non-autonomous AI-
89 decision-support tools have surprisingly high safety assurance requirements,⁵ understanding
90 the human-AI dynamic is key to assessing overall safety.

91

92 In this paper, we look at live safety assessment of AI for decision support in healthcare
93 because it combines many of the challenges that make AI safety assessment hard across the
94 board. Whilst AI recommendation engines have shown promising results on retrospective
95 data, the translation to the bedside has been slow due in part to concerns for patient safety.⁶
96 In situations where the optimal decision has historically been unclear,⁷⁻⁹ and with technologies
97 like reinforcement learning that aim at recommending decisions which surpass the standard
98 of care, the safety-assessment challenge becomes even harder. One such example is
99 cardiovascular management during sepsis where the optimal personalised doses of
100 intravenous (IV) fluids and vasopressors remain unknown.^{10,11}

101

102 Attempts have been made to improve the safety profile of AI-driven decision support in
103 retrospective intensive care settings.¹²⁻¹⁴ Still, the necessity of prospective and higher fidelity
104 evaluations involving clinical end users is clear from recent examples in other fields.¹³ For
105 instance, an acute kidney injury alert system showing good performance on retrospective data
106 was found to worsen outcomes when deployed in a real-world setting, illustrating the need for
107 a careful transition between retrospective testing and prospective deployment of digital
108 systems.¹⁵ Safely transitioning from "bytes to bedside" is a particularly complex challenge
109 because of the dynamic interaction with human users who are prone to biases and can behave
110 in unpredictable ways.¹⁶⁻¹⁸

111

112 In response to the growing emphasis on ecologically valid testing of AI systems,^{19,20} we run
113 our behavioural experiment in a physical simulation suite, a tool which has historically been
114 used as a widely accepted training tool for modelling high-fidelity situations and capturing
115 patterns of human behaviour with simulation now forming a core part of medical training.²¹
116 This immersive approach enables physicians to respond to bedside stimuli more realistically,
117 aligning their behaviours with actual clinical practice.²² Furthermore, this shift towards a more
118 realistic setting aligns with the evolving regulatory landscape surrounding AI, which
119 emphasises "human-centred AI" and the holistic evaluation of human-AI team
120 performance.^{23,24} AI safety assessment is not a mere problem of computer science but also
121 one of human-AI cooperation which should incorporate behavioural elements grounded in
122 human perceptual and decision-making studies.

123

124 We use the example of cardiovascular management in sepsis to study the behaviour of
125 physicians in response to AI recommendations. Here, we share the results of an observational
126 study of human-AI interaction in a high-fidelity simulation suite focusing on the influence of
127 safe and unsafe AI recommendations on treatment decisions. Using eye-tracking as a
128 behavioural marker, we show evidence that the attention placed by clinicians on AI
129 recommendations, as well as broader behavioural traits such as desire for senior advice,

130 depends on AI advice safety. We also demonstrate that most unsafe AI recommendations
131 would be appropriately rejected by the clinical team and recommend how clinical users of AI
132 should be trained to further improve their robustness to hazardous AI recommendations.

133 134 **RESULTS**

135 We conducted an observational human-AI interaction investigation within a high-fidelity
136 simulation facility. Our primary aim was to assess clinicians' ability to detect and appropriately
137 reject unsafe AI recommendations or seek senior assistance. Participants engaged in
138 simulated patient scenarios, prescribing fluid and vasopressor doses before and after
139 receiving AI guidance, see Figure 1. The scenarios encompassed safe, unsafe, and
140 "challenged" unsafe AI recommendations, with an experimenter acting as a bedside nurse
141 trying to change the clinician's decision in the latter case. Eye-tracking technology was used
142 to monitor participants' gaze patterns during simulations. Calibration and validation
143 procedures ensured accurate gaze data collection. We divided the visual field into regions of
144 interest (ROIs) to quantify attention patterns. Physicians were recruited from an ICU, and
145 ethical approval was obtained. Data analysis was conducted using Python. The full study
146 details, including the recruitment process and ethics approval, can be found in the Methods
147 section.

148
149 A total of 38 intensive care physicians took part in the experiment (Figure 2). This cohort
150 comprised 25 men (66%) and 13 women (34%), proportions in line with the national population
151 of intensivists.²⁵ The balance between junior and senior physicians (with less or more than 5
152 years of experience respectively) was even and 21% of participants reported having been
153 personally involved or having had experience, in AI research.

154
155 Each physician completed six (four safe, two unsafe) different patient scenarios leading to a
156 total of 228 recorded trials. Of these trials, 76 featured an unsafe AI recommendation and 152
157 were safe ones. See Supplementary Appendix D for the full trial matrix.

158
159 In total, unsafe AI recommendations were stopped more often than safe AI recommendations
160 (29% vs. 83%, $p < 0.0001$). The proportion stopping unsafe AI recommendations rose to 92%
161 ($p = 0.027$) when including physicians who asked for a senior opinion, which would most likely
162 lead to the unsafe AI recommendation being rejected (see Figure 3a). This analysis was
163 further expanded by categorising physicians into junior (<5 years of intensive care unit (ICU)
164 experience) and senior (≥ 5 years of ICU experience) practitioners. There was a non-significant
165 trend for junior physicians to stop AI recommendations less often than senior physicians (79%
166 vs. 83%, n.s.). Junior physicians asked more often for a second opinion than senior physicians
167 (65% vs. 25%, $p < 0.0001$), which led to more unsafe recommendations being stopped or
168 escalated by juniors (94% by juniors against 91% by seniors, n.s.).

169
170 Similarly, second-opinion requests rose from 40% before seeing any AI recommendation to
171 57% after seeing an unsafe AI recommendation ($p = 0.0056$) but the reduction in requests after
172 seeing a safe AI recommendation was not significant (figure 3b). Seeing an unsafe rather than
173 a safe AI recommendation triggered more senior/second opinion requests (57% vs. 36%,
174 $p = 0.0017$). Seeing unsafe AI recommendations therefore significantly increased the
175 proportion of requests for senior help.

176

177 As expected, prior to the AI recommendation being revealed, no significant difference in gaze
178 fixations on regions of interest (ROIs) was observed between safe and unsafe scenarios
179 regarding the three AI-independent regions (ICU data chart, vital signs monitor, and patient
180 mannequin), see Figure 3c. Subsequent to the disclosure of the AI recommendation, there
181 were more fixations on the AI screen in the unsafe scenarios (mean 960) versus safe
182 scenarios (mean 700) ($p=0.0015$, see Figure 3d). Finally, the number of gaze fixations on the
183 AI explanation ROI was not significantly different between safe and unsafe scenarios

184

185 The distributions of initial fluid and vasopressor dose prescriptions across participants in our
186 six scenarios are shown in Figure 4. These results show wide variation in clinical practice,
187 even when given the exact same information. Figure 4 suggests that the extent of the variation
188 in prescribing might depend on case-specific features (e.g. in scenario 2, the patient had
189 already had more fluids than in scenario 1 so physicians gave less fluid, or patient 5 had sepsis
190 related to infected and leaky heart valves so physicians were more careful with both fluid and
191 vasopressor), see Supplementary Appendix E for an extended discussion.

192

193 We also investigated the extent to which AI recommendations influenced prescription
194 decisions. Physicians changed their prescription (dose of fluids and/or vasopressors) in 46%
195 (105/228) of trials after seeing what the AI suggested. Both safe and unsafe AI
196 recommendations influenced human decisions to different extents: fluid doses shifted on
197 average by 80 ml/h (and vasopressor doses by 0.01 mcg/kg/min) after a safe AI
198 recommendation compared to 40 ml/h (and 0.08 mcg/kg/min) after an unsafe AI
199 recommendation. Figure 5 shows the shift of distribution in vasopressor prescriptions before
200 and after the AI recommendation was seen for two scenarios (split by whether the entire cohort
201 is considered or only those physicians who did not ask for senior/second opinion). In scenarios
202 (such as number two) where the unsafe AI recommendation was significantly influencing, this
203 did not seem apparent when exclusively considering the physicians who did not request senior
204 help (see Supplementary Appendix F for this plot over all scenarios).

205

206 Finally, each physician had one of the two unsafe scenarios extended with a “challenge”
207 section where the bedside nurse (a member of the experimental team) was given three
208 attempts to change the physician’s mind on whether or not to stop the AI recommendation
209 were it to be automatically implemented. In 95% of cases (36/38), the verbal input challenge
210 from the bedside nurse did not sway the physician's decision to accept or reject the automated
211 application of the AI recommendation. However, two participants (both junior) were persuaded
212 to change their minds from interrupting an unsafe recommendation to accepting it. Human-to-
213 human interactions can therefore also play a role in the inadvertent adoption or appropriate
214 rejection of unsafe recommendations.

215

216

217 **DISCUSSION**

218

219 Our findings confirm that AI recommendations can influence clinician behaviour and thereby
220 impact patient care. Unsafe AI recommendations, represented here as sudden under- or over-
221 dosing, were frequently (but not entirely) detected and appropriately mitigated by the clinical
222 team (by rejecting the AI recommendation). However, junior physicians more often deferred
223 the decision to senior colleagues when they were unsure about the safety of an AI

224 recommendation. This shows the importance of educating clinical teams who will interact with
225 a new AI recommender system on the correct intended use of the system, including target
226 patient population, indications, and limitations, as well as the importance of clinical context
227 when integrating the AI recommendations into their practice.
228

229 This study reinforces the call for more interdisciplinary and realistic human-AI interaction
230 studies on domain experts.^{26,27} Critically, our experimental design also allowed us to study
231 human-human behavioural dynamics during an encounter with AI decision support. This is
232 important as most clinical uses of AI-driven decision support tools will be in the context of
233 multi-disciplinary teams where humans other than the final decision-maker can still positively
234 or negatively influence the interaction between the final human decision-maker and the AI.
235 This is why our study was run in a simulation suite: an environment that reproduces natural
236 stimuli of bedside practice for clinicians without any risk to patients.
237

238 While eye-tracking is typically used in controlled environments,²⁸ this study demonstrates the
239 feasibility of using this behavioural phenotypic marker of attention in more realistic, less
240 constrained, environments. Our findings indicate that physicians fixated more on unsafe than
241 safe AI recommendations implying an appropriately higher level of allocated cognitive
242 attention. However, we also observed that physicians did not rely more on AI explanations in
243 the unsafe scenarios calling into question the use of explanations as a mitigation strategy for
244 unsafe AI. Nor did physicians devote significantly more attention to looking back at the
245 'traditional' (non-AI based) clinical data after seeing an unsafe AI recommendation to
246 understand why the recommendation might be unsafe (i.e. there was no outward evidence of
247 a desire to 'debug' the unsafe AI recommendation).
248

249 The influence of AI recommendations on clinical judgement has already been studied in
250 vignette-type experiments.³ This work goes one step closer to clinical deployment by studying
251 these interactions in a high-fidelity simulation environment. This setup enabled the study of
252 human-AI interaction with eye-tracking as well as the ability to investigate human-human
253 interactions as they relate to AI. Most studies of clinical decision support system safety use
254 medication error as the primary outcome measure and proxy for patient safety.²⁹ Here, we
255 look at systems that are not yet deployed in clinical practice, so measuring prescription error
256 rate directly (and correlating that to 'error' without a gold standard) is challenging. Therefore,
257 we took the problem from a different angle and aimed to estimate the ability of clinicians to
258 spot unsafe treatment recommendations from an AI tool.
259

260 However, the limitations of the study should also be acknowledged. First, as raised by many
261 physicians during the initial briefing, prescribing hourly fluid and vasopressor doses directly is
262 unusual for intensive care physicians who typically indicate blood pressure (and other
263 parameters) targets and let the bedside nurse titrate the actual doses within a reasonable
264 range to reach the set targets. Similarly, the simulation limited the physician's action space to
265 one specific aspect of patient care, preventing action plans that might go beyond the defined
266 possibilities. Moreover, making treatment decisions for the next hour is also less dynamic than
267 real clinical practice (where for example the ability to examine a real patient and use advanced
268 cardiac output monitors might add to the nuance of the overall clinical picture).
269

270 From a different perspective, one could challenge the definitions of safe and unsafe
271 recommendations used in the scenarios by arguing that there is no ground truth in sepsis

272 resuscitation and that they are therefore subjective. One might even go further and argue that
273 strategies that under- or over-dose in specific patients (compared to the ‘average’) could be
274 desirable in some cases. The scenarios used in this experiment were designed for the unsafe
275 recommendations to be inappropriate to a majority of clinicians and validated by an
276 independent panel of intensivists. The introduction of AI-driven decision support tools,
277 particularly those using reinforcement learning aims to improve patient outcomes beyond the
278 current standard of care.^{9,30} This means that such systems will give recommendations that
279 differ from what the clinical team would ordinarily have done but potentially without explanation
280 - a “mysterious oracle dilemma” where the AI oracle recommends actions that on average lead
281 to better outcomes but might occasionally be suboptimal, and the users do not get context on
282 the AI recommendation. It will therefore be essential for humans to exert critical thinking and
283 assess how reasonable the AI recommendation is to filter potentially novel but superior calls
284 by the AI from harmful recommendations.

285 As regulators push toward requiring clear intended purpose statements for software as
286 medical devices,³¹ our high-fidelity eye-tracking based approach to evaluating an AI-driven
287 decision support tool serves as a basis for promoting the generation of safety evidence.
288 Furthermore, the recent rise in popularity of generative AI (most notably as large language
289 models) highlights the safety concerns of hallucinatory outputs that might be acted upon in a
290 clinical setting and bring harm to patients.³² It is likely that an AI system that shows overall
291 superhuman performance in a given task will still show lower-quality performance in some
292 specific cases.³³ Solutions such as uncertainty-aware models or explainable AI might help
293 users differentiate between well-informed recommendations and flawed calls.^{34–36} The human-
294 AI interactions at the bedside, with a particular focus on high-pressure decision-making, would
295 also help to accelerate the safe translation of AI-based decision support tools to the bedside.

296
297

298 **CONCLUSION**

299

300 It is critical for clinician acceptance, regulatory compliance and real-world adoption that we
301 evaluate cooperation between clinical experts and AI decision support tools in high-fidelity
302 settings - in our case a simulated intensive care unit. This study demonstrates the influence
303 of AI recommendations on clinical behaviour and suggests that the vast majority of unsafe AI
304 recommendations are appropriately rejected by bedside clinicians. The findings on junior
305 physicians occasionally accepting an unsafe AI recommendation and their general willingness
306 to seek senior help when unsure should inform the intended use (i.e. some tools might need
307 to only be used by junior clinicians if they have access to senior advice). Uncertainty
308 awareness, novel forms of AI interpretability and a better understanding of human-human
309 interactions (i.e. team decisions) in the context of AI-driven decision support will help not only
310 with assuring safety from a regulatory perspective but also in fostering confidence and
311 approval from physician end-users.

312

313 **METHODS**

314

315 **Objective** - We conducted an observational human-AI interaction study in a high-fidelity
316 simulation facility. Our primary objective was to measure whether participants were able to
317 detect, and correctly reject, unsafe recommendations from an AI tool and/or ask for senior
318 help when appropriate. Secondary study objectives included: (i) quantifying the shift in fluid
319 and vasopressor doses induced by seeing an AI recommendation, and (ii) determining

320 whether or not gaze patterns varied differentially depending on the safety status of the AI
321 recommendation.

322

323 **Experimental design** - Participants (clinicians) were briefed on the experiment and completed
324 a pre-experiment questionnaire recording their demographics and prior experience with AI
325 (see Supplementary Appendix A for the full content of the briefing and questionnaire).
326 Participants were told that they would conduct a review of several adult patients with sepsis
327 within a simulation suite (Imperial College Simulation Centre) and that they would need to
328 prescribe appropriate doses of fluid and vasopressor for each patient both before and after
329 getting advice from an AI tool. Critically, physicians were told that the AI recommendation
330 engine had been successfully validated in multiple retrospective settings but had not been
331 prospectively evaluated. The simulation layout is shown in Figure 1a.

332

333 Each physician completed a total of six different patient scenarios, simulating a virtual “ward
334 round”. Each scenario started with physicians entering the simulation suite and conducting
335 their assessment of the patient as they saw fit. Data sources within the room included a
336 standard paper ICU bedside data chart with observations and blood results, an ICU handover
337 note including details of the patient’s presentation and medical history, a vital signs monitor
338 and a physical patient mannequin (Simman 3G, Laerdal Medical, Stavanger, Norway) which
339 could be examined (see Supplementary Appendix B for the details of each patient scenario).
340 All patient scenarios were crafted by clinical experts from the authors team and to fit the needs
341 of this simulation experiment, they do not come from real patients. A member of the research
342 team played the role of the bedside ICU nurse who could only give standardised responses to
343 any questions. Following their assessment of the patient, physicians were asked to
344 recommend a dose rate for fluids (ml/hr) and vasopressors (noradrenaline, mcg/kg/min) for
345 the coming hour (to match the format of AI recommendations). Physicians rated their
346 confidence on a 1-10 scale and whether or not they would like support for their decision from
347 a senior physician (or a second opinion if the physician was already senior themselves). They
348 were then shown the AI recommendation, asked to what extent they agreed with the
349 recommendation on a 5-point Likert scale (from completely disagree to completely agree), and
350 then the initial dosing-related questions again (what dose they would prescribe, their
351 confidence level, optional ask for senior help - see figure 1b).

352

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

358

359 The running of a single patient scenario from entry into the simulation suite to exit constituted
360 one trial. Trials were categorised by the nature of the AI tool’s recommendation provided to
361 the participant: safe, unsafe or “challenged” unsafe. In the latter, after the physician reported
362 whether or not they would stop the AI recommendation if it were to be automatically
363 administered, the bedside nurse was permitted three attempts (all following a standardised
364 script) to verbally try to convince them to change their mind (see Supplementary Appendix C
365 for the scripts). Each physician experienced four safe trials, one unsafe and one “challenged”
366 unsafe in a pseudo-randomised order (see the trial matrix in Supplementary Appendix D). The
367 first trial encountered by every physician was always in the safe condition to establish a

368 baseline level of trust with the AI tool and let the physician familiarise themselves with the
369 environment. The details of each patient scenario are presented in Supplementary Appendix
370 B.

371
372 All AI recommendations were synthetically generated by the research team for the purpose of
373 ensuring a standardised experimental format (i.e. they were not from a real AI system). The
374 definition of unsafe recommendations was based on extreme under- or over-dosing of fluid
375 and/or vasopressor as per previous work.¹³ All participating physicians were fully debriefed at
376 the conclusion of the study on the synthetic nature of the AI recommendations so as not to
377 bias their opinions of future interactions with AI-driven systems.

378
379 During each trial, all physician responses were recorded by a member of the research team
380 sitting in a dead angle in the simulation suite. This data, along with questionnaire answers was
381 reformatted and analysed in Python (code available online here:
382 <https://figshare.com/s/78c5ff5c6031f701c0d1>)
383

384 **Eye-tracking for gaze recording**
385 In this study, gaze was employed as an indicator of physicians' attentional focus during
386 simulations, with particular interest in whether this varied according to the safety of the AI
387 recommendation. Pupil and first-person videos were recorded with non-invasive commercially
388 available eye-tracking glasses (Pupil Core headset). The Pupil Labs software (Core, version
389 3.3) utilised both eye cameras to delineate the pupil and estimate the direction of gaze within
390 the recorded field of view.

391
392 Prior to the experiment, a two-part 2D calibration procedure was conducted. The initial stage
393 involved a static calibration using five screen markers on a laptop display (default Pupil Labs
394 'screen marker' calibration). Subsequently, a depth-based static exercise was performed,
395 requiring participants to focus on nine screen markers sequentially ('natural features' mode)
396 displayed on a 60-inch TV screen, initially at 1 metre and then at 2 2-metre distance. A laptop
397 (Lenovo Thinkpad) was connected to the eye-tracking glasses for the entire experiment. To
398 allow for unrestricted movement in the suite, the glasses were connected via USB to a battery-
399 powered laptop (Lenovo Thinkpad) worn by physicians in a lightweight backpack.

400
401 Because of variability in facial morphologies, 19/38 physicians passed the calibration
402 exercises and had their gaze-based attention data collected. Physicians were instructed to
403 point to where they were reading on the handover note at the start of each scenario as a final
404 level of validation that the eye tracking was appropriately calibrated.

405
406 We defined four key regions of interest (ROIs) (Figure 1c): the paper ICU data chart, the vital
407 signs monitor, the patient mannequin (Laerdal Simman 3G) and the AI display screen. Four
408 further sub-regions were identified within the AI screen ROI corresponding to four types of
409 explanation for the AI recommendation. April tags (simple QR codes) within the simulation
410 suite (see Figure 1c) were used to identify ROIs in post-processing. As is common practice in
411 eye-tracking literature^{37,38}, we used the number of gaze fixations per ROI—a fixation being
412 the predominant eye movement occurring when the foveal region of the visual field is held
413 stationary— as a proxy for participant attention.

414

415 **Participant recruitment and simulation facility** - Recruitment of ICU physicians made use
416 of both convenience sampling and targeted advertising to a local NHS trust (Imperial College
417 Healthcare NHS Trust) Inclusion criteria were: (i) practising physician, (ii) has worked for two
418 or more months in an adult ICU, (iii) currently works in ICU or has worked in ICU within the
419 last 6 months. Physicians were compensated for their time and each experiment lasted
420 approximately 60 minutes. The study was approved by the Research Governance and
421 Integrity Team (RGIT) at Imperial College London and the UK Health Research Authority (Ref:
422 22/HRA/1610).
423
424
425
426
427

428 **REFERENCES**

429

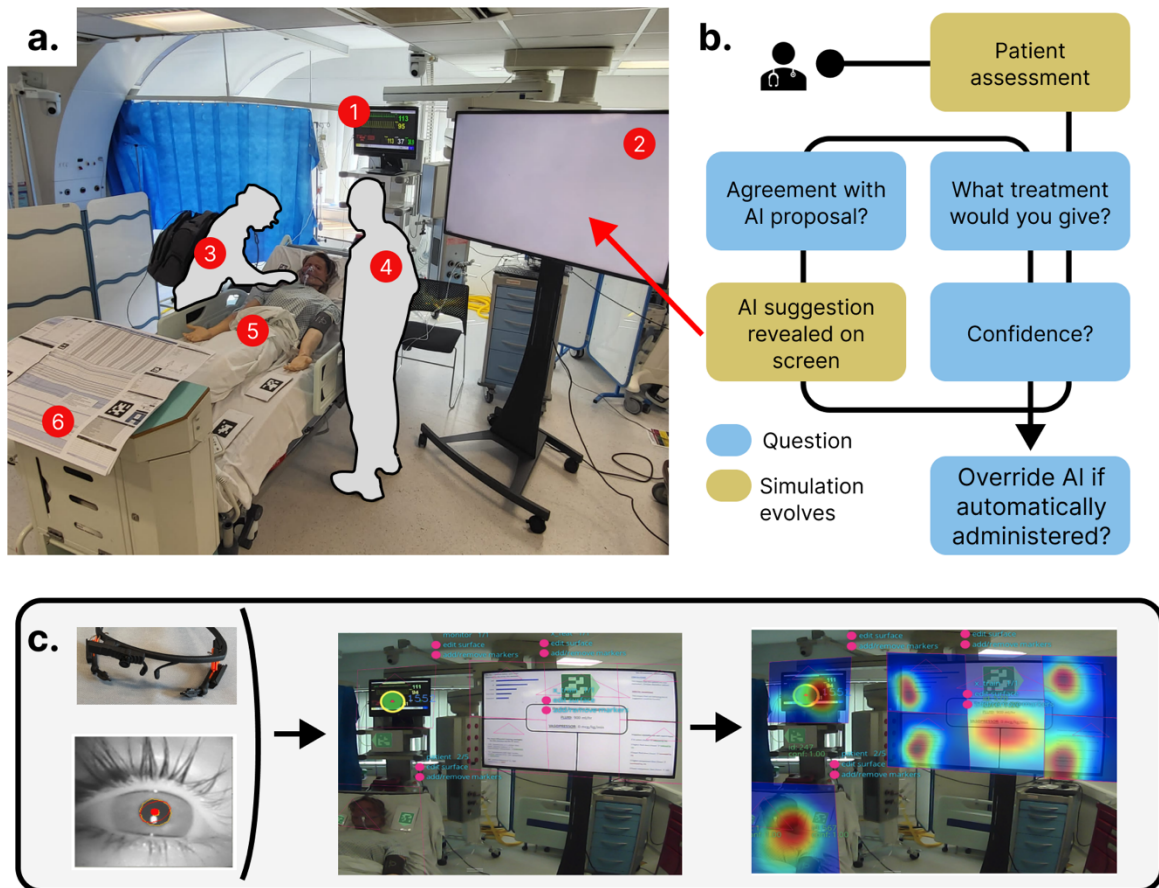
- 430 1. Wang G, Liu X, Ying Z, Yang G, Chen Z, Liu Z, et al. Optimized glycemic control of type
431 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nat Med*. 2023 Sep
432 14;1–10.
- 433 2. Zhan X, Xu H, Zhang Y, Zhu X, Yin H, Zheng Y. DeepThermal: Combustion
434 Optimization for Thermal Power Generating Units Using Offline Reinforcement
435 Learning. *Proc AAAI Conf Artif Intell*. 2022 Jun 28;36(4):4680–8.
- 436 3. Nagendran M, Festor P, Komorowski M, Gordon A, Faisal AA. Quantifying the impact
437 of AI recommendations with explanations on prescription decision making: an
438 interactive vignette study [Internet]. 2023 [cited 2023 Jun 19]. Available from:
439 <https://www.researchsquare.com>
- 440 4. Köbis N, Bonnefon JF, Rahwan I. Bad machines corrupt good morals. *Nat Hum Behav*.
441 2021 Jun;5(6):679–85.
- 442 5. Festor P, Habli I, Jia Y, Gordon A, Faisal AA, Komorowski M. Levels of Autonomy and
443 Safety Assurance for AI-Based Clinical Decision Systems. In Springer; 2021. p. 291–6.
- 444 6. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving
445 from bytes to bedside: a systematic review on the use of artificial intelligence in the
446 intensive care unit. *Intensive Care Med*. 2021;1–11.
- 447 7. Tejedor M, Woldaregay AZ, Godtliebsen F. Reinforcement learning application in
448 diabetes blood glucose control: A systematic review. *Artif Intell Med*. 2020 Apr
449 1;104:101836.
- 450 8. Prasad N, Cheng LF, Chivers C, Draugelis M, Engelhardt BE. A Reinforcement
451 Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units
452 [Internet]. arXiv; 2017 [cited 2023 Jan 6]. Available from:
453 <http://arxiv.org/abs/1704.06300>
- 454 9. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence
455 clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*.
456 2018;24(11):1716–20.
- 457 10. Yealy DM, Mohr NM, Shapiro NI, Venkatesh A, Jones AE, Self WH. Early Care of
458 Adults With Suspected Sepsis in the Emergency Department and Out-of-Hospital
459 Environment: A Consensus-Based Task Force Report. *Ann Emerg Med*. 2021;
- 460 11. van der Ven W, Schuurmans J, Schenk J, Roerhorst S, Cherpanath T, Lagrand W, et
461 al. Monitoring, management, and outcome of hypotension in Intensive Care Unit
462 patients, an international survey of the European Society of Intensive Care Medicine. *J*
463 *Crit Care*. 2022;67:118–25.
- 464 12. Jia Y, Burden J, Lawton T, Habli I. Safe reinforcement learning for sepsis treatment. In:
465 2020 IEEE International conference on healthcare informatics (ICHI). IEEE; 2020. p. 1–
466 7.
- 467 13. Festor P, Jia Y, Gordon AC, Faisal AA, Habli I, Komorowski M. Assuring the safety of
468 AI-based clinical decision support systems: a case study of the AI Clinician for sepsis
469 treatment. *BMJ Health Care Inform*. 2022;
- 470 14. Peng X, Ding Y, Wihl D, Gottesman O, Komorowski M, Lehman L wei H, et al.
471 Improving sepsis treatment strategies by combining deep and kernel-based
472 reinforcement learning. In American Medical Informatics Association; 2018. p. 887.
473 Available from:
474 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371300/pdf/2975959.pdf>
- 475 15. Wilson FP, Martin M, Yamamoto Y, Partridge C, Moreira E, Arora T, et al. Electronic
476 health record alerts for acute kidney injury: multicenter, randomized clinical trial. *BMJ*.
477 2021 Jan 18;372:m4786.
- 478 16. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with
479 medical decisions: a systematic review. *BMC Med Inform Decis Mak*. 2016 Nov
480 3;16(1):138.
- 481 17. Sujan M, Furniss D, Grundy K, Grundy H, Nelson D, Elliott M, et al. Human factors

- 482 challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care*
483 *Inform* [Internet]. 2019;26(1). Available from:
484 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7252977/>
- 485 18. Dawson NV, Arkes HR. Systematic errors in medical decision making: judgment
486 limitations. *J Gen Intern Med*. 1987;2(3):183–7.
- 487 19. Quinan PS, Padilla LM, Creem-Regehr SH, Meyer M. Towards ecological validity in
488 evaluating uncertainty. In: *Proceedings of Workshop on Visualization for Decision*
489 *Making Under Uncertainty (VIS'15)* [http://vdl.sci.utah](http://vdl.sci.utah.edu/publications/2015_vdmu_ecologicalvalidity)
490 [edu/publications/2015_vdmu_ecologicalvalidity](http://vdl.sci.utah.edu/publications/2015_vdmu_ecologicalvalidity) [Internet]. 2015 [cited 2023 Sep 29].
491 Available from: <https://miriah.github.io/publications/eco-validity-vdmu.pdf>
- 492 20. Madras D, Pitassi T, Zemel R. Predict responsibly: improving fairness and accuracy by
493 learning to defer. *Adv Neural Inf Process Syst*. 2018;31.
- 494 21. Cato DL, Murray M. Use of Simulation Training in the Intensive Care Unit. *Crit Care*
495 *Nurs Q*. 2010 Mar;33(1):44.
- 496 22. Chang M, Büchel D, Reinecke K, Lehmann T, Baumeister J. Ecological validity in
497 exercise neuroscience research: A systematic investigation. *Eur J Neurosci*. 2022
498 Jan;55(2):487–509.
- 499 23. Article 8 [Internet]. *Artificial Intelligence Act*. [cited 2023 Sep 29]. Available from:
500 <https://artificialintelligenceact.com/title-iii/chapter-2/article-8/>
- 501 24. Software and AI as a Medical Device Change Programme [Internet]. GOV.UK. 2023
502 [cited 2023 Oct 2]. Available from:
503 [https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-](https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme)
504 [change-programme](https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme)
- 505 25. Women in Intensive Care Medicine | The Faculty of Intensive Care Medicine [Internet].
506 [cited 2023 Jan 4]. Available from:
507 <https://www.ficm.ac.uk/careersworkforceworkforce/women-in-intensive-care-medicine>
- 508 26. Xu W, Dainoff MJ, Ge L, Gao Z. From human-computer interaction to human-AI
509 Interaction: new challenges and opportunities for enabling human-centred AI. *ArXiv*
510 *Prepr ArXiv210505424*. 2021;5.
- 511 27. Zhang Y, Liao QV, Bellamy RK. Effect of confidence and explanation on accuracy and
512 trust calibration in AI-assisted decision making. In: *Proceedings of the 2020 conference*
513 *on fairness, accountability, and transparency*. 2020. p. 295–305.
- 514 28. Cao S, Huang CM. Understanding User Reliance on AI in Assisted Decision-Making.
515 *Proc ACM Hum-Comput Interact*. 2022;6(CSCW2):1–23.
- 516 29. Ranji SR, Rennke S, Wachter RM. Computerised provider order entry combined with
517 clinical decision support systems to improve medication safety: a narrative review. *BMJ*
518 *Qual Saf*. 2014 Sep;23(9):773–80.
- 519 30. Zhang K, Wang H, Du J, Chu B, Arévalo AR, Kindle R, et al. An interpretable RL
520 framework for pre-deployment modelling in ICU hypotension management. *Npj Digit*
521 *Med*. 2022 Nov 18;5(1):1–10.
- 522 31. Crafting an intended purpose in the context of software as a medical device (SaMD)
523 [Internet]. GOV.UK. [cited 2023 Mar 28]. Available from:
524 [https://www.gov.uk/government/publications/crafting-an-intended-purpose-in-the-](https://www.gov.uk/government/publications/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd)
525 [context-of-software-as-a-medical-device-samd/crafting-an-intended-purpose-in-the-](https://www.gov.uk/government/publications/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd)
526 [context-of-software-as-a-medical-device-samd](https://www.gov.uk/government/publications/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd/crafting-an-intended-purpose-in-the-context-of-software-as-a-medical-device-samd)
- 527 32. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their
528 hallucinations. *Crit Care*. 2023 Mar 21;27(1):120.
- 529 33. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. *Dataset Shift in*
530 *Machine Learning*. MIT Press; 2022. 246 p.
- 531 34. Festor P, Luise G, Komorowski M, Faisal AA. Enabling risk-aware Reinforcement
532 Learning for medical interventions through uncertainty decomposition. *ICML*. 2021;
- 533 35. Trombley CM, Gulum MA, Ozen M. Evaluating Uncertainty-Based Deep Learning
534 Explanations for Prostate Lesion Detection. *MLHC*. 2022;
- 535 36. Shafti A, Derks V, Kay H, Faisal AA. The Response Shift Paradigm to Quantify Human
536 Trust in AI Recommendations [Internet]. *arXiv*; 2022 [cited 2023 Jan 19]. Available

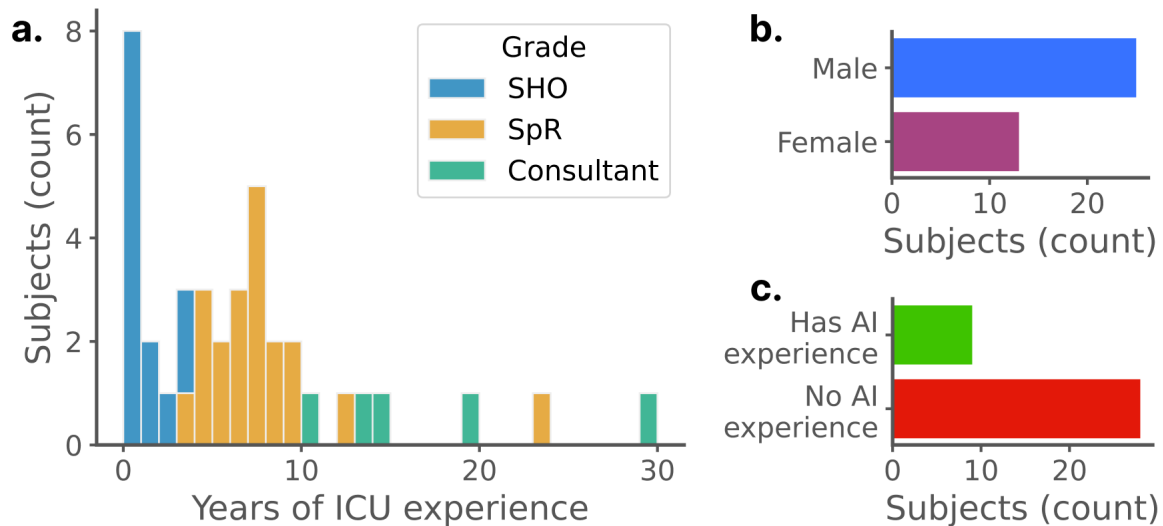
- 537 from: <http://arxiv.org/abs/2202.08979>
538 37. Harston JA, Faisal AA. Methods and Models of Eye-Tracking in Natural Environments.
539 In: Eye Tracking: Background, Methods, and Applications. Springer; 2022. p. 49–68.
540 38. Gidlöf K, Wallin A, Dewhurst R, Holmqvist K. Using eye tracking to trace a cognitive
541 process: Gaze behaviour during decision making in a natural environment. J Eye Mov
542 Res. 2013;6(1).
543

544 FIGURES

545

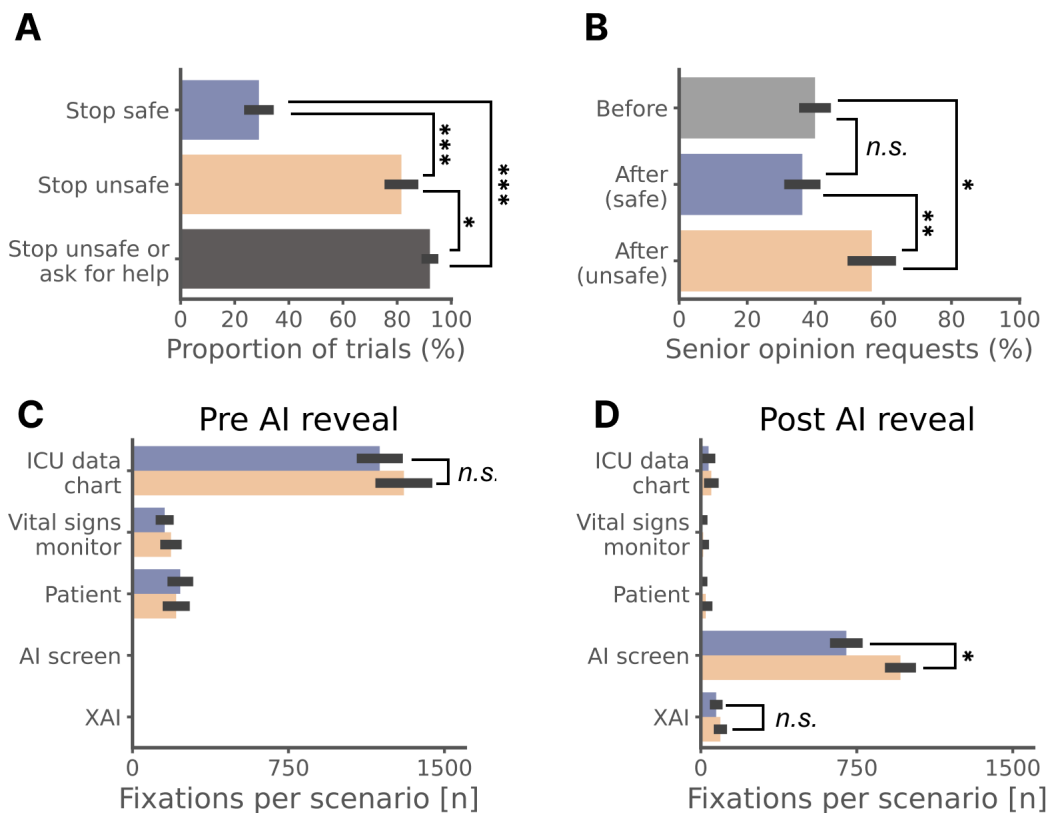


546
547 **Figure 1: Experimental design – A.** Photo of the simulation suite with: (1) Bedside monitor
548 (2) AI screen (3) Participant (4) Bedside nurse (5) Patient mannequin (6) Intensive care unit
549 (ICU) bedside information chart. **B.** Experimental protocol diagram. **C.** Gaze-based attention
550 extraction pipeline: eye-tracking glasses, pupil camera view, a recorded field of view with
551 April tags (QR codes) and reconstructed data with fixation heatmaps on the different regions
552 of interest (ROIs).
553



554
555
556
557
558
559

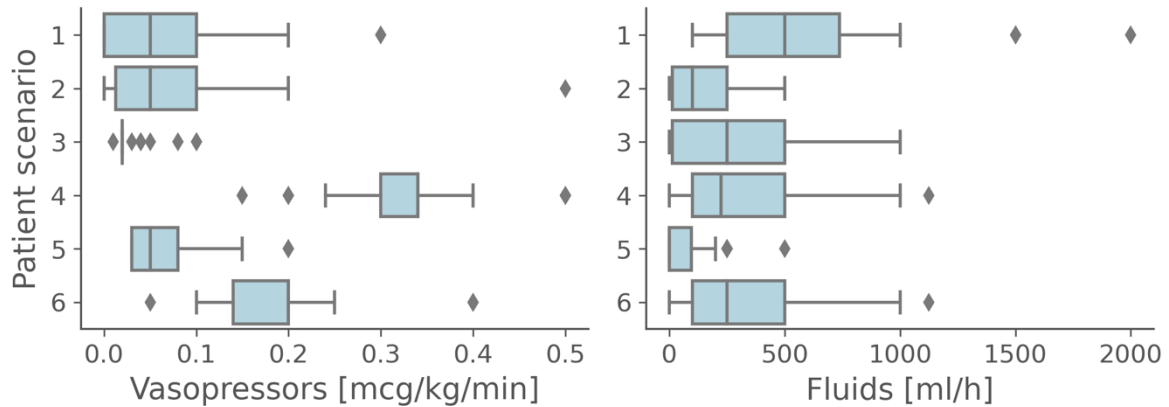
Figure 2: Recruited cohort demographics – A. Distribution of intensive care experience. **B.** Gender distribution. **C.** Proportion of physicians who had ever been involved in a research project involving AI. This cohort covers the whole range of experience levels, is in line with national gender ratios and contains both people who have and have not worked with AI.



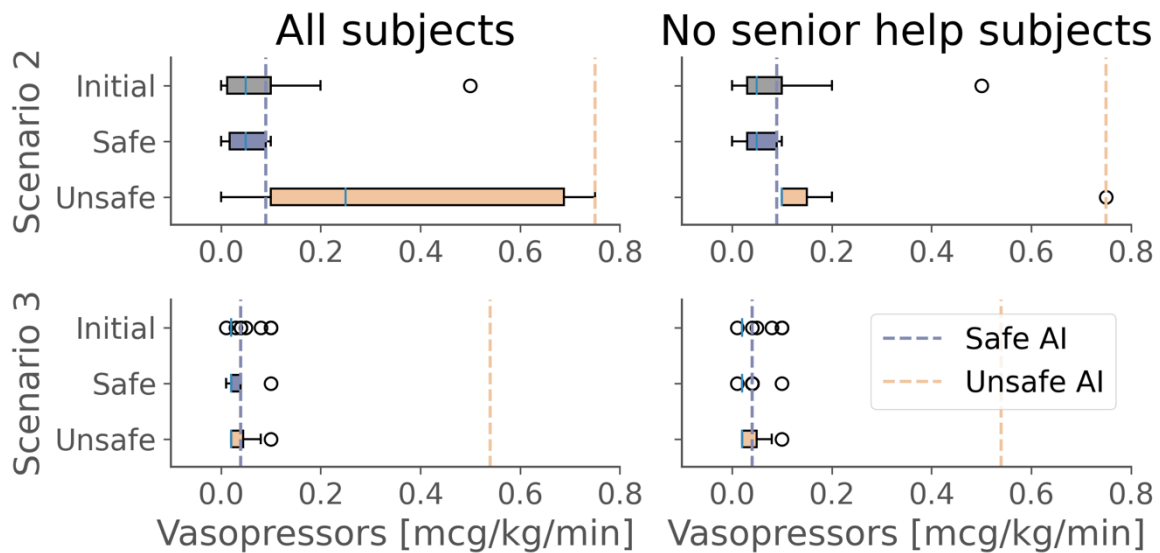
560
561
562
563
564
565

Figure 3: Impact of AI recommendation safety status on clinician decisions, and gaze fixations on each ROI. – A. Bar chart of the proportion of stopped safe recommendations, stopped unsafe recommendations and stopped or escalated unsafe recommendations. **B.** Proportions of requests for senior help before seeing any recommendation and after having seen a safe or an unsafe one. **C.** Number of gaze fixations on each ROI before revealing the

566 AI recommendation (i.e. there can be no fixations on the AI) **D.** Number of gaze fixations on
 567 each ROI after revealing the AI recommendation (when clinicians have already evaluated
 568 the non-AI information sources and so would be expected to look at these much less). ***
 569 $p < 0.0001$, ** $p < 0.005$, * $p < 0.05$, *n.s.* not significant.
 570



571
 572 **Figure 4: Clinical practice variability** – Distribution of initial (i.e. pre AI reveal) vasopressor
 573 (left) and fluid (right) prescriptions by physicians for each patient scenario.
 574



575
 576 **Figure 5: Shift analysis** – Vasopressors dose distribution before and after having seen a
 577 safe or unsafe AI recommendation for two patient scenarios across all physicians (left) and
 578 only those who did not ask for senior/second review (right). Unsafe AI recommendations do
 579 not always influence the final decision. Physicians most influenced by unsafe AI
 580 recommendations tend to ask for a second opinion, while those who do not ask for help are
 581 less influenced by unsafe recommendations.
 582

Appendix A - Pre-experiment questionnaire

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

Appendix B - Nurse challenge scripts

	If participant WOULD override unsafe AI suggestion	If participant would NOT override unsafe AI suggestion
Size 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?"

Appendix C - Patient scenarios

Patient 1 handover note for participants:

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

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

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.
- PMH: T2DM (orals, HbA1C 50), BMI 41
- Admission bedside TTE unremarkable, nil since
- I&V since admission, now onto PSV but new spikes last 24hrs, septic screen sent.
- PSV 10/6 with sats 93 on FiO2 0.45.
- Had 5 day 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.
- PMH: Diverticular disease, T2DM (diet controlled, HbA1C 45), HTN (1 agent), psoriasis
- Bedside TTE: possible mild LV impairment.
- Norad 0.34 (up from peak 0.21 in theatre)
- Fluid balance +ve 6.5L last 12 hrs
- Latest lac 5.8, UO 15ml over last 3 hrs

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

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
- PMH: T1DM (HbA1C 94), prev left big toe amputation
- ECG: sinus tachy
- CXR: clear (on admission)
- Bedside TTE: hyperdynamic LV
- Norad 0.14, started 3 hrs ago, rising
- Fluid balance +7.5L last 12 hrs
- Latest lac 8.3, UO 80-150 ml/hr last few hours

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
- All 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 D - Trial matrix

Subject n°	Decision point						Trial type
	1	2	3	4	5	6	
1	S	S	S	S	U	C	
2	S	S	S	U	S	C	
3	S	S	U	S	S	C	
4	S	U	S	S	S	C	
5	S	S	S	S	C	U	
6	S	S	S	C	U	S	
7	S	S	U	S	C	S	
8	S	U	S	S	C	S	
9	S	S	S	U	S	C	
10	S	S	S	C	U	S	
11	S	S	U	C	S	S	
12	S	U	S	C	S	S	
13	S	S	C	S	S	U	
14	S	S	C	S	U	S	
15	S	S	C	U	S	S	
16	S	U	C	S	S	S	
17	S	C	S	S	S	U	
18	S	C	S	S	U	S	
19	S	C	S	U	S	S	
20	S	C	U	S	S	S	
21	S	S	S	S	C	U	
22	S	S	S	U	S	C	
23	S	S	U	S	S	C	
24	S	U	S	S	S	C	
25	S	S	S	S	C	U	
26	S	S	S	U	C	S	
27	S	S	U	S	C	S	
28	S	U	S	S	C	S	
29	S	S	S	C	S	U	
30	S	S	S	C	U	S	
31	S	S	U	C	S	S	
32	S	U	S	C	S	S	
33	S	S	C	S	S	U	
34	S	S	C	S	U	S	
35	S	S	C	U	S	S	
36	S	U	C	S	S	S	
37	S	C	S	S	S	U	
38	S	C	S	S	U	S	

Appendix E - Initial dose discrepancies discussion

Figure 4 in the main text shows the distribution of doses initially decided by subjects by drug and scenario. In light of the patient scenarios described above, we here propose a discussion of Figure 4.

Patient 1 was given more fluids than patient 2 likely because it has received more fluids so far. Patient 5 was given significantly less fluids than others, most likely because subjects wanted to avoid causing harm to a patient with heart sepsis and potentially damaged valves.

Appendix F - Dose distribution shift for all scenarios

